Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek A

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume A

Inteligentni sistemi
Intelligent Systems

Uredila / Edited by

Rok Piltaver, Matjaž Gams

http://is.ijs.si

7. oktober 2015 / October 7th, 2015
Ljubljana, Slovenia

Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek A

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume A

Inteligentni sistemi

Intelligent Systems

Uredila / Edited by

Matjaž Gams, Rok Piltaver

7. oktober 2015 / October 7th, 2015
Ljubljana, Slovenia

Urednika:

Rok Piltaver
Odsek za inteligentne sisteme
Institut »Jožef Stefan«, Ljubljana

Matjaž Gams
Odsek za inteligentne sisteme
Institut »Jožef Stefan«, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič, Mitja Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2015

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

659.2:316.42(082)(0.034.2)
659.2:004(082)(0.034.2)

MEDNARODNA multikonferenca Informacijska družba (18 ; 2017 ; Ljubljana)
Inteligentni sistemi [Elektronski vir] : zbornik 18. mednarodne multikonference Informacijska družba - IS 2015, 7. oktober 2015, [Ljubljana, Slovenia] : zvezek A = Intelligent systems : proceedings of the 18th International Multiconference Information Society - IS 2015, October 7th, 2015, Ljubljana, Slovenia : volume A / uredila, edited by Rok Piltaver, Matjaž Gams. - El. zbornik. - Ljubljana : Institut Jožef Stefan, 2015

Način dostopa (URL): http://is.ijs.si/zborniki/!%20A%20-%20Inteligentni%20sistemi%20-%20ZBORNIK.pdf

ISBN 978-961-264-083-5 (pdf)
1. Gl. stv. nasl. 2. Vzp. stv. nasl. 3. Dodat. nasl. 4. Piltaver, Rok
28923943

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2015

Multikonferenca Informacijska družba (http://is.ijs.si) je z osemnajsto zaporedno prireditvijo osrednji srednjeevropski dogodek na področju informacijske družbe, računalništva in informatike. Letošnja prireditev traja tri tedne in poteka na Fakulteti za računalništvo in informatiko in Institutu »Jožef Stefan«.

Informacijska družba, znanje in umetna inteligenca se razvijajo čedalje hitreje. V vse več državah je dovoljena samostojna vožnja inteligentnih avtomobilov, na trgu je moč dobiti čedalje več pogosto prodajanih avtomobilov z avtonomnimi funkcijami kot »lane asist«. Čedalje več pokazateljev kaže, da prehajamo v naslednje civilizacijsko obdobje, hkrati pa so konflikti sodobne družbe čedalje težje razumljivi.

Letos smo v multikonferenco povezali dvanajst odličnih neodvisnih konferenc. Predstavljenih bo okoli 300 referatov v okviru samostojnih konferenc in delavnic, prireditev bodo spremljale okrogle mize in razprave ter posebni dogodki kot svečana podelitev nagrad. Referati so objavljeni v zbornikih multikonference, izbrani prispevki pa bodo izšli tudi v posebnih številkah dveh znanstvenih revij, od katerih je ena Informatica, ki se ponaša z 38-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2015 sestavljajo naslednje samostojne konference:

- Inteligentni sistemi
- Kognitivna znanost
- Izkopavanje znanja in podatkovna skladišča
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Vzgoja in izobraževanje v informacijski družbi
- Soočanje z demografskimi izzivi
- Kognitonika
- Delavnica »SPS EM-zdravje«
- Delavnica »Pametna mesta in skupnosti kot razvojna priložnost Slovenije«
- Druga študentska konferenca s področja računalništva in informatike za doktorske študente
- Druga študentska konferenca s področja računalništva in informatike za vse študente
- Osma mednarodna konferenca o informatiki v šolah: razmere, evolucija in perspektiva.

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija, SLAIS in Inženirska akademija Slovenije. V imenu organizatorjev konference se zahvaljujemo združenjem in inštitucijam, še posebej pa udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V 2015 bomo tretjič podelili nagrado za življenjske dosežke v čast Donalda Michija in Alana Turinga. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe bo prejel prof. dr. Jurij Tasič. Priznanje za dosežek leta je pripadlo dr. Domnu Mungosu. Že petič podeljujemo nagradi »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono je dobilo počasno uvajanje informatizacije v slovensko pravosodje, jagodo pa spletna aplikacija »Supervizor«. Čestitke nagrajencem!

Niko Zimic, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2015

In its 18th year, the Information Society Multiconference (http://is.ijs.si) remains one of the leading conferences in Central Europe devoted to information society, computer science and informatics. In 2015 it is extended over three weeks located at Faculty of computer science and informatics and at the Institute "Jožef Stefan".

The pace of progress of information society, knowledge and artificial intelligence is speeding up. Several countries allow autonomous cars in regular use, major car companies sell cars with lane assist and other intelligent functions. It seems that humanity is approaching another civilization stage. At the same time, society conflicts are growing in numbers and length.

The Multiconference is running in parallel sessions with 300 presentations of scientific papers at twelve conferences, round tables, workshops and award ceremonies. The papers are published in the conference proceedings, and in special issues of two journals. One of them is Informatica with its 38 years of tradition in excellent research publications.

The Information Society 2015 Multiconference consists of the following conferences:

- Intelligent Systems
- Cognitive Science
- Data Mining and Data Warehouses
- Collaboration, Software and Services in Information Society
- Education in Information Society
- Facing Demographic Challenges
- Cognitonics
- SPS EM-Health Workshop
- Workshop »Smart Cities and Communities as a Development Opportunity for Slovenia«
- 2nd Computer Science Student Conference, PhD Students
- 2nd Computer Science Student Conference, Students
- 8th International Conference on Informatics in Schools: Situation, Evolution, and Perspective.

The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM, SLAIS and the Slovenian Engineering Academy. In the name of the conference organizers we thank all societies and institutions, all participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

For 2013 and further, the award for life-long outstanding contributions will be delivered in memory of Donald Michie and Alan Turing. The life-long outstanding contribution to development and promotion of information society in our country is awarded to Dr. Jurij Tasič. In addition, a reward for current achievements was pronounced to Dr. Domnu Mungosu. The information strawberry is pronounced to the web application "Supervizor, while the information lemon goes to lack of informatization in the national judicial system. Congratulations!

Niko Zimic, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki-Špetič
Mitja Lasič
Robert Blatnik
Mario Konecki
Vedrana Vidulin

Programme Committee

Nikolaj Zimic, chair
Franc Solina, co-chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič Grega
Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah

KAZALO / TABLE OF CONTENTS

Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek A

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume A

Inteligentni sistemi

Intelligent Systems

Uredila / Edited by

Matjaž Gams, Rok Piltaver

7. oktober 2015 / October 7th, 2015
Ljubljana, Slovenia

PREDGOVOR

Konferenca Inteligentni sistemi je med tradicionalnimi konferencami multikonference »Informacijska družba« tudi letos, tako kot vsa pretekla leta od 1997 dalje. Konferenca se ukvarja z inteligentnimi sistemi in inteligentnimi storitvami v informacijski družbi oziroma konkretnimi tehničnimi rešitvami v inteligentnih sistemih, možnosti njihove praktične uporabe, pa tudi trendi, perspektivami, prednostmi in slabostmi, priložnostmi in nevarnostmi, ki jih v informacijsko družbo prinašajo inteligentni sistemi.

Nesluteni razvoj informacijske družbe in zlasti umetne inteligence se nadaljuje, čeprav povprečni državljan o tem malo ve. Nekoč utopične ideje Raya Kurtzweila o točki singularnosti in preskoku v novo človeško ero se zdijo čedalje bližje.

Tudi letos konferenca Inteligentni sistemi sestoji iz mednarodna dela in delavnice; prispevki so tako v slovenskem kot angleškem jeziku. V letu 2015 je bilo okoli 30 prispevkov. Prispevki so bili recenzirani s strani vsaj dveh anonimnih recenzentov, avtorji pa so prispevke popravili po navodilih recenzentov. V ločeni sekciji pa so predstavljeni prispevki Delavnice E9. Večina prispevkov obravnava raziskovalne dosežke v Odseku za inteligentne sisteme Instituta Jožef Stefan. Hkrati s predstavitvijo poteka tudi aktivna analiza prispevkov vsakega predavatelja in diskusija o bodočih raziskavah.

Rok Piltaver in Matjaž Gams, predsednika konference

PREFACE

The conference Intelligent Systems remains a traditional part of the multiconference Information Society since its beginnings in 1997. The conference addresses important aspects of information society: intelligent computer-based systems, the corresponding intelligent services; it addresses technical aspects of intelligent systems, their practical applications, as well as trends, perspectives, advantage and disadvantages, opportunities and threats that are being brought by intelligent systems into the information society.

In 2015, the pace of progress in information society and intelligent systems further increases, oblivious to an average citizen. Once regarded as utopist, the ideas of Ray Kurtzweil promoting the point of singularity where the human civilization will embrace a new, intelligent era, are becoming widely accepted.

As previously, the conference Intelligent Systems 2015 consists of an international event and the workshop, and presents around 30 papers written in both English and Slovenian language. The papers have been reviewed by at least two anonymous reviewers, and the authors have modified their papers according to the remarks. In a separate section, papers from the E9 workshop are presented. Most of them present research at the Department of Intelligent Systems at the Jožef Stefan Institute, Ljubljana, Slovenia. Each presentation consists of the classical paper report, and further includes analysis of the researcher’s achievements and future research of each presenter in the workshop manner.

Rok Piltaver and Matjaž Gams, Conference Chairs

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Rok Piltaver, IJS (co-chair)
Matjaž Gams, IJS (co-chair)
Marko Bohanec, IJS
Tomaž Banovec, Statistični zavod
Cene Bavec, UP, FM
Jaro Berce, UL, FDV
Marko Bonač, ARNES
Ivan Bratko, UL, FRI in IJS
Dušan Caf, FIŠ
Bojan Cestnik, Temida
Aleš Dobnikar, CVI
Bogdan Filipič, IJS
Nikola Guid, FERI
Borka Jerman Blažič, IJS
Tomaž Kalin, DANTE
Mario Konecki, FOI
Marjan Krisper, FRI
Marjan Mernik, FERI
Vladislav Rajkovič, FOV
Ivo Rozman, FERI
Nikolaj Schlamberger, Informatika
Tomaž Seljak, IZUM
Miha Smolnikar, IJS
Peter Stanovnik, IER
Damjan Strnad, FERI
Peter Tancig, RZ
Pavle Trdan, Lek
Iztok Valenčič, GZ
Vasja Vehovar, FDV
Martin Žnidaršič, IJS

Analiza signalov EKG z globokimi nevronskimi mrežami

Jani Bizjak
Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova cesta 39
1000 Ljubljana, Slovenija
+386 1 477 3147
jani.bizjak@ijs.si

Igor Kononenko
Laboratorij za kognitivno modeliranje, Fakulteta za računalništvo in informatiko
Večna pot 113
1000 Ljubljana, Slovenija
+386 1 479 8230
igor.kononenko@fri.uni-lj.si

POVZETEK
V tem prispevku je na kratko obrazloženo delovanje srca ter zaznavanje srčnih bolezni s pomočjo elektrokardiografije. Predstavljena je ideja globokih nevronskih mrež, natančneje povratnih nevronskih mrež ter nekaj optimizacijskih metod, ki so nujno potrebne za učenje velikih mrež. V zadnjem delu prispevka so predstavljeni rezultati klasifikacije kompleksa QRS, kjer povratna nevronska mreža doseže 99 odstotno klasifikacijsko točnost.

Ključne besede
povratne nevronske mreže, RNN, elektrokardioagrafija, EKG

1. UVOD
Kljub napredku medicine v zadnjem stoletju, je še vedno kar 30 odstotkov vseh smrti povezanih s kardiološkimi težavami. Ena glavnih metod za zgodnje odkrivanje srčnih obolenj je uporaba elektrokardiograma (EKG). Kljub temu, da je diagnostika s pomočjo elektrokardiograma uspešna, se veliko simptomov pokaže skozi daljše časovno obdobje opazovanja, na primer, ko je telo dlje časa v stresu. Da bi lahko zanesljivo ugotovili, ali ima oseba okvaro na srcu bi tako bilo potrebno daljše (več dnevno) snemanje EKG. Samo snemanje v tem primeru ni problem, problem je količina podatkov, ki jih tako snemanje vrne. Medicinskemu osebju bi preverjanje več urnih (dnevnih) izpisov iz EKG vzelo preveč časa, poleg tega bi zaradi monotonega dela in posledično popuščanja koncentracije prihajalo do številnih napak. Reševanja problema se je potrebno lotiti s pomočjo strojnega učenja.

Čeprav je področje zaradi pomembnosti zelo raziskano, trenutni state-of-the-art algoritmi še vedno ne dosegajo dovolj visoke klasifikacijske točnosti, ki je še posebej pomembna v medicini. V tem prispevku smo predstavili pristop k analizi signala EKG s pomočjo globokih nevronskih mrež, natančneje povratnih nevronskih mrež (ang. Recurrent neural network - RNN).

Najprej smo se lotili enostavnejšega problema detekcije kompleksov QRS v signalu EKG. Kompleks QRS predstavlja najvišjo amplitudo nihanja sledi v signalu EKG. S pomočjo kompleksa QRS je najlažje razbrati število utripov na minuto (ang. Beats per minute – BPM), katero se poleg ostalih simptomov pogosto uporablja pri diagnozi bolezni.

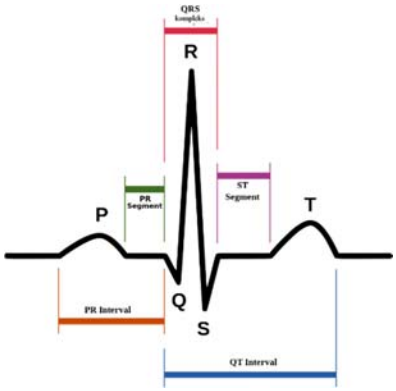
Delovanje mreže smo testirali na podatkovni bazi MIT-BIH [1]. Podatkovna baza vsebuje 48, 30 minutnih posnetkov, ki so bili posneti med letom 1975 in 1979 v sklopu laboratorija za aritmijo BIH na 47 različnih ljudeh. 23 posnetkov je bilo izbranih naključno, ostalih 25 iz enake skupine ljudi pa tako, da so vsebovali redkejše primere aritmij. Skupno je v bazi označenih približno 110.000 primerov na kanalu V1 in II odvodu.

2. ELEKTROKARDIAGRAM
Srce s prehajanjem ionov v in iz celice spreminja električni potencial ter s tem tvori električni pulz, ki potuje skozi srce in povzroča krčenje srčnih mišic. Ta vrsta električnega pulza je tako močna, da jo lahko zaznamo tudi na koži po celem telesu. Elektrokardiogram je naprava, ki s pomočjo elektrod pritrjenih na kožo meri električne spremembe, ki so posledica bitja srca.

Odvisno od števila ter pozicije elektrod kardiogram prikazuje srce iz različnih perspektiv. Najbolj pogosta je uporaba 10 elektrod, ki delovanje srca prikazuje v 12 sledeh (ang. leads).

Signal na posnetku EKG tvori značilne valove imenovane PQRST (Slika 1). Srčni utrip se začne z valom P, ki predstavlja depolarizacijo atrija. Kompleks QRS je najznačilnejši del signala, nastane zaradi krčenja ventriklov, ki kri črpajo nazaj v telo.

S pomočjo kompleks QRS lahko ugotovimo BPM ter nekatere bolezni srca, ki popačijo obliko kompleksa. V tem prispevku bomo obravnavali le iskanje normalnega kompleksa QRS in neenakomeren utrip, ki oblike kompleksa QRS ne spreminja.



Slika 1: Kompleks PQRST [2].[1]

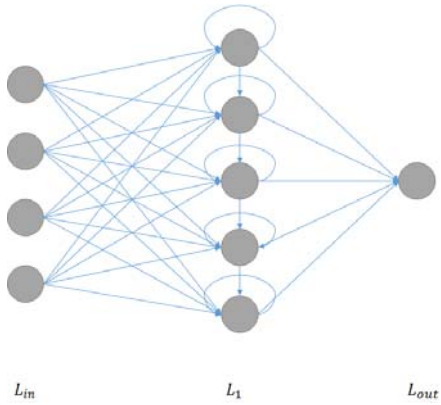
3. ARHITEKTURA MREŽE
Pri analizi signalov EKG niso pomembne le značilke v trenutnem oknu temveč tudi tiste, ki so se zgodile nekaj sekund pred tem. Standardne nevronske mreže (npr. večplastni perceptron) niso zmožne prepoznavati časovne odvisnosti med podatki. Povratne nevronske mreže so posebna vrsta globokih nevronskih mrež, ki so se zmožne naučiti časovnih odvisnosti med podatki. To omogočajo posebne povratne plasti, kjer so nevroni znotraj ene plasti povezani med seboj ter sami nase (Slika 2).

Zaradi arhitekture, je RNN podvržen problemu izginjajočega (ang. vanishing) ali eksplozivnega (ang. exploding) gradienta. To se

zgodi, ko se zaradi globine mreže in napak pri zaokroževanju vrednost gradienta približa ničli, hkrati pa se poveča napaka, kar onemogoči učenje takšnih mrež. RNN zato ni zmožen učenja (uporabe) podatkov, ko so le ti preveč narazen. Ta problem rešuje tako imenovane kratko-dolgo ročna spominska mreža (ang. long-short term memory – LSTM).

3.1 Učenje povratnih nevronskih mrež

Za učenje povratnih nevronskih mrež je potrebno spremeniti standardno vzvratno razširjanje tako, da deluje na rekurzivnih povezavah povratne plasti [3]. Povratno plast si lahko predstavljamo kot več standardnih skritih nevronskih plasti, ki so skrčene v eno plast. Vsaka naslednja plast v tem konstruktu predstavlja podatke iz prejšnjih učenj (t_k). Pri učenju tako najprej razširimo povratno plast za k časovnih enot, k v tem primeru predstavlja število časovnih enot, ki si jih mreža lahko zapomni. Na razširjeni mreži lahko nato uporabimo standardno vzvratno razširjanje. Ko konstrukt nato ponovno skrčimo, povprečimo uteži vseh plasti konstrukta ter povprečje uporabimo za uteži povratne plasti.



Slika 2: Nevroni so v RNN povezani znotraj ene plasti (L1).

3.2 Optimizacija gradientnega spusta

Zaradi dolgih časov učenja globokih nevronskih mrež je nujno potrebna optimizacija posodabljanja uteži, ki pripomore k hitrejši konvergenci.

Splošno uporabljena optimizacija je stohastični gradientni spust (SGD), ki za vsak vhodni podatek in vsak parameter posebej posodablja vrednosti uteži. Takšnega načina učenja, se kljub temu, da je relativno dober, ne da paralelizirati zato je čas učenja zelo dolg. To se lahko do neke mere reši z uporabo blokovskega gradientnega spusta. Ta vrednosti gradientov ne računa za vsak par podatkov posebej, temveč izračuna povprečje za manjši blok.

SGD tekom učenja uporablja fiksno stopnjo učenja (α) vse parametre pa posodablja globalno. S pristopom ohlajanja (ang. annealing) lahko pri SGD spreminjamo stopnjo učenja - navadno je v začetku učenja večja proti koncu pa konvergira k 0.

$$\theta = \theta + \alpha(\Delta f)$$

Metoda ADAGrad spreminja velikost gradienta za vsak parameter posebej. To stori z uporabo normalizacije na podlagi velikosti gradientov.

$$v = v + (\Delta f)^2$$

$$\theta = \theta + \alpha \frac{\Delta f}{\sqrt{v} + \varepsilon}$$

Pri tem nastane problem, ko so gradienti zelo visoki, navadno se to pojavlja v začetku učenja ter po določenem času učenja, saj se gradienti akumulirajo iz prejšnjih iteracij. Rahlo spremenjena metoda ADADelta [4] to popravi z vpeljavo dodatnih uteži v normalizacijsko funkcijo ter omejitvijo intervala, ki ga uporablja akumulator. Če v spodnji enačbi za m vzamemo vrednost iz intervala [0,1], se bo vpliv preteklih vrednosti eksponentno zmanjševal. Prav tako z uporabo prirejene sigmoidne funkcije, močno zmanjšamo vpliv prvih nekaj normalizacijskih uteži.

$$v = v * m + \frac{1}{1 + e^{-\frac{t}{5}}}(\Delta f)^2; m \in (0,1)$$

Na podoben način deluje tudi metoda RMSProp [5], ki v akumulator gradienta doda parameter razpada β (ang. decay) ter s tem omili vpliv normalizacije. V naših testih se je izkazalo, da RMSProp najhitreje konvergira k rešitvi, razlika v klasifikacijski točnosti pa je minimalna.

$$v = \beta v + (1 - \beta)(\Delta f)^2$$

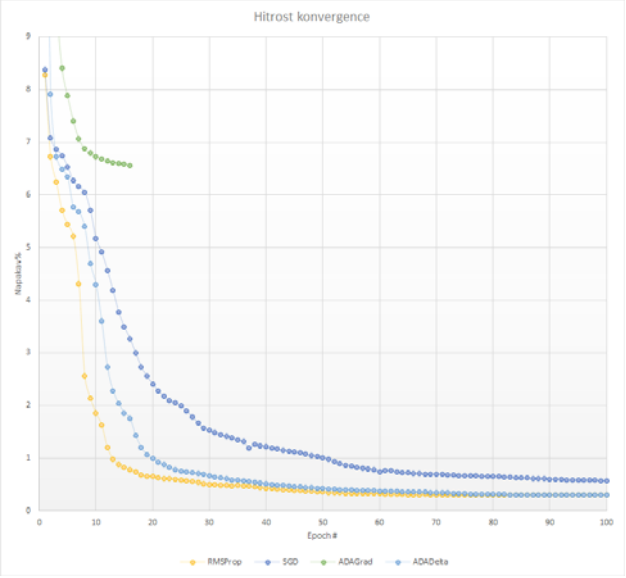
$$\theta = \theta - \alpha \frac{\Delta f}{\sqrt{v} + \varepsilon}$$

4. EKSPERIMENTALNI REZULTATI

4.1 Primerjava optimizacijskih metod

Čas učenja je, kljub izvajanju algoritmov na grafični kartici s 1000 jedri kjer je procesorska zasedenost večino časa 100%, zelo dolg. Za mrežo z 200 skritimi nevroni ter učno množico z 400.000 okni je učenje potekalo več dni. Preverili smo, kako se obnesejo različne metode za optimizacijo gradientnega spusta.

Za najhitrejšo se je izkazal RMSProp, ki je konvergiral skoraj 10 krat hitreje od bločnega SGD. ADAGrad se je zaradi visokih gradientov v začetku učenja ustavil predčasno, modifikacija ki smo jo naredili za ADADelta pa se je izkazala za relativno dobro. Na Slika 3 je prikazana hitrost konvergence za prvih 100 iteracij učenja.

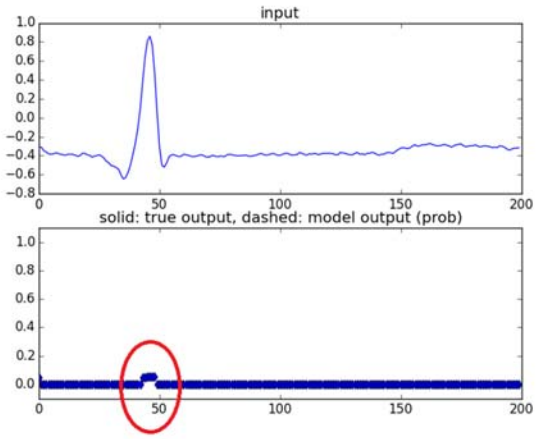


Slika 3: RMSProp konvergira najhitreje.

4.2 Iskanje kompleksa QRS

Podatke iz posameznih posnetkov smo razdelili v okna velikosti 200. Pri tem smo morali biti pozorni, da oken med sabo nismo premešali, saj so podatki časovno odvisni med seboj. Za testiranje natančnosti smo uporabili princip prečnega preverjanja leave-one-patient-out. Učili smo se na vseh posnetkih, brez posnetka na katerem smo testirali natančnost učenja. Zaradi časovne zahtevnosti učenja smo od vsakega posnetka vzeli le prvih 5.000 oken s po 200 signali, za testno množico pa smo vzeli prvih 40.000 oken posnetka.

Za klasifikacijo v en oz. drugi razred, smo preverili točnost metode pri različnih mejah verjetnosti, za najboljšo se je izkazala meja 6%, za klasifikacijo v pozitivni razred. V splošnem so bile verjetnosti s katerimi je mreža prepoznala kompleks QRS zelo majhne. Primer pravilne klasifikacije je prikazan na Slika 4.



Slika 4: Detekcija kompleksa QRS.

V Tabela 1 so prikazani rezultati testiranja. Kljub visoki povprečni točnosti 99,7% vidimo, da sta preciznost in senzitivnost zelo nizki. Pri nekaterih posnetkih se mreža zelo slabo nauči prepoznavati komplekse QRS, zaradi veliko večjega števila negativnih primerov v signalu (vse kar ni QRS vrh) je točnost kljub vsemu visoka.

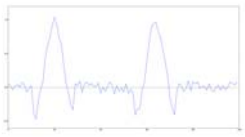
Tabela 1 Rezultati prečnega preverjanja detekcije kompleksov QRS

Datoteka	Preciznost	Senzitivnost	Specifičnost	Točnost
100	1	1	1	1
101	1	0,9739	1	0,9999
102	0,9688	0,7266	0,9999	0,9989
103	1	1	1	1
104	0,9660	0,9930	0,9999	0,9998
106	0,9826	0,9496	0,9999	0,9998
107	0,8014	0,9915	0,9991	0,9991
108	1	0,0099	1	0,9971
109	1	0,0067	1	0,9956
111	0	0	1	0,9958
113	1	0,9891	1	1
114	0	0	1	0,9974
115	0,9905	1	1,0000	1,0000
116	1	0,9776	1	0,9999
117	0,8929	0,9843	0,9996	0,9995
119	0,6752	0,9550	0,9985	0,9984
121	0	0	1	0,9969
122	1	0,9793	1	0,9999
123	1	0,7059	1	0,9993
124	0,3481	0,6267	0,9947	0,9931
201	1	0,9603	1	0,9998
202	1	0,9882	1	1,0000
203	1	0,4136	1	0,9966
205	1	0,0067	1	0,9956
207	0,1338	0,9661	0,9662	0,9662
209	1	1	1	1
210	1	0,1477	1	0,9962
212	1	1	1	1
213	1	1	1	1
214	0,1764	1	0,9733	0,9734
217	0,0859	0,1339	0,9946	0,9914
219	0,7561	0,9394	0,9988	0,9986
220	1	1	1	1
221	1	0,7836	1	0,9991
222	0,8658	1	0,9994	0,9994
228	0,2143	0,2903	0,9961	0,9935
230	1	0,9420	1	0,9998
231	1	0,9189	1	0,9997
232	0	0	1	0,9971
233	1	1	1	1
Povprečje:	0,7714	0,7090	0,9980	0,9969
Mediana:	1	0,9576	1	0,9993

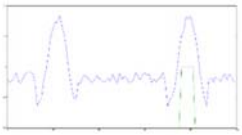
4.3 Detekcija neenakomernega utripa

Pri detekciji neenakomernega utripa (PAC) se je izkazalo, da so si kompleksi QRS preveč narazen, da bi se jih naša mreža lahko naučila. Pogosto se je dogajalo, da je gradient postal ničen po le nekaj iteracijah učenja. Poleg tega smo imeli težave z velikostjo učne množice. Odločili smo se, da bomo to vrsto podatkov sintetizirali. Naredili smo umetne QRS-komplekse, ki so bili med seboj oddaljeni približno polovico manj kot v pravem signalu. Razdaljo med dvema kompleksoma smo določili naključno, znotraj 40% dolžine (umetnega) intervala RR. Na Slika 5a je prikazan primer normalnega utripa, na Sliki 5b pa primer zakasnelega utripa. Zelena črta predstavlja verjetnost, da gre za zakasneli utrip. Prva

slika zaradi negativnih vrednosti, ki jih signal doseže zaradi šuma, izgleda kot da ima verjetnost višjo kot 0.



Slika 5a: Normalen utrip.

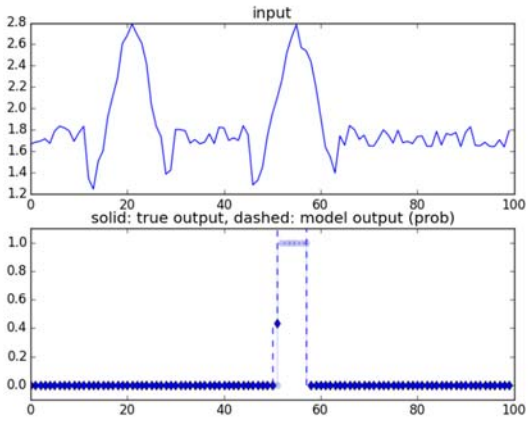


Slika 5b: Zapozneli utrip.

V primeru ko sta kompleksa QRS preblizu ali predaleč (20% okrog robov naključnega intervala), vrh označimo kot neenakomeren utrip. Izkazalo sej je, da se je RNN sposoben naučiti najprej prepoznavati komplekse QRS, nato pa še, ali je razdalja med dvema sosednjima kompleksoma pravilna. V Tabela 2 so prikazani rezultati pri prepoznavanju neenakomernega utripa. Tudi tukaj opazimo večjo razliko med senzitivnostjo ter specifičnostjo. Na Slika 6 je prikazan primer pravilne klasifikacije prezgodnjega utripa. V zgornjem delu slike je prikazan signal, v spodnjem delu pa verjetnost, da gre za neenakomeren utrip.

Tabela 2 Točnost prepoznavanja neenakomernega utripa

Št. Nevronov	Senzitivnost	Specifičnost	Klas. Točnost
150	0,928	0,965	0,929
200	0,945	0,974	0,947
400	0,918	0,960	0,918



Slika 6: Pravilno klasificiran prezgodnji utrip.

5. Zaključek

Pokazali smo, da se povratne nevronske mreže lahko uporabljajo za analizo signalov EKG. Poleg prepoznavanja statičnih lastnosti (kompleks QRS) lahko prepoznavajo tudi časovne odvisnosti med podatki (PAC). Izkaže se, da so podvržene problemu dolgotrajnega spomina, njihova klasifikacijska točnost pa nekoliko zaostaja za ostalimi state-of-the-art metodami [6],[7]. Kot možno rešitev tega problema predlagamo uporabo LSTM, ki omogoči poljubno dolgo hranjenje informacij.

6. Reference

[1] Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. Engineering in Medicine and Biology Magazine, IEEE, 20(3), 45-50.

[2] Wikimedia Commons, Schemantic diagram of normal sinus rhythm for a human heart as seen on ECG (2007) [obiskano 15.9.2015], URL https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg

[3] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550-1560.

[4] M. D. Zeiler, ADADELTA: an adaptive learning rate method, CoRR abs/1212.5701. URL http://arxiv.org/abs/1212.5701

[5] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning 4.

[6] Ka, A. K. (2011). ECG beats classification using waveform similarity and RR interval. arXiv preprint arXiv:1101.1836.

[7] Yu, S. N., & Chou, K. T. (2008). Integration of independent component analysis and neural networks for ECG beat classification. Expert Systems with Applications, 34(4), 2841-2846.

Decision Support Systems for Parkinson's Disease: State of the Art and the "PD_manager" Approach

Marko Bohanec
Jožef Stefan Institute, Department of Knowledge Technologies
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
marko.bohanec@ijs.si

ABSTRACT

We present the results of a literature-based study of medical decision support systems (DSS), with focus of Parkinson's disease (PD) management. The study was motivated by the needs of the EU H2020 project "PD_manager", which aims to develop innovative, mobile-health, patient-centric platform for PD management. The core element of the platform will be a DSS for supporting the physician and other caregivers in their monitoring of patients and deciding about their medication plans. In the present study, we describe the state-of-the-art of clinical DSSs in general, and specifically those related to PD. On this basis, we also propose the main design principles and functionality of the envisioned PD_manager DSS.

Categories and Subject Descriptors

H.4.2 [**Types of Systems**]: Decision support.
J.3 [**Life and Medical Sciences**]: Medical information systems.

Keywords

Decision Support Systems, Decision Modeling, Parkinson's Disease, Health Care, Multi-Criteria Models, Expert Modelling.

1. INTRODUCTION

Parkinson's disease (PD) [10] is a neurodegenerative disorder predominantly characterized by motor symptoms: tremor, rigidity, bradykinesia and postural instability. PD is also associated with non-motor symptoms, such as loss of taste and sense of smell, sleep disturbances, gastrointestinal complications, and many others. PD requires complicated, individual and long-term disease management in order to ensure that the patient retains his/her independence and continues to enjoy the best quality of life possible. The EPDA Consensus Statement [10] proposes to manage PD in a multidisciplinary way with special emphasis on accurate diagnosis, access to support services, continuous care, and actively involving PD patients in managing their illness.

"PD_manager" [22] is an EU Horizon 2020 project aimed at developing an innovative, mobile-health, patient-centric platform for PD management. Primary motor symptoms such as tremor, bradykinesia and postural imbalance, and non-motor symptoms, such as sleep, speech and cognitive disorders, will be evaluated with data captured by light, unobtrusive, co-operative, mobile devices: sensor insoles, a wristband and the patient's or caregiver's smartphone. Data mining studies will lead to the implementation of a Decision Support System (DSS) aimed at making suggestions for modifications in the medication, which is the key for maintenance and prolongation of patients' independence and improved quality of life.

In the initial stage of the project, we carried out an extensive analysis of the state-of-the-art of various topics relevant to PD management, including signal processing methods, studies for the monitoring, detection and evaluation of motoric symptoms, cognitive assessment tests, research for speech disturbances, PD nutrition and physiotherapy aspects, data mining studies, and decision support systems [16]. In this paper, we focus on decision support systems and present the findings from this perspective. After explaining the concept of DSS, we review the trends and main accomplishments in the area of clinical DSSs, including those addressing PD. On this basis, we propose the methodological approach to the development of the PD_manager DSS.

2. DECISION SUPPORT SYSTEMS

Decision Support Systems (DSSs) are interactive computer-based systems intended to help decision makers utilize data and models to identify and solve problems and make decisions [23, 29]. Their main characteristics are:

- DSSs incorporate both data and models;
- they are designed to assist decision-makers in their decision processes in semi-structured or unstructured tasks;
- they support, rather than replace, managerial judgment;
- their objective is to improve the quality and effectiveness (rather than efficiency) of decisions.

DSSs used in medicine are often referred to as Clinical DSS (CDSS). They are aimed at providing clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care [3].

Traditionally, DSSs are categorized according to the prevailing aim, functionality and employed approach [23]:

- *Communication-driven DSS*: aimed at supporting user collaboration, typically employing a web or client server.
- *Data-driven DSS*: these rely on databases to provide the desired decision-support information and facilitate seeking specific answers for specific purposes; typical technologies used include databases, data warehouses, query systems and on-line analytical processing methods.
- *Document-driven DSS*: their purpose is to store documents, which can be accessed through a set of keywords or search terms; the functionality may include advanced semantic and language processing tools.
- *Knowledge-driven DSS*: such DSSs store and apply knowledge for a variety of decision problems, including classification and configuration tasks, risk management and application of policies; the approach often relies on artificial intelligence and statistical technologies.
- *Model-driven DSS*: such DSSs are complex systems that help analyze decisions or choose between decision alternatives; they are characterized by employing different kinds of

quantitative or qualitative models, such as algebraic, financial, optimization, simulation and evaluation models.

3. STATE OF THE ART

From the historical perspective, the concept of DSS is fairly old, mature and well-developed. First DSSs can be traced back to 1960s [23], and first large-scale CDSS were developed in 1970s (for instance, MYCIN [31] and INTERNIST-I [21], to mention just two). Nonetheless, the area of DSS has been very prolific since then, and still is. The concepts and, particularly, technologies and architecture of DSS evolved dramatically in the attempts to follow rapid technological change and to satisfy ever increasing decision-makers' needs for data, information and knowledge. The 1980s witnessed the specialization of DSSs into Management Information Systems (MIS), Executive IS (EIS), Expert Systems (ES) and many others. The emphasis in 1990s was on data warehousing, data mining and on-line analytical processing (OLAP). The new millennium brought more attention to web-based DSS and business intelligence. According to Power [24], the attributes of contemporary analytical and decision support systems typically include the following:

1. Access capabilities from any location at anytime.
2. Access very large historical data sets almost instantaneously.
3. Collaborate with multiple, remote users in real-time using rich media.
4. Receive real-time structured and unstructured data when needed.
5. View data and results visually with graphs and charts.

We carried out a literature review about recent CDSS, particularly those addressing PD. The literature indicates lots of activities in DSS development, particularly since 2010. The focus is shifting towards Mobile Health applications, which can be characterized with the following three dimensions [9]:

- *Domains*: wellness and prevention, diagnosis, treatment and monitoring, stronger health-care systems.
- *Technology*: applications, sensors, devices.
- *Target groups*: healthy people, hospital patients, chronically ill patients.

On these grounds, DSS functionality and architectures are facing major transformations, most notably:

- from centralized to distributed and mobile architectures,
- from traditional databases to cloud computing,
- from medical institutions to patients' homes,
- from supporting a relatively small number of expert physicians to providing service to many individual patients,
- from providing general answers and solutions towards more personalized advice.

The prevalent state-of-the-art DSS architecture involves multiple components, which are combined and integrated in a variety of ways:

- *patient data*, where traditional clinical databases are combined with real-time data, obtained by telemonitoring of the patient;
- *models* – in the sense of knowledge-driven and model-driven DSS – that propose solutions to various aspects of the decision problem;
- *communication-driven network infrastructure*, which supports the exchange of data and information between patients and medical workers (in both directions);
- *user modules* that convey information to the DSS user and facilitate the exploration of solutions.

In the literature, there are many examples of DSSs that follow this architecture, but they are mainly addressing other diseases than Parkinson's. For example, there are DSSs for cancer recurrence prediction [13], heart failure diagnosis and treatment [33, 28], and management of chronic disease [2]. For PD, DSSs are still more at the level of exploring various modelling and data mining approaches, and creating DSS prototypes. Several DSSs for PD diagnosis were built around the UCI PD dataset [11, 12, 17]. A web-based approach was used to design the DSS for selecting PD patients for deep brain simulation [34]. Notable examples of recent efforts include decision support based on data mining for PD diagnosis and therapy [14] and a series of systems for monitoring and diagnosing PD patients developed at Dalama University [20, 35,19].

Considering the types of models used in CDSSs, the situation is extremely diverse. In their study of decision-analytic models used in relation with PD, Shearer et al. [30] identified 18 model-based evaluations of interventions in PD. Among the 18 models, 14 used Markov modelling, 3 decision trees and one a simulation model. In our view, this prevalence of Markov modelling is somewhat surprising, as much more variety is indicated in other literature. Particularly abundant are approaches based on data mining and machine learning, which involve methods such as decision trees, decision rules, artificial neural network models, support vector machines, and Bayesian models (see [16] for review). Another important branch is based on expert modelling, i.e., involving experts in the creation decision models. A variety of expert-system, knowledge-based and rule-based approaches are used here, such as fuzzy rule-based modelling [1], fuzzy cognitive maps [18], ontologies [25], and methods based on experts' feedback [27]. Other notable approaches include multi-criteria decision analysis [8] and semantic technologies [26].

Considering user modules and providing DSS services to decision makers, many of the reviewed DSSs seem rather weak. Namely, there are many DSSs whose development had only reached the stage of constructing a decision model, which was only verified on some data set, without considering the end user at all. In our view, a fully developed DSS should duly consider its user: the physician, medical staff and patients. The DSS should provide information that is considered useful by the users and helps them in their decision-making process. Furthermore, a good DSS should also provide methods and tools that facilitate an active user-initiated exploration of relevant information, possible solutions and expected consequences of decisions.

4. PD_MANAGER APPROACH TO DSS

The PD_manager's DSS will be aimed at supporting the physician and other caregivers in their monitoring of PD patients and deciding about their therapies. The primary emphasis will be on providing suggestions for adjusting the medication plan for the patient. Later, the decision support for nutrition, exercise and physiotherapy will be gradually added. Two decision-support functionalities will be implemented:

- Providing the relevant information about the patient to the physician, who actually makes the decision (e.g., action,

therapy prescription). Here, the emphasis is on contents, clarity, and form of information presented to the user. The system should present the information to the user in a compact, visual and easy to comprehend way.

- Proposing diagnostic and therapeutic solutions to the users. Here, the emphasis is on models that transform input information about the patient to decisions. Models, together with appropriate algorithms, also provide mechanisms to explore, explain and justify proposed solutions.

The PD_manager DSS architecture will be based on a combination of communication-, data- and model-driven approaches. The main DSS development activities will be carried out through:

1. reviewing and collecting existing PD-related models, and developing new ones through data mining,
2. selecting, adapting, integrating and developing models for the desired DSS functionality, and
3. implementing the DSS with special emphasis on user modules and user-oriented functionality.

The expected result of the first stage will be a set of models, potentially useful for the PD_manager DSS. The set will include models from the literature and other related projects [e.g., 11, 12, 14, 17, 19, 20, 35], and the models developed in PD_manager from patient-collected data. In the second stage, we will review those models and decide about their potential inclusion in the DSS according to the following criteria:

- *Operability*: are the necessary conditions for using the model, such as data availability and quality, satisfied?
- *Fitness for purpose*: does the model provide answers required for the addressed decision support task?
- *Accuracy*: are the results, provided by the model, good enough?
- *Transparency*: are the model and its results easy to understand, comprehensible to the user and sufficiently easy to visualize and explain?
- *Flexibility*: to which extent is it possible and how difficult is it to personalize the model to individual patients?

We expect that the selected models will only partly satisfy the needs of the DSS. In our previous DSS-development projects SIGMEA [5] and Co-Extra [7], we found out that data mining models typically covered just a part of the problem space, while the remaining gaps had to be completed through expert modelling techniques. Thus, we expect that it is going to be necessary to supplement the data-mining models with expert-developed ones. The main method used for this purpose will be qualitative modelling method DEX [6], which will be, if necessary, combined with other methods, such as model revision [36], and fuzzy rule-based modelling [1, 18, 33].

DEX (Decision EXpert) is a qualitative multi-criteria modelling approach, aimed at the assessment and analysis of decision alternatives. DEX belongs to a wider class of multi-attribute (or multi-criteria) models [15]. DEX models have a hierarchical structure, which represents a decomposition of some decision problem into smaller, less complex sub-problems. DEX differs from most conventional multi-attribute decision modeling tools in that it uses qualitative (symbolic) attributes instead of quantitative (numeric) ones. Also, aggregation (utility) functions in DEXi are defined by *if-then* decision rules rather numerically by weights or some other kind of numerical value function.

DEX has already been successfully used in health care [4, 32] and is considered suitable for PD_manager because of the transparency of its models. DEX is supported by software DEXi (http://kt.ijs.si/MarkoBohanec/dexi.html), which is available for free and provides methods for acquiring expert knowledge, maintaining the consistency and completeness of models, and carrying out exploratory analysis of decision alternatives and their consequences. Once developed and properly verified, DEX models are guaranteed to be complete (providing answers for all possible input data combinations) and consistent (results obey the principle of dominance, so that the model's overall value function is monotone).

5. CONCLUSION

In this paper, we reviewed the state-of-the-art of DSS for PD management. The findings clearly indicate that the area of clinical DSS is very vivid and evolves rapidly. The DSS architectures are facing major transitions toward real-time, distributed and mobile architectures, based on telemonitoring and supporting a variety of users, including patients at their homes. While there are already several existing CDSSs that employ these architectures for various other diseases, the support for PD is still scarce and insufficient in providing substantial assistance to the end users – clinicians and patients. We expect this will be one of the main contributions of the forthcoming PD_manager DSS.

The development of PD_manager DSS is in progress, with the prototype expected at the end of 2016. To date, we have designed the system's architecture, which will combine the principles of communication-, data- and model-driven approaches. The main decision modelling method will be DEX, which was chosen because of transparency of its hierarchical rule-based models, positive experience with previous applications in health care, accessibility of the supporting software, and good support for ensuring the completeness and consistency of models.

In further work, we expect two major difficulties: (1) incomplete coverage of the decision-problem space by the models developed by data mining, and (2) the need to personalize the DSS models to characteristics of individual patients. We intend to alleviate them by combining data mining with expert-modelling and model-revision techniques.

6. ACKNOWLEDGMENT

The work of the author was supported by the PD_manager project, funded within the EU Framework Programme for Research and Innovation Horizon 2020, under grant number 643706.

7. REFERENCES

[1] Ahmed MU, Westin J, Nyholm D, Dougherty M, Groth T. A fuzzy rule-based decision support system for Duodopa treatment in Parkinson. In: P. Eklund HL, M. Minock, editor. SAIS 2006, p. 45–50.

[2] Bellos C, Papadopoulos A, Rosso R, Fotiadis DI. Heterogeneous data fusion and intelligent techniques embedded in a mobile application for real-time chronic disease management. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE; 2011.

[3] Berner ES. Clinical Decision Support Systems: State of the Art. Rockville: Agency for Healthcare Research and Quality, AHRQ Publication No. 09-0069-EF; 2009.
[4] Bohanec M, Zupan B, Rajkovič V. Applications of qualitative multi-attribute decision models in health care, International Journal of Medical Informatics 2000; 58-59:191-205.
[5] Bohanec M, Messéan A, Scatasta S, Angevin F, Griffiths B, Krogh PH, et al. A qualitative multi-attribute model for economic and ecological assessment of genetically modified crops. Ecological Modelling. 2008 Jul;215(1-3):247–61.
[6] Bohanec M, Žnidaršič M, Rajkovič V, Bratko I, Zupan B. DEX Methodology: Three Decades of Qualitative Multi-Attribute Modeling. Informatica 2013; 37(1).
[7] Bohanec M, Bertheau Y, Brera C, Gruden K, Holst-Jensen A, Kok EJ, et al. The Co-Extra decision support system: A model-based integration of project results. Genetically Modified and Non-Genetically Modified Food Supply Chains: Co-Existence and Traceability. Wiley-Blackwell, 2013;459–89.
[8] Cunningham C. Development of an electronic treatment decision aid for Parkinson's disease using mutli-criteria decision analysis. Ph.D. Thesis. Cardiff University; 2008.
[9] Dehzad F, Hilhorst C, de Bie C,Claassen E. Adopting Health Apps, What's Hindering Doctors and Patients? Health 2014 6:2204-17.
[10] European Parkinson's Disease Association (EPDA). *The European Parkinson's Disease Standards of Care Consensus Statement, Volume I*. 2011.
[11] Ericsson A, Lonsdale MN, Astrom K, Edenbrandt L, Friberg L. Decision support system for the diagnosis of Parkinson's disease. Image Analysis. Springer; 2005, 740–9.
[12] Eskidere Ö, Ertaş F, Hanilçi C. A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Systems with Applications. 2012 Apr;39(5):5523–8.
[13] Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence. IEEE Transactions on Information Technology in Biomedicine. 2012 Nov;16(6):1127–34.
[14] Exarchos TP, Tzallas AT, Baga D, Chaloglou D, Fotiadis DI, et al. Using partial decision trees to predict Parkinson's symptoms: A new approach for diagnosis and therapy in patients suffering from Parkinson's disease. Computers in Biology and Medicine. 2012 Feb;42(2):195–204.
[15] Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys. New York: Springer; 2005. 1045 p.
[16] Gatsios, D, et al. *State of the art*. PD_manager project, Deliverable D3.6, 2015,
[17] Gil D, Johnson JB. Diagnosing parkinson by using artificial neural networks and support vector machines. Global Journal of Computer Science and Technology; 2009.
[18] Groumpos PP, Anninou AP. A theoretical mathematical modeling of Parkinson's disease using Fuzzy Cognitive Maps. Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference. IEEE; 2012.
[19] Khan T, Memedi M, Song W and Westin J. A case study in Healthcare Informatics: a telemedicine framework for automated Parkinson's disease symptom assessment. In Proceedings of the International Conference for Smart Health (ICSH2014), Beijing, China, July, 2014.
[20] Memedi M, Westin J, Nyholm D, Dougherty M and Groth T. A web application for follow-up of results from a mobile device test battery for Parkinson's disease patients. Computer Methods and Programs in Biomedicine 2011; 104: 219-226.
[21] Miller RA, Pople HE Jr, Myers JD. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982 Aug 19;307(8):468-76.
[22] *PD_manager: m-Health platform for Parkinson's disease management*. EU Framework Programme for Research and Innovation Horizon 2020, Grant number 643706, 2015–2017, http://www.parkinson-manager.eu/
[23] Power DJ. Decision support systems: concepts and resources for managers. Greenwood Publishing Group; 2002.
[24] Power DJ. Decision support, analytics, and business intelligence. New York: Business Expert Press; 2013.
[25] Riaño D, Real F, López-Vallverdú JA, Campana F, Ercolani S, Mecocci P, et al. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. Journal of Biomedical Informatics. 2012 Jun;45(3):429–46.
[26] Rodríguez-González A, Alor-Hernández G. An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. Computers in Biology and Medicine. 2013 Jan;43(1):51–62.
[27] Rodríguez-González A, Torres-Niño J, Valencia-Garcia R, Mayer MA, Alor-Hernandez G. Using experts feedback in clinical case resolution and arbitration as accuracy diagnosis methodology. Computers in Biology and Medicine. 2013 Sep;43(8):975–86.
[28] Sonawane JS. Survey on Decision Support System For Heart Disease. International Journal of Advancements in Technology. 2013;4(1):89–96.
[29] Sharda R, Delen D, Turban E, Aronson J, Liang TP. Business intelligence and analytics: Systems for decision support. Pearson Education; 2014.
[30] Shearer J, Green C, Counsell CE, Zajicek JP. The use of decision-analytic models in Parkinson's disease. Applied health economics and health policy. 2011;9(4):243–58.
[31] Shortliffe EH. Computer-Based Medical Consultations: MYCIN. New York: Elsevier/North Holland; 1976.
[32] Šušteršič, O, Rajkovič, U, Dinevski, D, Jereb, E, Rajkovič, V. Evaluating patients' health using a hierarchical multi-attribute decision model. Journal of international medical research 2009; 37(5):1646-1654.
[33] Tsipouras MG, Karvounis EC, Tzallas AT, Goletsis Y, Fotiadis DI, Adamopoulos S, Trivella MG. Automated knowledge-based fuzzy models generation for weaning of patients receiving Ventricular Assist Device (VAD) therapy. Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE; 2012.
[34] Vitek M, Pinter M, Rappelsberger A, Hayashi Y, Adlassnig K-P. Web-Based Decision Support in Selecting Patients with Parkinson's Disease for Deep Brain Stimulation. CBMS'07 Computer-Based Medical Systems, 2007, p. 79–84.
[35] Westin J, Schiavella M, Memedi M, Nyholm D, Dougherty M and Antonini A. Validation of a home environment test battery for supporting assessments in advanced Parkinson's disease. Neurological Sciences 2012; 33: 831-838.
[36] Žnidaršič M, Bohanec M. Automatic revision of qualitative multi-attribute decision models. Foundations of Computing and Decision Sciences 2007; 32(4):315-326.

Prepoznavanje in napovedovanje hiperglikemij in hipoglikemij na neinvaziven način

Božidara Cvetković
Institut "Jožef Stefan", Odsek za inteligentne sisteme
Jamova cesta 39
1000 Ljubljana, Slovenija
+386 1 4773498
boza.cvetkovic@ijs.si

Urška Pangerc
Institut "Jožef Stefan", Odsek za inteligentne sisteme
Jamova cesta 39
1000 Ljubljana, Slovenija

Mitja Luštrek
Institut "Jožef Stefan", Odsek za inteligentne sisteme
Jamova cesta 39
1000 Ljubljana, Slovenija
+386 1 4773380
mitja.lustrek@ijs.si

POVZETEK

V tem prispevku je predstavljen pristop strojnega učenja za razpoznavanje in napovedovanje abnormalnih stanj glukoze (hiperglikemija, hipoglikemija) pri bolnikih z diabetesom tipa I in tipa II. Algoritme strojnega učenja smo uporabili na podatkih prsnega merilnika elektrokardiogramskega signala, ki poleg delovanja srca spremlja tudi dihanje. Signal smo obdelali z algoritmom, ki izlušči pomembne parametre, katere smo uporabili pri grajenju modelov za klasifikacijo. Dosegli smo 84% točnost pri napovedovanju glikemij v primeru bolnikov z diabetesom tipa I in 88,5% pri bolnikih z diabetesom tipa II. Pri prepoznavanju smo dosegli 78,05% točnost v primeru bolnikov z diabetesom tipa I in 75,81% pri bolnikih z diabetesom tipa II.

Ključne besede

EKG, glukoza, hiperglikemija, hipoglikemija, napovedovanje, razpoznavanje.

1. UVOD

Diabetes je kronična bolezen, ki se pojavi bodisi kadar trebušna slinavka ne proizvaja zadosti inzulina ali kadar telo ne more učinkovito izrabiti proizvedenega inzulina. Diabetes tipa 1 je avtoimunska bolezen, v kateri imunski sistem uniči celice v trebušni slinavki, ki proizvajajo inzulin, diabetes tipa 2 pa se največkrat razvije zaradi odpornosti celic na inzulin, k čemur precej prispeva nezdrav način življenja. Po pričanju svetovne zdravstvene organizacije je število obolelih za diabetesom v letu 2014 preseglo 9% odrasle svetovne populacije [1] in še narašča.

Bolniki z diabetesom so primorani celo življenje spremljati in vzdrževati nivo glukoze na ustrezni ravni, saj je lahko povečanje ali zmanjšanje glukoze vzrok za vrsto bolezni, med katerimi je najbolj izpostavljeno tveganje za obolenje srca in ožilja. Svetovna zdravstvena organizacija trdi, da je vzrok smrti vsakega drugega bolnika z diabetesom povezan z odpovedjo srca ali možgansko kapjo.

Za zmanjšanje tveganja smrti je pomembno poleg ravni glukoze spremljati tudi delovanje srca. V medicinski literaturi lahko najdemo, da je hipoglikemija povezana z zmanjšanjem srčnega utripa in da je hiperglikemija močno povezana s polarizacijo in depolarizacijo srčnega prekata, ki ga opisuje tako imenovani interval QT, ki ga lahko pridobimo iz signala elektrokardiograma (EKG). Sprememba intervala QT je prav tako povezana z aritmijo, ki poveča tveganje za nenadno smrt.

Dostopnost komercialnih naprav za zvezno spremljanje signala EKG lahko bistveno olajšajo prepoznavanje in napovedovanje tako anomalij pri delovanju srca, kot tudi anomalij zaradi spremembe ravni glukoze.

V tem prispevku bomo predstavili pristop za razpoznavanje in napovedovanje hipoglikemij in hiperglikemij na neinvaziven način in sicer z uporabo komercialnega prsnega merilnika EKG [2], ki poleg delovanja srca meri tudi dihanje.

2. SORODNO DELO

Večina raziskav se ukvarja z iskanjem relacij med hipoglikemijo ter boleznimi srca in ožilja in s tem povezano umrljivostjo. Raziskave temeljijo na povezavi med parametri, izračunanimi iz signala EKG, ter izmerjeno ravnjo glukoze. Raziskave na tem področju uporabljajo signale iz kliničnega merilnika EKG in rezultati temeljijo na meritvah, opravljenih pod nadzorovanimi kliničnimi pogoji.

Hanefeld et. al. [1] sistematično predstavijo dosežke na tem področju. V raziskavah zveznega spremljanja glukoze pri bolnikih z diabetesom tipa I in II so ugotovili, da pride do spremembe dolžine intervala QT v primeru hude hipoglikemije (tudi nočne), kar sproži anomalijo srčnega ritma in lahko doprinese k višjemu tveganju za smrt. Potrditev teze podaljšanja intervala QT v primeru hipoglikemije lahko najdemo tudi v drugih neodvisnih raziskavah [4][5].

Povezava med hiperglikemijo in parametri srčnega signala so raziskovali Singh et al. [6]. Ugotovili so, da se variabilnost srčnega utripa zmanjša v primeru hude hiperglikemije.

Detekcijo hiperglikemije in hipoglikemije iz preostalih parametrov, izluščenih iz signala EKG (pri bolnikih z diabetesom tipa I), so raziskovali Nguyen et al. [7]. Ugotovili so, da je hitrost srčnega utripa (narašča) povezana izključno s hipoglikemijo in sprememba intervala PR, ki ga dobimo iz signala RKG, izključno s hiperglikemijo.

Raziskave, kjer bi uporabili metode strojnega učenja pri napovedovanju ali odkrivanju stanja glukoze, zaenkrat ne vsebujejo informacij, pridobljenih iz signala EKG, ampak napovedujejo nivo glukoze glede na kompleksne dinamične modele. priučene iz zgodovinskih podatkov o ravni glukoze pri določenem bolniku [8].

3. ZBIRKA PODATKOV IN PRISTOP

3.1 Zbirka podatkov

Zbiranje podatkov je potekalo šest tednov na dveh lokacijah, na Poljskem in v Italiji. Na Poljskem smo zbrali podatke 30 bolnikov z diabetesom tipa II in v Italiji 22 bolnikov z diabetesom tipa I. Vsi bolniki so bili opremljeni s prsnim merilnikom, ki zajema signal EKG, dihanja in pospeška, z merilnikom glukoze, z merilnikom krvnega tlaka, s tehtnico ter s pametnim telefonom, preko katerega so se podatki pošiljali v centralno bazo.

V trenutni raziskavi smo uporabili signal EKG, signal dihanja in podatke o izmerjenem nivoju glukoze. Signal EKG in signal dihanja smo najprej obdelali s filtri, ki so odstranili neberljive in šumne dele signalov in hkrati ohranili njihovo morfologijo. Očiščene dele signala smo razdelili na 30-minutne segmente, ki so sovpadali z meritvijo glukoze, in sicer:

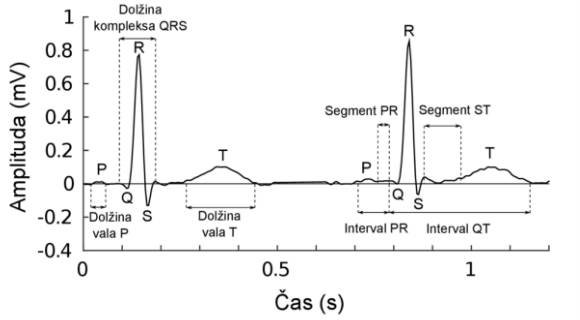
- od 45 do 15 minut pred meritvijo za potrebe napovedovanja nivoja glukoze (hiperglikemija, hipoglikemija, normalna glikemija)
- 15 minut pred in po meritvi za potrebe prepoznavanja nivoja glukoze (hiperglikemija, hipoglikemija, normalna glikemija)

Signal EKG smo obdelali z algoritmom, ki iz signala izloči 13 parametrov, ki opisujejo obliko srčnega utripa. Parametri signala so predstavljeni na sliki 1 in so sledeči:

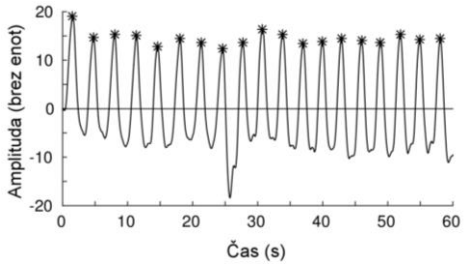
- Segment PR - čas med koncem vala P in začetkom kompleksa QRS
- Interval PR - čas med začetkom vala P in začetkom kompleksa QRS
- Interval QS - čas med začetkom in koncem kompleksa QRS
- Segment ST - čas med koncem kompleksa QRS in začetkom vala T
- Interval QT - čas med začetkom kompleksa QRS in koncem vala T
- Dolžina vala P - čas med začetkom in koncem vala P
- Dolžina vala T - čas med začetkom in koncem vala T
- Vrednost Q - amplituda vala Q
- Vrednost R - amplituda vala R
- Vrednost S - amplituda vala S
- Vrednost P - amplituda vala P
- Vrednost T - amplituda vala T
- Interval RR - čas med dvema zaporednima valoma R

Za vsak parameter smo v 30-minutnih segmentih izračunali srednjo vrednost, standardno deviacijo in trend (naklon linearne aproksimacije).

Signal dihanja smo prav tako razdelili na 30-minutne segmente, ki sovpadajo s časom meritve glukoze. Iz signala smo izluščili število vdihov na minuto ter izračunali standardno deviacijo in trend. Primer signala dihanja z označenimi razpoznanimi vdihi je predstavljen na sliki 2.



Slika 1. Parametri, pridobljeni z obdelavo signala EKG.



Slika 2. Signal dihanja z označenimi razpoznanimi vdihi.

Meritev glukoze vsebuje poleg same vrednosti nivoja glukoze tudi informacijo o tem, ali je bila izmerjena pred jedjo, po jedi, pred spanjem, ponoči ali ostalo. Glede na vrednost nivoja glukoze smo meritve razdelili v tri skupine:

1. Hipoglikemija – glukoza < 4 mmol/l
2. Hiperglikemija – glukoza > 7 mmol/l
3. Normalno stanje – 4 mmol/l < glukoza > 7 mmol/l

3.2 Pristop za napovedovanje in prepoznavanje glikemij

Za napovedovanje in prepoznavanje glikemij smo uporabili klasičen pristop strojnega učenja, kjer instanco sestavljajo izračunani in izbrani parametri enega ali več signalov. Tako sestavljanje instance obdelamo z algoritmom strojnega učenja za reševanje problema klasifikacije in točnost ovrednotimo z 10-kratnim prečnim preverjanjem.

Pristop smo zastavili na dva načina:

1) Začnemo z najbolj relevantnim signalom (EKG) in nato dodamo še dodaten vir informacij kot je dihanje ter čas v dnevu (pred, po jedi, pred spanjem, ponoči in ostalo).

2) Čas v dnevu uporabimo za razdelitev prostora napovedovanja na dva dela. Izdelamo dva modela: enega, ki napoveduje glikemije po jedi in enega pred jedjo.

Absolutne vrednosti izračunanih atributov (srednje vrednosti, standardne deviacije in trende trinajstih atributov iz signala EKG ter število vdihov, standardno deviacijo in trend dihanja) smo preračunali tudi v relativne vrednosti glede na posameznega bolnika. Sestavili smo sledeče štiri množice atributov:

A1. Vsi atributi (absolutne in relativne vrednosti)

A2. Absolutne vrednosti atributov

A3. Relativne vrednosti atributov

A4. Najboljših 20 atributov glede na algoritem ReliefF [9]

Množica atributov predstavlja instanco označeno s trenutnim stanjem glukoze (hiperglikemija, hipoglikemija, normalno stanje).

4. EKSPERIMENTI IN REZULTATI

Skupno smo izvedli 18 eksperimentov za napovedovanje glikemije in 18 eksperimentov za razpoznavanje glikemije. Za vsako množico atributov iz prejšnje sekcije smo naredili eksperiment z naslednjimi modeli:

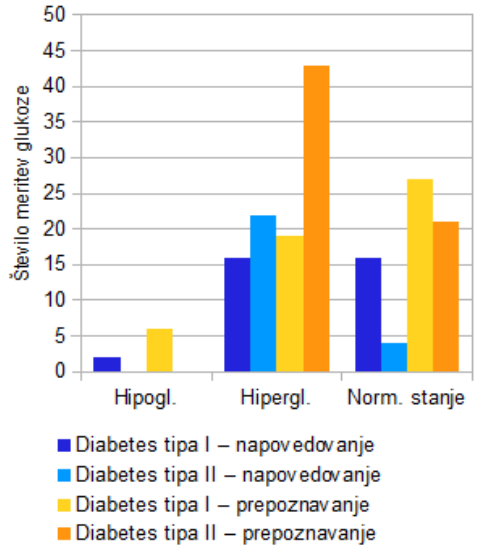
S1. Model, zgrajen z atributi signala EKG,

S2. Model, zgrajen z atributi signala EKG in signala dihanja,

S3. Model zgrajen z atributi signala EKG, signala dihanja ter atributom »čas meritve glukoze« (pred jedjo, po jedi, pred spanjem, ponoči ali ostalo)

S4. Po atributu »čas meritve glukoze« smo razdelili množico na dva dela. Modele smo zgradili z atributi iz točke S2.

Slika 3 predstavlja porazdelitev razredov v zbirki podatkov, kjer lahko vidimo, da imamo največ primerkov hiperglikemije in le nekaj primerkov hipoglikemije v primeru bolnikov z diabetesom tipa I. Bolniki z diabetesom tipa II niso imeli primerov hipoglikemije.



Slika 3. Porazdelitev primerkov glikemij glede na tip diabetesa in glede na tip klasifikacijskega problema.

V eksperimentu pod točko 4 smo instance delili glede na parameter »čas meritve glukoze«, ki imajo vrednost pred jedjo ali po jedi. Porazdelitev po vrednosti je prikazana v tabeli 1 Zaradi premajhnega števila instanc napovedovanja stanja glukoze v krvi nismo testirali v naslednjih skupinah: bolniki z diabetesom tipa I po jedi ter bolniki z diabetesom tipa II pred jedjo in po jedi.

Vsako množico atributov smo testirali z desetimi klasifikacijskimi algoritmi strojnega učenja, ki so na voljo v paketu Weka [9]: naivni Bayes, logistična regresija, SVM, IB3, AdaBoostM1 z RepTree, Bagging z RepTree, JRip, J48, Random Forest in ZeroR kot izhodiščni algoritem, ki vedno vrne večinski razred. Vsak eksperiment je bil ocenjen in preverjen z uporabo 10-kratnega prečnega preverjanja.

Tabela 1. Število instanc glede na stanje glukoze v krvi in glede na čas meritve glukoze (pred jedjo ali po jedi).

Stanje	Napovedovanje				Prepoznavanje			
	Diab. I		Diab. II		Diab. I		Diab. II	
	Pr.	Po	Pr.	Po	Pr.	Po	Pr.	Po
Hipogl.	1	-	-	-	4	2	-	-
Hipergl.	10	-	-	-	8	7	9	32
Normalno	8	-	-	-	12	8	6	15

Rezultati eksperimentov so prikazani v tabeli 2 in tabeli 3, pri čemer tabela 2 vsebuje rezultate napovedovanja glikemij, tabela 3 pa rezultate prepoznavanja glikemij. Prva vrstica označuje tip eksperimenta (S1, S2, S3 ali S4), levi del tabele predstavlja rezultate pri bolnikih z diabetesom tipa I in desni del tabele rezultate pri bolnikih z diabetesom tipa II. Prvi stolpec obeh razdelkov vsebuje oznako množice uporabljenih atributov (Sekcija 3.2), ki je vrnila najboljši rezultat klasifikacije.

Pri napovedovanju glikemij se je izkazala množica A4 (izbranih 20 atributov glede na algoritem Relieff) za najboljšo v primeru diabetesa tipa I in množica A2 (absolutne vrednosti vseh atributov) za primer diabetesa tipa II. Rezultat z najvišjo točnostjo 84,21% pri bolnikih z diabetesom tipa I dobimo z logistično regresijo in z delitvijo množice glede na čas meritve glukoze. V primeru bolnikov z diabetesom tipa II je najboljši rezultat pridobljen z algoritmom IB3 in sicer 88,46%. Domnevamo, da so se bolniki tipa II merili predvsem ob slabem počutju, saj meritve niso izvajali po jedi ali pred jedjo ampak različno skozi cel dan, zato rezultatov za eksperiment S4 z dodanim atributom čas meritve glukoze ni bilo mogoče izvesti.

Tabela 2. Rezultati napovedovanja hipoglikemij in hiperglikemij v primerjavi z izhodiščnim rezultatom.

	Diabetes tipa I			Diabetes tipa II		
	A	ZeroR (%)	Točnost (%)	A	ZeroR (%)	Točnost (%)
S1	A4	41.18	61.76	A2	84.62	88.46
S2	A4	41.18	64.71	A2	84.62	88.46
S3	A4	41.18	79.41	A2	84.62	**88.46**
S4	A4	52.63	**84.21**	-	-	-

Tabela 3. Rezultati prepoznavanja hipoglikemij in hiperglikemij v primerjavi z izhodiščnim rezultatom.

	Diabetes tipa I			Diabetes tipa II		
	A	ZeroR (%)	Točnost (%)	A	ZeroR (%)	Točnost (%)
S1	A3	51.92	63.46	A2	67.19	70.31
S2	A4	51.92	69.23	A2	67.19	71.88
S3	A4	51.92	74.15	A2	67.19	73.44
S4	A4	48.78	**78.05**	*	66.13	**75.81**

* A4 za model pred jedjo in A1 za model po jedi

Pri prepoznavanju glikemij (

Tabela 3), se je delitev po času meritve glukoze izkazal za najboljši pristop pri obeh tipih diabetesa. Za diabetes tipa I je

uporaba atributov iz množice A4 in algoritem SVM za model pred jedjo in logistična regresija za model po jedi vrnila najboljši rezultat in sicer 78,05%. Za diabetes tipa II je najboljši rezultat vrnila uporaba atributov iz množice A4 in algoritem SVM za model pred jedjo in atributov iz množice A2 in algoritem Bagging za model po jedi, in sicer 75,81%.

5. ZAKLJUČEK

V prispevku smo predstavili pristop za napovedovanje in razpoznavanje anomalij (hiperglikemija, hipoglikemija) pri bolnikih z diabetesom tipa I in tipa II. Uporabili smo splošni pristop strojnega učenja, kjer smo iz atributov, pridobljenih iz signala EKG in iz signala dihanja, zgradili modele za reševanje obeh klasifikacijskih problemov.

Eksperimente smo izvedli na podatkih 30 bolnikov z diabetesom tipa I in 22 bolnikov diabetesa tipa II. Ugotovili smo, da je najboljši pristop tako pri napovedovanju kot pri razpoznavanju gradnja dveh modelov, enega za napovedovanje ali razpoznavanje glikemij pred jedjo in drugega za napovedovanje ali razpoznavanje po jedi.

Z omenjenim pristopom smo dosegli 84,21% točnost pri napovedovanju glikemij pri bolnikih z diabetesom tipa I. Enakega postopka nismo mogli uporabiti pri bolnikih tipa II, saj je bilo premalo podatkov, označenih s časom merjenja glukoze. Z uporabo tako pomanjkljivih podatkov smo dosegli točnost 88,46%. Z enakim pristopom smo pri razpoznavanju dosegli točnosti 78,05% pri bolnikih s diabetesom tipa I in 75,81% pri bolnikih tipa II. Glede na to, da se razpoznavanje zdi lažja naloga od napovedovanja, je ta rezultat nekoliko presenetljiv in si ga bomo v prihodnje prizadevali pojasniti.

V nadaljnjem delu bomo tudi ocenili, ali je razvita metoda primerna za praktično prepoznavanje in napovedovanje glikemij, ki bi ga lahko uporabili, da bi bolniku svetovali, naj si izmeri raven glukoze v krvi z običajnim invazivnim merilnikom. Poleg tega bomo ugotovili signifikantnosti posameznih parametrov in se ukvarjali z nepravilnostmi delovanja srca.

6. ZAHVALA

Raziskava je bila delno financirana iz evropskega projekta COMMODITY12 (www.commodity12.eu).

7. REFERENCE

[1] WHO, http://www.who.int/mediacentre/factsheets/fs312/en/

[2] Zephyr BioHarness, http://www.zephyranywhere.com/

[3] Hanefeld, M., Duetting, E., Bramlage, P. 2013. Cardiac implications of hypoglycaemia in patients with diabetes – a systematic review. *Cardiovascular Diabetology*, 135, 12 (Sep. 2013). DOI=10.1186/1475-2840-12-135.

[4] Frier, M. B., Schernthaner, G., Simon R. Heller, R. S. 2011. Hypoglycemia and Cardiovascular Risks. *Diabetes Care*, 34 (May 2011). 132-137. DOI=10.2337/dc11-s220.

[5] Snell-Bergeon, J. K., Wadwa, R. P. 2012. Hypoglycemia, Diabetes, and Cardiovascular Disease. *Diabetes Technology & Therapeutics*, 14, 51–58. DOI= http://doi.org/10.1089/dia.2012.0031

[6] Singh, P.J., Larson, G. M., O'Donnell, J. C., Wilson F. P., Tsuji, H., Lloyd-Jones, M. D., Levy, D. 2000. Association of hyperglycemia with reduced heart rate variability (The Framingham Heart Study). The American Journal of Cardiology, 86, 3, (Aug 2000), 309-312, DOI=http://dx.doi.org/10.1016/S0002-9149(00)00920-6.

[7] Nguyen, L. L, Su, S., Nguyen, H.T. 2012. Identification of Hypoglycemia and Hyperglycemia in Type 1 Diabetic patients using ECG parameters, *in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Sand Diego, USA, August 28 – September 1, 2012), EMBC, IEEE, 2716-2719. DOI = 10.1109/EMBC.2012.6346525

[8] Plis,K., Bunescu, R., Marling, C., Shubrook, J., Schwartz, F. 2014. A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management. *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, H. I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 1, 2009.

Learning From Microarray Gene Expression Data

Aleksandar Dimitriev
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
ad7414@student.uni-lj.si

Zoran Bosnić
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
zoran.bosnic@fri.uni-lj.si

ABSTRACT

Gene expression microarrays are an ever-more abundant source of information for patients and doctors. Today, there are thousands of data sets containing tens of thousands of features, or gene expression levels, each. Their format is suitable for machine learning and data mining, but care must be taken to avoid the pitfalls of the extremely high features-to-samples ratio. Some algorithms are also more suited than others to extract information from this high-dimensional data. We present an overview of supervised methods that are being applied to microarrays, as well as feature-selection methods, which is a vital step in the pre-processing of these data sets. We compare a number of feature selectors, as well as supervised classification algorithms. We found no statistically significant difference between feature selection techniques, but among the classifiers, random forests outperformed the others, which indicate that they might be more suitable for gene expression analysis.

Keywords

Genetic expression, Cancer detection, Classification algorithms, Machine Learning, Bayes methods

1. INTRODUCTION

Gene expression microarrays track gene expression levels across different environmental conditions. A standard use for these microarrays is using them for studying cancerous tissues (or another disease) and control samples, e.g. non-cancerous tissues. After obtaining gene expression levels for each gene and each tissue, the resulting data can be used for both supervised and unsupervised machine learning. The main idea is to use the genes to distinguish between the two (or more) different tissues, or classes. Because microarrays provide expression measurements for tens of thousands of genes, the number of features is orders of magnitude higher, which is opposite to the usual features-to-samples ratio of data sets in the machine learning community. Due to the dimensionality and sparsity of the data, care must be taken to avoid overfitting to noise and ensuring that the many features that are uncorrelated with the dependent variable do not hinder the performance of the algorithms. Many feature selection and dimensionality reduction techniques have been developed to combat this problem, some of which we overview in this paper.

2. BACKGROUND AND RELATED WORK

The central dogma of biology explains the flow of genetic information in the cell, or more specifically, between DNA, RNA and proteins. It states that, most of the time, genes are being transcribed into RNA, and subsequently translated to proteins. Gene expression is then defined as the amount of protein being produced by the gene. To measure this gene expression, the first microarray technology was created by Brown et al. [11]. Microarrays, commonly visualized as heatmaps, shown in Fig. 1, are mainly used to measure and compare gene expression between cancerous and non-cancerous tissue. Other experiments include monitoring the gene expression of a single tissue under different conditions or monitoring the expression of a tissue chronologically (e.g. over the course of a day).

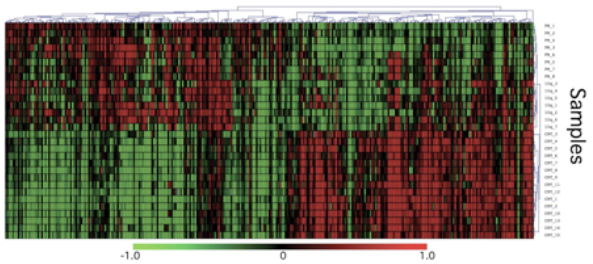


Figure 1: Heatmap of gene expression levels. Green denotes greater expression in the control tissue, whereas red denotes increased expression of that gene in the cancerous tissue.

The literature on supervised microarray analysis is extensive. One of the first experiments was done by Derisi et al. [5], as well as Brown et al. [2], which uses Support Vector Machines (SVMs) to predict gene function. Xing et al. [15] compare feature selection as an alternative to regularization and find the former to be superior, indicating that our focus on feature selection is valid. Another comparison of machine learning algorithms by Pirooznia et. al. [10] also finds that feature selection is an important pre-processing step. Statnikov et al. [13] compare single and ensemble classifiers on cancerous data sets with more than 40 cancer types and find that SVMs outperform k-nearest neighbors (kNN), ANN and ensemble classifiers. A more recent comprehensive review, also by Statnikov et al. [14], compares random forests and SVM on 22 microarray data sets, and find SVMs superior. Diaz et al. [6], however, use random forests for both feature selection and subsequent classification, and find them comparable in accuracy to SVMs and kNN. The most related works are probably by Li et. al [8], who compare

Table 1: Data sets used in the analyses.

Name	Genes	Samples	GEO id	Classes
Lung	54675	58	GDS3627	2
BCP-ALL1	54675	12	GDS4779	2
BCP-ALL2	54675	197	GDS4206	3
Tobacco	24526	183	GDS3929	2

Table 2: Between data set classifiers' adjusted pairwise p-values. Bold indicates statistical significance (p<0.05).

	SVM	NB
NB	1	-
RF	0.103	**0.025**

multiclass classification models and feature selection techniques, as well as Liu et. al. [9], who similarly overview a number of feature selection and classification algorithms.

3. METHODS

We compare multiple feature selection techniques and multiple classifiers on multiple data sets. Because most algorithms can not deal with the thousands of features and very few samples from microarrays, we have chosen to compare Empirical Bayes [12], Student's t-test, and Information Gain as a preprocessing step on the train fold before training. For the classifiers we chose Support Vector Machines [3], Naive Bayes, and Random Forests [1]. The models were compared with the Friedman test and post-hoc pairwise Nemenyi tests with Bonferroni correction, which are the recommended non-parametric multiple hypothesis tests by Demšar [4].

4. RESULTS

We performed our analysis on four different gene expression microarray data sets, available from public repository Gene Expression Omnibus [7]. The first data set contains two types of non-small lung cancer tissues: adenocarcinoma and squamous cell carcinoma. The second and third data set are analyses of B-cell precursor acute lymphoblastic leukemia (BCP-ALL), the second having two classes: short or long time until relapse, and the third having three: early, late, or no relapse. All three data sets have the same genes. The last data set measures the effect of tobacco smoke on maternal and fetal cells and has two classes: smoker and non-smoker, each comprised of three types of tissues. Other information is shown in Table 1. The diverse data sets were chosen to compare the analysis of both cancerous and non-cancerous tissues, binary and multi-class outcomes, different cancers, and different number of samples ranging from 12 to almost 200. In Figure 4 we can see the results of multi-dimensional scaling of the four data sets. It seems that the lung cancer data set can easily be projected into two-dimensional space, which can not be said for the others. It is interesting to note that the tobacco data set has three very clear clusters, which are not the smoker and non-smoker classes but the tissue types, neonatal cord blood, placenta, and maternal peripheral blood.

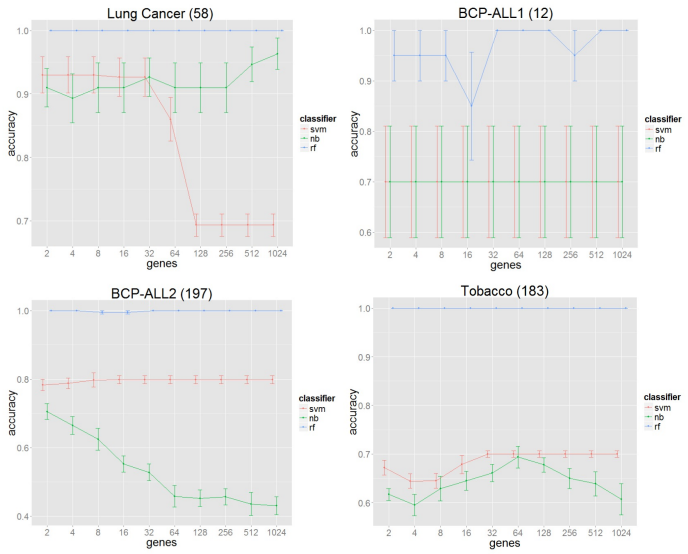


Figure 2: Number of genes used vs. accuracy for all classifiers on all data sets. Red, green, and blue denote SVM, NB, and RF, respectively.

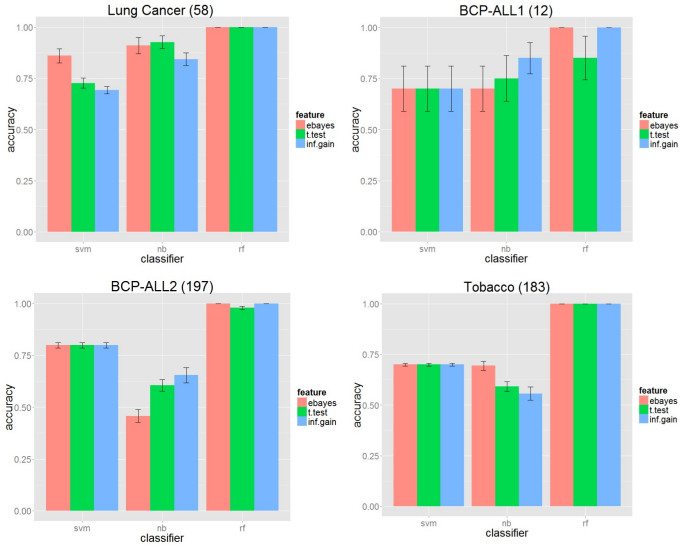


Figure 3: Comparison of the classifiers across the four datasets and with the three different feature selection techniques.

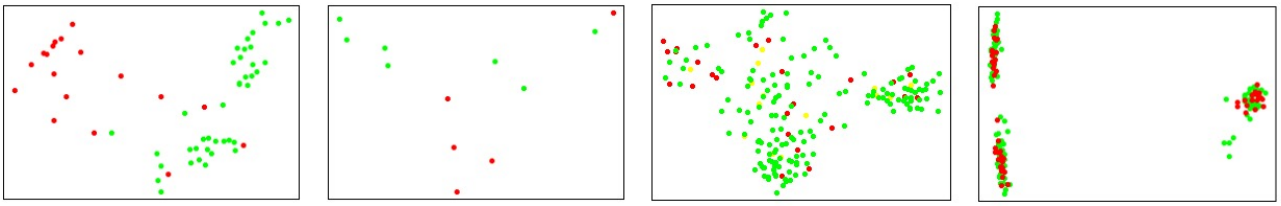


Figure 4: Multidimensional scaling of the four data sets. From left to right: lung, bcp-all1, bcp-all2, tobacco. Green, red and yellow colors denote control, cancer, and early relapse (only present in the third data set).

Table 3: Average 10-fold CV classification accuracy across datasets.

Data set/Approach	svm.ebayes	svm.t.test	svm.inf.gain	nb.ebayes	nb.t.test	nb.inf.gain
lung(58)	0.83 ± 0.014	0.81 ± 0.014	0.76 ± 0.012	0.92 ± 0.011	0.94 ± 0.009	0.85 ± 0.014
bcp-all(12)	0.70 ± 0.033	0.75 ± 0.034	0.72 ± 0.036	0.70 ± 0.033	0.70 ± 0.036	0.77± 0.035
bcp-all(197)	0.80 ± 0.004	0.80 ± 0.003	0.80 ± 0.004	0.53 ± 0.013	0.67 ± 0.012	0.7 ± 0.010
tobacco(183)	0.68 ± 0.004	0.68 ± 0.005	0.69 ± 0.003	0.64 ± 0.007	0597 ± 0.010	0.58 ± 0.009

Data set/Approach	rf.ebayes	rf.t.test	rf.inf.gain
lung(58)	1 ± 0.000	0.99 ± 0.002	1 ± 0.000
bcp-all(12)	0.96 ± 0.015	0.92 ± 0.025	0.99 ± 0.011
bcp-all(197)	0.99 ± 0.001	0.99 ± 0.002	0.99 ± 0.001
tobacco(183)	1 ± 0.000	1 ± 0.000	1 ± 0.000

Table 4: Post-hoc pairwise Nemenyi tests' Bonferonni-adjusted p-values. Bold indicates statistical significance (p<0.05).

	svm.ebayes	svm.t.test	svm.inf.gain	nb.ebayes	nb.t.test	nb.inf.gain	rf.ebayes	rf.t.test
svm.t.test	1	-	-	-	-	-	-	-
svm.inf.gain	1	1	-	-	-	-	-	-
nb.ebayes	**0.032**	**0.028**	0.073	-	-	-	-	-
nb.t.test	1	1	1	1	-	-	-	-
nb.inf.gain	**0.277**	**0.244**	0.559	1	1	-	-	-
rf.ebayes	**0**	**0**	**0**	**0**	**0**	**0**	-	-
rf.t.test	**0**	**0**	**0**	**0**	**0**	**0**	1	-
rf.inf.gain	**0**	**0**	**0**	**0**	**0**	**0**	1	1

We also computed the accuracy of the classifiers for different numbers of genes (across all data sets), shown in Figure 2. Random forests achieve excellent accuracy across the spectrum, outperforming bot SVM and NB. Most importantly, we tested classifier performance on multiple data sets and with different feature selection techniques. We set the number of features to the best $N = 50$ genes according to each feature selector, which results in three different sets of features. Thus, we obtain 9 classifiers, SVM, NB and RF each with 3 different sets of features selected, and we perform 10-fold cross-validation on all four data sets. The results are shown in Table 3, and summarized in Figure 3.

Comparing the results from the classifiers, the Friedman test p-value is $p < 10^{-16}$, which indicates that further pairwise tests are needed. The adjusted p-values are shown in Table 4, and show significant difference in the average classification accuracies for many pairs. Notably, RFs outperform both SVM and NB, whereas most comparisons between SVM and NB are not significant. It is also interesting to note that different selection methods do not seem to outperform each other. Finally, for each classifier and the three data sets that share the same genes, we train a classifier on one of the data sets and test on the other two, shown in Table 5. Not surprisingly, RF comes out on top again, with almost perfect classification accuracy. An interesting result is the between-data-set accuracy of the three algorithms. It appears that the information that is used to discriminate a cancer in one dataset can also be subsequently used to predict cancerous tissues in another data set for a different cancer. Once again, random forests seem to outperform both algorithms, with Friedman $p - value < 10^{-16}$. The

Table 5: Between data set accuracy for SVM, NB and RF, respectively.

train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	0.948	0.583	0.331
bcp-all(12)	0.690	1	0.438
bcp-all(197)	0.218	0.341	0.848
train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	0.931	0.333	0.277
bcp-all(12)	0.707	1	0.532
bcp-all(197)	0.690	0.418	0.792
train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	1	1	0.703
bcp-all(12)	1	1	0.703
bcp-all(197)	1	1	1

subsequent pairwise p-values, however, are significant only for RF vs. NB, as can be seen in Table 2.

5. DISCUSSION

The results reveal a number of things. First, Fig. 4 indicates that some data sets, like the lung data set, are probably easier to model, since the data is almost linearly separable in only two dimensions, which is definitely not the case for the last two data sets. Second, the increase (or decrease) of the number of genes plays a role in classifier performance and indicates which algorithms cannot deal with correlated data. For example, as we increase the number of features to hundreds, we can see that Naive Bayes' assumption of conditional independence of the features is violated, and the performance degrades. On the other hand, random forests do not seem to have a problem with the number of features, or the feature selector used, and clearly outperform, with statistical significance, both SVM and NB on all four data sets, as can be seen in Table 3. The choice of feature selector, however, does not seem to be statistically significant, which indicates that the performance largely depends on the classifier.

6. CONCLUSION

Due to today's availability of gene expression microarray data, many feature selection techniques and machine learning algorithms have been applied on them. We overviewed three such feature selectors: empirical Bayes, Student's t-test, and information gain, but there was no obvious correlation between a classifier and feature selector and no overall significantly better method. On the other hand, the statistical comparison of the three supervised learning algorithms: random forests (RF), support vector machines (SVMs), and naive Bayes (NB), finds RF as clearly superior to both SVM and NB, regardless of the number of features (genes) used, the feature selection technique, or the data set. We conclude that, although feature selection is a vital step in the process of microarray data analysis, most tecniques find appropriate gene subsets, and the difference in performance is mostly due to the classifiers. Future work would entail comparing other data sets, machine learning models, feature selection techniques, as well as developing appropriate statistical tests for comparing multiple classifiers on multiple data sets.

7. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[5] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature genetics*, (14):457–60, 1997.

[6] R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[7] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[8] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[9] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[10] M. Pirooznia, J. Yang, M. Q. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(Suppl 1):S13, 2008.

[11] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[12] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[13] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[14] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319, 2008.

[15] E. P. Xing, M. I. Jordan, R. M. Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer, 2001.

INVISIBLE SMART SYSTEMS FOR MOTION DETECTION AND ANALYSIS OF WALKING

David A. Fabjan
Jožef Stefan Institute
Jamova c. 39, Ljubljana
david.fabjan@ijs.si

ABSTRACT

With miniaturizations and low power consumption of the hardware components, we are seeing the uptake of the Internet of Things (IoT). Different dedicated sensors, often with already embedded processors and memory, can connect wirelessly to remote central controllers and actuators.

They are becoming ubiquitous with their numbers increasing rapidly in numbers and different ambient signals they are able to detect and process. Such devices are easily placed in familiar environment or on the human body and used to increase comfort, detect medical conditions and used when dealing with various issues of personal and public security. This overview paper is exploring novelty ways in detecting human behavior by introducing this new technologies, with emphasis on invisibility, privacy, and ambient based sensors for non-intrusive, and kinetic measurement of walking.

Keywords

Ambient sensors, Wearables, Non-intrusive, Motion, Gait analysis, Walking, Invisible, Smart floors, Smart Carpets, Sensors

1. INTRODUCTION

Motions and gait analysis using wearable sensors is seen as convenient and efficient manner of providing useful information for the user in various applications and environments. Such tasks can be achieved by systems of various complexity and portability, depending on the context in which they are deployed. The user's situation and needs often require even better privacy or invisibility for the dwellers. Important factors when deploying non-intrusive systems are also reduced costs and elimination of the need to wear a device with dedicated wearable sensors (accelerometer, tilt sensors, RFIT tags, etc.).

2. OUTLOOK ON THE SYSTEMS FOR THE MOTION DETECTION

There are a number of different motion detectors applicable for various users' needs that are currently available on the market and are being used in different general and living setups. Universally we can group such appliances into three main categories: Wearable and portable devices, Ambient devices and mixed, hybrid systems.

Wearable and portable devices (wearables) are sensors embedded in necklaces, bracelets, tags, smart phones, belts, etc. They are using sensors of various purposes such as: accelerometers, gyroscopes, blood pressure and blood oxygen saturation measurements to sensors for simple positioning and motion detection and are usually attached to end users physical body. Wearables can thus have diagnostic, as well as monitoring applications [1]. They can be used either indoor or outdoor environments, latter typically together with the use of smart phones.

Wearables are typically connected to smart phones, which are used for collecting the raw or partially pre-processed signal from the sensor, and acting as a hub for all body sensors. Examples of such setups range from simple motion detection to detection of neurodegenerative medical conditions or can be narrowly focused on motion and analysis of chosen extremity of the human body in various sport related needs. Smart mobile devices can also be used without other wearables, as standalone devices. They usually incorporate enough relevant sensors to measure and assess various energy expenditure activities or positioning and movement requirements for the user. In general, such mobile devices can act as remote computer with wireless connectivity, powerful application processing and storage capabilities, thus has options to act independently, to alert or inform end users on events with no connection to mobile networks. Usually smart phones are connected to mobile networks, with access to back end servers, which enables for better processing, analyzing and decision making.

Ambient devices are devices such as cameras or sensors built into furniture and other general appliances [2]. They perform a wide range of functions, such as health monitoring, gait analysis, location detection, and various security issues, similarly as wearable devices. They are used for the activity recognition, as help with daily activities, communication with friend and family, alerting medical or security personnel, etc.

Downside of ambient sensing technologies is stationarity, personal intrusion when using cameras or to install the sensors into home furniture, on the other hand there is no limitations of currently available battery technology and no stigmas, discomfort and they eliminates the need to remember to wear a device. Furthermore, ambient sensing can be made completely invisible, nonintrusive and have the ability to monitor many persons at the same time with just one sensor.

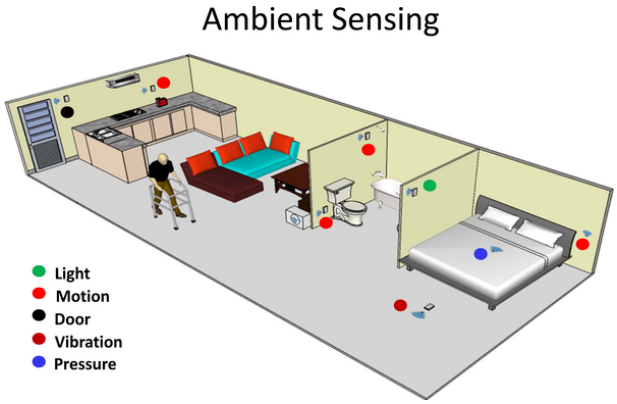


Figure 1: Ambient Sensing [3]

3. INTRODUCING AMBIENT FOOTSTEPS DETECTORS

One of the trending and most promising non-intrusive solutions from the ambient devices group are smart floors or active floors with sensors build into physical floor or into thin smart carpets usually made of polymers and /or textile with build in sensors. These devices must be able to capture or scan, process and present the information on human footsteps.

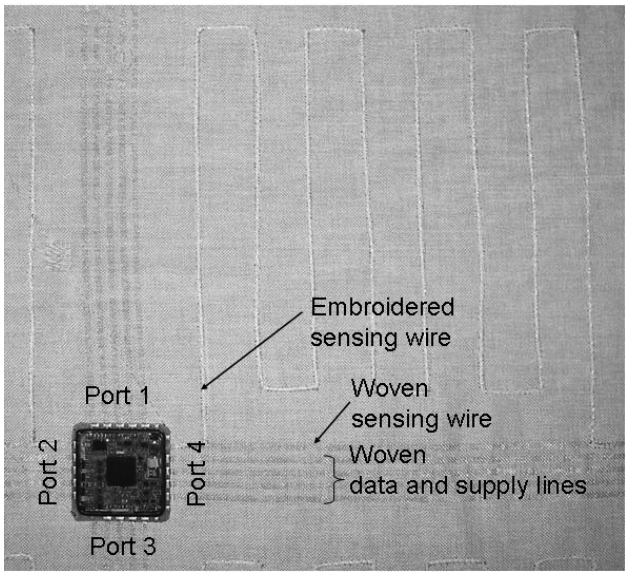


Figure 2: Example of smart carpet [9]

Capturing or scanning of the human footsteps has been in development since the 1980s. They involved techniques and sensors involving piezoelectric, resistive, force, capacitive, seismic and acoustic sensing principles. The utilized sensing technology has included load cells, switches, electromechanical films, optical fibers, flatbed scanners, resistive loads, accelerometers and RFID tags as well as combinations thereof [4], [5]. Signals or raw data from these sensors can then be analyzed to show the image of the footprint and to identify gradual changes in walking behavior.

Generally, human walking is a periodic movement of the body segments and includes repetitive motions. To understand this periodic walking course better and easier, the gait phase must be used to describe an entire walking period. Gait analysis is the systematic study of human locomotion. This type of analysis involves the measurement, description, and assessment of quantities that characterize human locomotion [6]. Through gait analysis, the gait phase can be identified, the kinematic and kinetic parameters of human gait events defined.

The data gained from the ambient sensors is sent to back end servers where various machine learning methods are applied on the activated sensor outputs. This enables us to identify footsteps of the individual subject and estimate the trajectory taken by the subject taken during his walk on the smart floor or carpet. To identify the footsteps applied can be clustering algorithms based on Maximum Likelihood Estimate (MLE) [7] and Rank Regression [8] analysis. Such a system can, for example, be used in medical and security setups [9].

4. CONCLUSIOS

Intelligent environments which identify and track the behavior of occupants in a room are gaining immense attention. Smart floors and smart carpets can be made invisible, easily tailored to various shapes and used in home, medical or security setups. They present best solution when applied in the indoor environments, without shortcomings of the wearable sensors. In the case of smart carpets, they can be tailored to any flor plan and wirelessly connected to computer, thus made truly non-intrusive and invisible to the users.

5. REFERENCES

[1] Bonato P: Wearable sensors and systems. From enabling technology to clinical applications. IEEE Eng Med Biol Mag 2010, 29:25-36

[2] Gjoreski, Hristijan, et al. "Competitive Live Evaluations of Activity-Recognition Systems." Pervasive Computing, IEEE 14.1 (2015): 70-77.

[3] Patel, Shyamal, et al. "A review of wearable sensors and systems with application in rehabilitation." J Neuroeng Rehabil 9.12 (2012): 1-17.

[4] Cantoral-Ceballos, Jose, et al. "Intelligent Carpet System, Based on Photonic Guided-Path Tomography, for Gait and Balance Monitoring in Home Environments." Sensors Journal, IEEE 15.1 (2015): 279-289.

[5] Torres, Roberto Luis Shinmoto, et al. "What if Your Floor Could Tell Someone You Fell? A Device Free Fall Detection Method." Artificial Intelligence in Medicine. Springer International Publishing, 2015. 86-95.

[6] Ghoussayni S., Stevens C., Durham S., Ewins D. Assessment and validation of a simple automated method for the detection of gait events and intervals. Gait Posture. 2004; 20:266–272.

[7] R. Fisher, “A mathematical examination of the methods of determining the accuracy of an observation by the mean

error, and by the mean square error," in Notices of the Royal Astronomical Society 80 758 770, 1920, vol. 1.Z.

[8] R. Bradley and M. Terry, "The rank analysis of incomplete block designs: I. the method of paired comparisons," in Biometrika, 1952, vol. 39, pp. 324 – 345.

[9] Savio, Domnic, and Thomas Ludwig. "Smart carpet: A footstep tracking interface." Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on. Vol. 2. IEEE, 2007.

Application for sexually transmitted infection risk assessment

dr. Gašper Fele-Žorž[1*], Karolina Počivavšek[2**], Jaka Konda[1], Ana Marija Peterlin[2], Alen Ajanovič[1], Ana Prodan[2], Saša Rink[2], dr. Anton Gradišek[3], prof. dr. Matjaž Gams[3], prof. Mojca Matičič, MD, PhD[4***]

[1]Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana *polz@fri.uni-lj.si

[2]Medical Faculty, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana **karolina.pocivavsek@gmail.com

[3]Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia

[4]Clinic for Infectious Disease and Febrile Illnesses, University Medical Centre Ljubljana, Japljeva 2, 1525 Ljubljana, Slovenia ***mojca.maticic@kclj.si

ABSTRACT

We present a web application to detect risks related to sexually transmitted diseases. The application works as a questionnaire about sexual behaviour of the individual and, based on the answers, calculates the risk of being infected. The application also works as an informational tool with educating about STD and uses a combination of approaches from computer science and psychology to deliver a usable, clean interface with which the user feels safe.

Keywords

STI, web application, questionnaire, diagnosis, risk level assessment, information, teenagers

1. INTRODUCTION

Sexually transmitted infections (STIs) present an important public health issue, since daily, nearly a million people contract at least one STI, including the human immunodeficiency virus (HIV) [1]. Prevention of STIs is crucial as the viral ones, such as herpes simplex virus (HSV), human papillomavirus (HPV), hepatitis B virus (HBV) and HIV infections are incurable. Chlamydia, gonorrhoea and syphilis are still a major cause of disability and death, despite the fact that effective antibiotic treatment is available. The majority of STIs remains asymptomatic presenting no signs and symptoms of disease for a long period and therefore remains undetected. However, chronic STIs can lead to serious consequences including infertility, ectopic pregnancy, cervical cancer and several others [2]. Moreover, asymptomatic infection can easily be transmitted to sexual partner and foetus or neonatus in pregnant women.

Young people are particularly vulnerable to STIs. According to World Health Organization (WHO) rates of infection are the highest among 20-24 year-olds, followed by those at 15-19 years of age [3]. One of the most important reasons is lack of proper sex education, especially acquiring information on STI prevention and counselling on safe sexual behaviour [4]. Other prominent factors that place youth at risk are multiple sex partners, greater biologic susceptibility, insufficient screening, lack of access to healthcare and confidentiality concerns [5-8].

The Internet is emerging as a powerful tool for learning as well as more or less reliable source on health information. Young people are especially likely to seek health information online, when they are unable to confide in others or feel uncomfortable disclosing certain information to their general practitioner. When searching for medical information, an estimated 25% of adolescents turn to web browsing, which suggests that online contexts present an open and safe space in which young people can express themselves and promote healthy habits [9,10].

To take advantage of an opportunity that the Internet clearly provides, we decided to develop a web-based application offering concise, medically confirmed and evaluated information on prevention, characteristics and treatment of most common STIs. Besides, we set to create a questionnaire to assess the probability of having STI based on a user's history of sexual behaviour and other health issues that might have occurred. Using the application the visitors who engage in risky sexual practices are supposed to be motivated for visiting available STI clinics to get proper management and counselling and to prevent the further spread of infection.

The application is available at aspo.mf.uni-lj.si

2. METHODS

2.1 FUNCTIONAL REQUIREMENTS

The goal of the project within which the ASPO website was created is to educate young people - between 15 and 25 years of age - with regard to STIs and safe sexual practices. To reach the target population, the relevant information must be made available on mobile devices as well as personal computers.

The information must be presented in an interesting way. One way to achieve this is to make the site visually appealing. Another is to make it interactive in some way - for example by including a questionnaire.

To ensure that the information is trustworthy, it should be prepared and/or checked by medical professionals. This means that most of the content will be created by people who are not web developers. A mechanism is needed to make the adding of new content easy for people without a computing background.

2.2 THE INFORMATION PRESENTED

The information which is of interest to our target population can be separated into a few topics. First, information regarding safe and risky sexual behaviour. A part of this information can be presented in the form of a questionnaire. Second, general information regarding STIs. Third, instructions on where and how one can get professional help.

3. ETHICAL CHALLENGES

Since sexual relations are a sensitive subject, care must be observed when dealing with people who are seeking information regarding various sexual practices. For example, even though some questions in the questionnaire regarding safe sexual practices might seem like something a physician might ask their patient, the feedback must never be presented in a way that could be interpreted by a visitor as a medical diagnosis. Because some of the questions asked are very personal, the answers should never be stored in a way that would allow anyone to connect the answers to the person using the questionnaire.

When describing STIs, one way to present the symptoms is through the use of pictures which some people might find offensive. It is difficult to weigh the benefits of educating the public against the possibility that graphic images of sexual organs made available to teenagers might cause a public outcry.

4. RESULTS

4.1 TECHNICAL CONSIDERATIONS

During the initial phases of our project, we considered multiple possible solutions to satisfy the requirement of making the information available to users of mobile devices. One possibility was to build multiple applications - one for the web and one for each supported mobile platform, such as iOS and Android. By building a separate application for each platform, the user interface could be adapted to the expected screen size. By optimizing the application, the drain on the battery could have been kept to a minimum. One downside to this approach would have been the amount of work required to implement at least three versions of the application. Also, because applications on most mobile platforms are rarely updated automatically, we would have faced a choice of either presenting the user with possibly outdated information or requiring a constant connection to the internet, thereby negating most of the possible battery savings. In the end, we decided to only build a web-based version of our application, but to make the design responsive. This way, the codebase is shared between mobile and desktop clients. Also, any updates the site may receive are made immediately available on all platforms.

4.2 FRONT-END

The website is built as a single-page application (SPA) using the Angular JS library made by Google [11]. By using AngularJS, we were able to simply create a dynamic and modern web site, making it more attractive for the primarily targeted population. By building the site as an SPA, the transition of the web page to a mobile web application is greatly simplified.

The frontend follows a strict MVC (Model View Controller) design pattern which makes our website extremely modular by separating data from code logic and code logic from the user interface.

Figure 1: First page of our website where main menu is visible at the top, large button that attracts users to take a test and in the footer organizations participated in the making of our project.

4.2.1 GENERAL INFORMATION

The main site is separated into seven sections. Information regarding safe and risky sexual behaviours is made available through the questionnaire and the section on condom use. Information regarding STIs is presented in the section containing descriptions of common STIs, the section describing the most common symptoms of STIs, and the section describing a number of medical cases. Information on where one can get professional medical help regarding STIs is presented in the section containing a description of a medical examination at a clinic which offers testing for STIs and a list of all clinics and other public medical establishments offering such testing in Slovenia. The final section is about the project itself.

From a technical standpoint, the most interesting part of the site is the questionnaire, which is described in the following subsection.

4.2.2 THE QUESTIONNAIRE

The questionnaire is the central part of the web page. Its intent is that users submit their habits about sexual life and the output is the risk level of their particular scenario.

The questionnaire was composed based on the questionnaire a patient may receive at the University Medical Centre Ljubljana. One of the problems we faced while choosing the questions was the fact that the symptoms of many of STIs are similar. The questionnaire covers most of the common ones (HIV/aids, syphilis, hepatitis…), but all of them have very similar symptoms (such as rash, warts, discharge, etc.) making them impossible to distinguish between each other without additional (laboratory) tests.

Another factor considered while choosing the questions was the psychological effect. The questionnaire on our website is intended for people who suspect or notice any symptoms or behaviour which indicates they may have an STI. With that in mind, the questions asked are slightly intimidating, which gives the user an additional reason to visit the nearest doctors for further examination.

4.2.2.1 The questionnaire from a user's standpoint

When entering the questionnaire, a user is greeted by a message indicating approximately how long the whole test will take. It is divided into three main parts which are seamlessly joined together. The first part consists of questions, the purpose of which is to collect some of the basic user information (age, sex, etc.). The second part consists of questions the user's sexual activity

(number of sexual partners, past sexual activity, etc.). The last question set is targeted directly at the symptoms the user may be experiencing (rashes, pain, etc.).

This particular structure is effective because it begins in a very casual manner that is not too intrusive. As the questions continue they become progressively more serious to give off the impression of urgency. The questions also serve as informational sources and provide the user with some more background information about the case he might have. If the user notices that a question contains a symptom which they might have once had or have noticed on any of their partners, they can get access to more information about the infection by visiting other parts of the website.

The order in which the questions are presented is dynamic, so a user only answers the questions which are applicable to them. For example, if a user answers the question "Do you have a rash?" with "yes", the next question might be "How would you describe your rash". If they answer the first question with "no", they are not presented with the second question at all. This way, users who do not exhibit certain symptoms do not have to answer questions regarding problems which they obviously do not have. The same system is also used to tailor the questions according to the user's sex. When the user finishes answering all the questions, the system determines a risk level based on the answers given. The risk level is presented as one of three colours (green, yellow, red) where green represents a low level of risk and red represents a high level of risk. Each end result is also accompanied by a description of the severity of the situation and instructs the user on what to do next.

All questions are loaded onto the client as soon as the user visits our web page. This ensures the responsiveness of the questionnaire in the case of an unreliable Internet connection and reduces the load on our server.

4.2.2.2 Questionnaire implementation

Originally, the questionnaire was supposed to have been implemented in the form of a decision tree. In this tree, each node would represent a question and each link would represent a possible answer. The answer to each would determine the path through the rest of the questions. The questions would also be ordered according to their relevance.

After manually building this type of decision tree on paper, it became obvious that most of the questions are shared among all possible paths a person might take through the quiz. The questions in our questionnaire are now stored in a simple list. First, the question at the start of the list is presented to the user and removed from the list. When the user answers the question, all questions which are made redundant by the answer are removed from the list. This procedure is repeated until no questions remain in the list.

In the first versions of the questionnaire, many answers would cause some of the questions to be removed. In the latest version, only three such answers remain. The main reduction of questions which may be removed after an answer happened when we changed the questions from being gender -specific to being grammatically gender-neutral. For example, the original questions "have you noticed any lesions on your penis" and "have you noticed any lesions on your vagina" has been replaced by "have you noticed any lesions on your sexual organs".

The calculation of results is separated from the question order.

To determine the threat level, seven bits are used to track all possible outputs. Four bits represent the three possible threat levels plus an additional bit which is used to separate yellow and red threat assessments. One bit is used to track drug use and the remaining two are meant for special cases like a marital status. A simple example of this would be the fact that if a user is married and has one partner (their spouse) without protection, the threat level would default to yellow, but if we decide to use the marital flag, it would change to green. This allows us to circumvent some of the natural anomalies which occur by the system we use and this way we can add an extra level of user protection by having specialized messages for certain cases. In all remaining cases, the most alarming threat level from all answers is returned.

One example of a green diagnosis would be when the user tells us that they only had one regular sexual partner and that they used a condom. Another clear example of a green diagnosis would be when the user never indulged in a sexual activity in the first place.

The middle or the yellow diagnosis would be when the user has acted in a way where they might have been exposed to a sexually transmitted disease, but the chances of it happening were still relatively low.

The highest risk diagnosis is considered when the user has acted in a very sexually promiscuous way, when the user indulged in a very high risk sexual activity (many different partners, no protection, etc.).

An additional message is designated for people who consume drugs. Even though they might not have sexual intercourse, STDs could be transmitted through a needle shared between people. A message is displayed, warning about such options. Like mentioned before, this is also tracked by our system through a special bit.

4.2.3 INFORMATIVE SUBSECTIONS

Most of the information on the site is in the form of text on various pages describing various sexually transmitted diseases. These pages were written by medical professionals who are not fluent in HTML. The texts are written in the Markdown domain-specific language [12], which is easy enough for non-programmers to learn quickly. Instead of rendering the pages into HTML on the server, the pages are stored and transferred in the format in which they are written and rendered in the client's browser. Those pages which require advanced formatting, such as the list of clinics, are still edited by programmers.

4.3 BACK-END

When the project within which the application was created started, our goals were much higher than what was created in the end. Originally, the data driving the questionnaire results was supposed to be derived through machine learning. The questionnaire was supposed to be used to create a large dataset of symptoms and diagnoses related to STIs. Because Weka [13], one of the most popular machine-learning libraries available is written in Java, we decided to use that platform on our back-end. Java also had the advantage of being the language most of the programmers working on the project have the most experience working with.

The current backend is written using Java EE. While most of the questionnaire logic is implemented in the front-end, the questions themselves are still stored server-side in a relational database. The back-end uses the Java persistence API as the object-relational mapper [14]. The service serving questionnaire data uses the high-level interfaces and annotations used to create RESTful service resources [15]. The questionnaire data is serialized using the

GSON library by Google [16]. The application does not depend on any specific relational database. During the development, we used MySQL [17] to store the questionnaire data. As of the moment of this writing, we are using a Percona Server relational database [18] to store the data. The application is deployed on a Wildfly 9.0 Java EE application server [19] hidden behind an nginx [20] web server.

In parallel with the main back-end which is written in Java, a second back-end was written in PHP [21]. This enabled us to create a prototype site before the main back-end was finished. Like the Java back-end, the PHP back-end is also fully functional, so our application could be deployed on any server web running, for example, Apache [22], MySQL and PHP.

5. DISCUSSION

Because were not able to find a single large dataset of symptoms and diagnoses of STIs to learn from, it turned out to be impossible to create the questionnaire using machine learning. Even if such a questionnaire were created, presenting the results to the users in a way which could be interpreted by the users as a form of diagnosis would have been unacceptable without an actual examination by a medical professional. This made the first reason for choosing our platform irrelevant. We also ran into some problems because, although they were fluent in Java, none of the programmers involved in our project were familiar with all the technologies used to actually implement a web application in Java, making the second reason for choosing Java irrelevant.

The front-end was also originally designed to be as flexible as possible. The order of questions could be made flexible. Initially, we intended for the user to only answer the minimal number of questions necessary to produce a diagnosis. In the end, because the questions themselves are supposed to get the user to think about their behavior, very few questions are actually avoided. Without much loss with regard to functionality, the quiz could have been implemented as a simple list of questions.

The photographic images showing examples of STI symptoms are very graphic pictures of genitalia. While designing the site, we feared that they may offend some of our visitors or, since some of our visitors are expected to be minors, their parents. Multiple options were considered. We immediately agreed that the pictures should only be displayed after the user has been warned about the graphical nature of the images. One option was to remove the images altogether. Another was to replace the photographs with pictures. In the end, after much deliberation, we decided to present the photographs unaltered. We believe that normal young people even at the lower end of our target age group are perfectly capable of finding sexually explicit material on their own. The benefits of educating them about the symptoms of STIs therefore outweigh the danger of presenting them to aesthetically non-pleasing images of sexual organs.

5.1 SUMMARY

By developing the application, we wanted to approach an audience that is most vulnerable to an ever rising habit of poor sexual education. We wanted to tackle that obstacle by developing a user friendly, psychologically inviting website and application that serves to show the user where he can seek medical advice and help if they ever encounter a situation in which they find themselves compromised. By supplementing the site with actual facts and procedures that the user will find themselves encountering in a real life world scenario, the site tries to emulate the feeling of being actually interviewed by a certified doctor and imply the sense of urgency so the user will take the necessary precautions to minimize the risk of the disease affecting them in the long term.

6. ACKNOWLEDGMENTS

The development of application to identify risks for sexually transmitted infections was partly funded by the European Union through the European Social Fund. The project is implemented under the Operational Programme for Human Resources Development for the period 2007-2013, Priority axis 1 "Promoting entrepreneurship and adaptability" 1.3. "Scholarship Scheme", within the approved operations "A creative path to practical knowledge."

7. REFERENCES

[1] WHO | Report on global sexually transmitted infection surveillance 2013: 2014. *http://www.who.int/reproductivehealth/publications/rtis/stis-surveillance-2013/en/*. Accessed: 2015- 08- 25.

[2] WHO | Global strategy for the prevention and control of sexually transmitted infections: 2006 - 2015: 2007. *http://www.who.int/reproductivehealth/publications/rtis/9789241563475/en/*. Accessed: 2015- 08- 25.

[3] WHO, 1995. *Sexually Transmitted Diseases Three Hundred and Thirty-three Million New, Curable Cases in 1995*.

[4] WHO | Sexually Transmitted Infections among adolescents: 2005. *http://www.who.int/maternal_child_adolescent/documents/9241562889/en/*. Accessed: 2015- 08- 25.

[5] NSFG - 2006-2010 NSFG - Public Use Data Files, Codebooks and Documentation: 2015. *http://www.cdc.gov/nchs/nsfg/nsfg_2006_2010_puf.htm*. Accessed: 2015- 08- 25.

[6] CDC | 2012 STD Surveillance: 2015. *http://www.cdc.gov/std/stats12/*. Accessed: 2015- 08- 25.

[7] Satterwhite, C., Torrone, E., Meites, E., Dunne, E., Mahajan, R., Ocfemia, M., Su, J., Xu, F. and Weinstock, H. 2013. Sexually Transmitted Infections Among US Women and Men. *Sexually Transmitted Diseases*. 40, 3 (2013), 187-193.

[8] Tao, G., Hoover, K., Leichliter, J., Peterman, T. and Kent, C. 2012. Self-Reported Chlamydia Testing Rates of Sexually Active Women Aged 15–25 Years in the United States, 2006–2008. *Sexually Transmitted Diseases*. 39, 8 (2012), 605-607.

[9] Ybarra, M., Emenyonu, N., Nansera, D., Kiwanuka, J. and Bangsberg, D. 2007. Health information seeking among Mbararan adolescents: results from the Uganda Media and You survey. *Health Education Research*. 23, 2 (2007), 249-258.

[10] Ngo, A., Ross, M. and Ratliff, E. 2008. Internet influences on sexual practices among young people in Hanoi, Vietnam. *Culture, Health & Sexuality*. 10, sup1 (2008), S201-S213.

[11] AngularJS *https://angularjs.org/* Accessed: 2015-09-02.

[12] Markdown *http://daringfireball.net/projects/markdown/* Accessed: 2015-09-02.

[13] Holmes, G., Donkin, A. and Witten, Ian H. 1994. WEKA: a machine learning workbench. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems.*, 357-361

[14] Java persistence API

http://www.oracle.com/technetwork/java/javaee/tech/persistence-jsp-140049.html Accessed: 2015-09-02.

[15] High-level interfaces and annotations used to create RESTful service resources.

https://docs.oracle.com/javaee/7/api/javax/ws/rs/package-summary.html Accessed: 2015-09-02.

[16] Google-GSON

https://github.com/google/gson Accessed: 2015-09-02.

[17] MySQL

https://www.mysql.com Accessed: 2015-09-02.

[18] Percona

https://www.percona.com/software/mysql-database/percona-server Accessed: 2015-09-02.

[19] Wildfly

http://wildfly.org Accessed: 2015-09-02.

[20] nginx

http://nginx.org Accessed: 2015-09-02.

[21] PHP

http://php.net Accessed: 2015-09-02.

[22] Apache

http://httpd.apache.org/ Accessed: 2015-09-02.

Prelisičenja in pravi pomen Turingovega testa

Matjaž Gams,
Lana Zemljak,
Vesna Koricki-Špetič,
Blaž Mahnič
Jožef Stefan Institute
Jamova 29, Ljubljana, Slovenia
matjaz.gams@ijs.si

POVZETEK

Desetletja se raziskovalci umetne inteligence, kognitivnih znanosti in filozofi trudijo, da bi rešili Turingov test, v katerem mora človek odkriti, kdaj komunicira z računalnikom in kdaj s človekom. Pred nekaj leti je prišlo do novih trditev, da je program prelisičil preiskovalce, pa se je izkazalo, da je bil sistem izveden nestrokovno. V tem prispevku opisujemo značilnosti Turingovega testa in izvajamo eksperiment, v katerem skušamo ugotoviti, ali ljudje uspešno prepoznavamo vizualno podobo (androidna verzija Turingovega testa). Meritve nakazujejo, da imamo ljudje težave s prepoznavanjem celo izrazito enostavnih verzij Turingovega testa, če nismo zbrani.

Ključne besede

Turingov test, prepoznavanje, androidna lutka

1. UVOD

Turingov test (TT) je eden najbolj znanih znanstvenih testov inteligence. Osnova testa je detektivsko izpraševanje računalnika, ki se pretvarja, da je človek. Izvirni test je bil nekoliko drugačen – Turing je leta 1950 predlagal, da izpraševalec komunicira preko jezikovne iterakcije (npr. preko tipkovnice) z dvema subjektoma, eden je človek in drugi računalnik. Izpraševalec ima na voljo le nekaj minut in ugotoviti mora z veliko verjetnostjo, kdo je človek in kje je računalnik: »an average interrogator will not have more than 70 percent chance of making the right identification (as between human and computer) after five minutes of questioning.« Turing je verjel, da se bo to zgodilo leta 2000, ko bodo računalniški pomnilniki presegli 10^9 bitov.

Test je Turing opisal leta 1950 v članku "Computing Machinery and Intelligence," [1, 2] (Turing, 1950; p. 460). Članek se začne z: "I propose to consider the question, 'Can machines think?'" Turing je test predlagal kor razložitev oz. praktično uporabo definicije za razmišljajoče računalnike. Zato je vprašal: "Are there imaginable digital computers which would do well in the imitation game?"

Razne verzije Turingov testov in debate o njih so zelo aktualne pri sedanjem izrednem razvoju umetne inteligence in so se pojavile tudi na zadnji konferenci IJCAI2015.

V tem prispevku obravnavamo testiranje verzije Turingovega testa v Odseku za inteligentne sisteme Instituta »Jožef Stefan«. Znanstvena hipoteza je bila, da so ljudje precej »naivni« in da jih je možno prevarati z dokaj enostavnimi prijemi. To lastnost je potrebno upoštevati pri dejanskem izvajanju Turingovega testa, če seveda testiranje v tem prispevku potrjuje hipotezo.

2. TURINGOVI TESTI

Osnovna verzija Turingovega testa je bila prilagojena inačica tedanje priljubljene družabne igrice. Izpraševalec je poskušal na osnovi vprašanj in odgovorov preko tretje osebe ugotoviti, kdo od dveh je moški in kdo ženska. Oba, tako moški kot ženska, sta se pretvarjala, da sta ženska. Ob vprašanjih in odgovorih o specialnosti ženske ali moškega se je skupina veselo zabavala. Morda imate kakšno zamisel, kaj bi vi vprašali moškega o ženskah, da bi pokazal svoje neznanje, ženske pa bi takoj vedele, koliko je ura?

Obstaja tudi vrsta drugih »zabavnih« oz. »satiričnih« verzij Turingovega testa. Gunderson [3] je prvi omenil, da gre za funkcionalni test in satirično predlagal, da ima izpraševalec dostop do dveh sob, samo da namesto izpraševanja pomoli nogo ali roko v sobo, stroj v eni in človek v drugi pa v ta ud mečeta kamenje. Ali lahko človek prepozna razliko? Ali bi lahko rekli, da je stroj inteligenten kot človek, če izpraševalec ne bi mogel ločiti razlike?

Odgovor že nakazuje pravo naravo Turingovega testa: gre za »vrtanje« po mentalnih procesih neznanega subjekta, ki naj bi odgovarjal pošteno (človek ne sme dajati odgovore kot računalnik). Pri tem je seveda možno marsikaj in tudi marsikatera dilema ostane odprta, kar nam kmalu odkrije globino Turingovega testa. Če recimo izpraševalec ne loči med moškim in žensko v predhodniku Turingovega testa, ali je res omenjeni moški ženska? Ali mu ne priznamo človečnosti? Ali nerodni izpraševalec, ki ne zna postavljati prava vprašanja, lahko kaj ugotovi? Ali je računalnik, ki uspešno prestane Turingov test človek? Semantika dogajanja pa je, da test pri kvalitetnem izpraševanju loči, če je razlika dovolj velika. Če se konkretna moški in ženska ne ločita po svojih odgovorih, potem sta si mentalno dovolj podobna? Če sta računalnik in človek dovolj podobna v svojih odgovorih, potem sta mentalno primerljiva ? Če se naivni izpraševalec izgubi v nesmiselni komunikaciji, lahko priskoči izkušeni in postavi vprašanje, ki ga lahko reši le človek (»računski stroj« ni kalkulator, ampak naprava, ki zna obdelovati informacije in preračunavati z njimi), potem je računski stroj, ki odgovori kvalitetno kot človek, na podobnem nivoju kot človek?

Med verzijami Turingovega testa omenimo totalni Turingov test (TTT), kjer je poleg umske funkcionalne podobnosti potrebna tudi podobna funkcionalna sposobnost, ali pa samo omenjena fizična sposobnost. Pri tem ne gre za to, da tudi izgleda kot človek, ampak izpraševalec postavi neko fizično nalogo, npr. premik fizičnih objektov, npr. kocko na kocko. Izpraševana oseba oz. naprava lahko hkrati komunicira v naravnem jeziku ali pa samo skrita opazuje premike fizičnih objektov in jih reproducira,

izpraševalec pa ne sme gledati med premikanjem, ampak po končani nalogi.

Še bolj zapletena verzija je (to je test za androide oz. humanoide) totalno totalna verzija Turingovega testa (TTTT), kjer lahko izpraševalec opazuje in interaktira z izpraševancem med celotnim postopkom tako jezikovno kot fizično. Najbolj ekstremna verzija testa pa je opazovanje populacije objektov s TTT testi.

Med drugimi verzijami test omenimo tudi CAPTCHA, to je grafični prikaz nekaj simbolov, ki jih človek loči, računalniki pa ne. To je tudi najbolj enostaven in najhitrejši test v praksi, ki ločuje med ljudmi in agenti na spletu.

Običajne pripombe glede TT so:

- gre za komunikacijske in ne intelektualne sposobnosti

  Ta argument temelji na možnosti, da bi lahko stroj zbral oz. našel odgovore na vprašanja, ne da bi jih v resnici znal vsebinsko razložiti. Tak bi bil npr. Google »bodočnosti« oz. nek super tabelarični pristop. Obstajajo pa znanstveni prispevki, da to niti teoretično ni možno. S Turingovimi besedami: »the question-answer method seems to be suitable for introducing almost any of the fields of human endeavor that we wish to include«

- gre za funkcijsko in ne globinsko analizo

  Argument tu je podoben kot pri prejšnjem ugovoru: različne interne strukture imajo različne funkcijske sposobnosti in če nek subjekt pokaže primerne funkcijske sposobnosti, potem njegova interna struktura niti ni pomembna – z drugimi besedami: inteligenco bi priznali tako inteligentnim živalim kot vesoljcem.

- superinteligenten subjekt ne bi opravil TT, ker bi bil drugačen od človeka

  Ta argument je sploh čuden, saj gre pri testu za ugotavljanje inteligence, tj. ali zna stroj za silo reševati podobne intelektualne naloge kot človek. V primeru superinteligence bi TT pač pokazal, da je ne samo dosežen, ampak tudi presežen nivo človeka oz. bi se superinteligenca brez težav pretvarjala, da je človek [4].

- potrebno je spremeniti verzijo Turingovega testa, npr. jo poenostaviti, da bomo lahko merili napredek

Taka je npr. verzija Lobnerjevega testa [5], ki se izvaja letno. Začela se je 8. 11. 1991. Obstaja več verzij testov z omejeno ali odprto komunikacijo in različnimi nagradami, recimo za najbolj uspešen sistem, četudi ne uspe rešiti testa z zadostno verjetnostjo.

Kritiki kot menijo, da gre pri TT le za varianto P. T. Barnuma, ki je masovno zavajal oz. sleparil ljudi z raznovrstnimi prevarami, temelječ na naivni popularni človeški psihologiji, npr. tako, da je svojcem umrlim prodajal biblije z izgovorom, da jih je pokojni naročil. No, v tistih časih je bilo to možno, danes pa bi moral pri roki imeti raznovrstne knjige od korana dalje. Je pa ta ideja o človeški naivnosti del eksperimentov z lutko, opisanem v nadaljevanju.

Verjetno najbolj znan program za komunikacijo z ljudmi, ki je uspešno zapletel določene posameznike v intimne pogovore po več deset minut, se je imenoval ELIZA in je igral žensko terapevtko od leta 1966 dalje. Tudi v našem odseku smo imeli izredno uspešno verzijo ELIZE, vendar je bila njena popularnost razkrinkana slučajno po nekaj letih uporabe in smo jo umaknili iz prometa. Trik je bil v tem, da je včasih po nesreči zamešala odgovore. Dokler so bili izpraševalci le mladoletniki, ki so v angleščini povpraševali zlasti glede intimnosti, erotike in opojnih substanc, ni bilo težav. Ko pa je nekoč poštena profesorica dobila nekaj umazanih odgovorov, ji je prekipelo in je na to opozorila javnost. Problem se je pojavil zato, ker naša verzija ni ločevala med odgovori z iste inštitucije in je zamešala odgovore, če je komunikacija enega posameznika časovno prehitela predhodno.

ELIZA je občasno dajala res smešne odgovore. Primer:

Q: Do you want to have sex with me?
ELIZA: Please go on.

Ne glede na vse povedano pa je pred 40 leti povprečen amaterski šahist brez težav premagal računalnik, medtem ko danes niti najboljši šahisti nimajo nobenih možnosti proti šahovskim programom, v nasprotju s tem pa se povprečni časi, potrebni za razkrinkanje računalnika v Turingovem testu, niso kaj dosti spremenili. Če bi bila sposobnost računalnika kot pri šahu odvisna od računske moči, bi morali časi, potrebni za razkrinkanje računalnika pri Turingovem testu, drastično rasti. Pa ne padajo. Kako je to mogoče [6,7],?

Tudi na IJCAI 2015 (http://ijcai-15.org/) so organizirali debate na temo Turingovega testa. Tak je bil panel “Rethinking the Turing Test”. Večina govorcev se je strinjala, da potrebujemo boljši test, ker sedanji ne kaže nobenega napredka, ki ga v resnici dosega umetna inteligenca, ker to ni test inteligence, ampak človeške psihologije, imitacije itd. Najuspešnejši so programi, ki zavedejo ljudi z raznimi zvijačami, to pa kaže na neprimernost testa, ki bi moral kazati napredek pri razvoju umetne inteligence. Predlagali so vrsto testov, s katerimi bi nadomestili Turingovega, recimo shema Winograda, kjer računalniki dosegajo okoli 70 (povsem naključno bi pa 50). Drugi predlagajo test Captcha, tretji test Lovelace. Zanimivi so testi, kjer podajamo računalniki nesmiselne slike ali trditve in ugotavljamo odstotek uspešnih prepoznav. Najbolj boleča trditev je bila od otroka enega od panelistov: »Kako bi izgledal naš svet, če Turingovega testa ne bi bilo?”

Po drugi strani je potrebno omeniti, da ima Turingov test določene probleme, ker je binaren: Da ali Ne. Če ne drugega, bi si želeli nek test, ki bi bil kvantitativen, npr. med 0 in 1. Na IJCAI je večina panelistov predlagala drugačne teste. Eden redkih drugače mislečih je bil Stephen Muggleton, ki je omenil, da gepardi tečejo hitreje kot ljudje, pa vendar niso tako inteligentni kot ljudje, pri meritvah hitrosti pa bi se vseeno dobro odrezali. Če bo predlagan kakšen drug test, ne bo meril prave inteligence, ampak nekaj drugega.

Mnenje prvega avtorja je, da ni nič narobe s predlaganjem drugih testov, samo ne bodo merili tiste prave človeške inteligence, ki bi jo empirično pokazal/dokazal TT. Predlagani drugi testi bodo pač merili nekaj drugega, kar ni samo po sebi nič narobe, samo ne smemo se slepiti, da bodo merili ključne človeške lastnosti podobno kot TT. Pri TT gre za izjemno zapletena vprašanje in interpretacije, ki šele s svojo zapletenostjo pokažejo vso moč človeškega uma. Kot primer navedimo izpite pri predmetu »Kognitivne znanosti«, kjer prvi avtor tega referata vnaprej obvesti študente, da bodo postavljali vprašanja kot bi jih pri TT.

Študentje večinoma pridejo na izpite s pripravljenimi vprašanji, zato je potrebno zamenjati vsebino in na nekem drugačnem področju analizirati stopnjo razumevanja pri študentih. Ta vprašanja jim povzročajo kar precej težav, kar kaže na kompleksnost TT.

Osnovni problem pri neuspešnem računalniškem izvajanju TT je v tem, da računalniki nimajo praktično nič ključnih človeških lastnosti kot razumevanja, semantike, zavesti itd. Če postavimo dva vprašanja in se drugi navezuje na razumevanje prvega, računalnik le po slučaju ugane odgovor. V primeru več možnih odgovorov pa po zakonu verjetnosti skoraj vedno zgreši. Za vsak slučaj je potem potrebno postaviti še tretje vprašanje in dokaj verjetno je računalnik razkrinkan.

Podobno je tudi vabljeni predavatelj Koch [8], s svojo enačbo pokazal, da računalniki zaradi svojega HW niso sposobni biti zavestni. Zato je TT s svojimi rezultati pokazatelj, da AI programi kljub svoji izredni uspešnosti še vedno niso sposobni izvajati najbolj ključnih človeških lastnosti, skoraj nič bolje kot pred desetletji. Drži tudi nasprotno: Test, ki bi kazal napredek pri razvoju ključnih človeških lastnosti kot zavesti ali pomenu, bi bil nerealno zavajanje.

In odgovor na vprašanje, kakšen bi bil svet brez Turingovega testa? Podobno kot o znanju o črnih luknjah nimamo neposredne koristi. Hkrati pa bi ob nepoznavanju TT in recimo ob čedalje bolj uspešnem igranju šaha mislili, da so računalniki pametnejši kot ljudje, pri reševanju IQ testov pa bi bili nekje primerljivi. Sklepali bi, da so računalniki pametnejši kot ljudje, kar pa bi bilo povsem nerealno. To bi bilo tako zavajajoče, kot če bi razvijali boljše in boljše vzmeti za potovanje v vesolju in se veselili vsakega novega metra, pa čeprav obstoječa tehnologija ne bi niti teoretično omogočala odlepitev od našega planeta. Zaradi TT pa vsaj vemo, da bo potrebno razviti nove metode, da se bodo računalniki lahko približali ljudem tudi v teh sposobnostih.

Za primer ene zadnjih »rešitev« Turingovega testa navedimo primer poročanja medijev iz leta 2014: “The 65-year-old iconic Turing Test was passed for the very first time by the computer programme Eugene Goostman during a Turing Test in 2014, held at the renowned Royal Society in London on Saturday. 'Eugene' simulates a 13-year-old boy and was developed in Saint Petersburg, Russia.” Mediji so povzeli trditve organizatorja Kevin Warwick, da je računalnik uspešno rešil Turingov test, ker ga ljudje niso prepoznali (http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx). Čeprav je bilo nekaj prisotnih in nekaj pozneje vprašanih strokovnjakov ogorčenih, se je šele čez nekaj časa pojavil znanstveni članek, ki je podrobno razložil vse zvijače pri izvajanju testa. Za nekoga so morda te zvijače sprejemljive, vendar s tem postavljajo tak test na neverodostojno raven. Ko so npr. program, ki je igral trinajstletnika iz Ukrajine vprašali kaj zahtevnega, se je izgovoril, da ne zna angleško, da je imel težko otroštvo itd. Izpraševalci si niso mogli pomagati, da ne bi začeli sočustvovati z Eugenom. Pri tem so spregledali tako očitne nesmisle kot vprašanja, koliko nog ima konj ali muha – v obeh primerih je program odgovoril 4-6.

V angleški Wikipedii je napisano takole: “The validity and relevance of the announcement of Goostman's pass was questioned by critics, who noted … the bot's use of personality quirks and humor in an attempt to misdirect users from its non-human tendencies and lack of real intelligence…”. Several other scientists were appalled by the level of mass-media propelling self-propagating individuals.

To je značilni »nevtralni« stil poročanja razvitega zahodnega sveta, namesto da bi konkretno povedali, za kaj gre. Recimo: takole: "It's nonsense," said Prof, Stevan Harnad. “We are not even close." Zdi se neverjetno, kako so lahko mediji in nekateri recimo »zmerni« strokovnjaki s področja računalnika lahko povzeli vznesene nerealne trditve. Naj še enkrat povsem jasno povemo, da je prišlo do formalne rešitve testa zato, ker je bilo dovoljeno programu »zavajati«, tj. početi nekaj prepovedanega.

Lahkota, s katero so zavajali testne izpraševalce, je sprožila raziskovalno eksperimentalno tezo tega prispevka: **Če bomo med naključno mimo hodeče ljudi postavili negibno lutko, ali jih bo večina takoj prepoznala?** Dodatna vprašanja so bila npr.: **Ali bodo nekateri posamezniki zavedeni za dalj časa? Kako bodo reagirali ljudje ob stiku z lutko?**

3. TURINGOV TEST Z LUTKO NA IJS

Test smo izvajali na Institutu »Jožef Stefan« s pomočjo 160 cm visoke lutke mladoletnice. Lutka je bila androidna s premično žico v sredini udov, s prsti, z žametno površino. Celo oči je imela žametne. Na glavo smo ji namestili lasuljo, v prvem delu črno s prameni in v drugem temno rjavo. Na nogah je imela obute čevlje. Preoblekli smo jo v nekaj preoblek, večino prvega testa je bila v obleki kot na sliki 1, večino drugega testa pa v obleki kot na sliki 2. Na obeh lokacijah je imela lutka skrit obraz (naslonjen na steno) in skrite prste – roke je stiskala pod obrazom v steno. Ker je bila nagnjena rahlo naprej, je bila kljub majhni teži dokaj stabilna, je pa dokaj hitro padla ob dregljajih s strani. Testiranje je potekalo tako, da je bila lutka trajno (24/7) nameščena na lokaciji, testiranci so anonimno vpisovali priložene anketne liste, izvajalci testa pa smo občasno opazovali spremembe na lutki in odzive zaposlenih in obiskovalcev. Snemanje ni bilo dovoljeno zaradi kršitve privatnosti. Testiranci so anketne liste izpolnjevali brez prisotnosti izvajalcev testa, da ne bi prišlo do kakršnegakoli vplivanja.

Prva lokacija: nasproti stranišča v prvem nadstropju centralne stavbe Instituta »Jožef Stefan«, Jamova 39, Ljubljana (slika 1). Lutka je bila obrnjena proti omari, tako da so jo videli mimoidoči oz. obiskovalci moškega in ženskega stranišča.

Druga lokacija: pri kavomatu oz. vodomatu, tako da so se ji željni vode ali kave približali in jo opazili (slika 2). Ta lokacija je na istem hodniku kot prva, tako da bi se lutki lahko videli med seboj, če bi bili živi.

Testiranje je potekalo od konca julija do konca avgusta, tj. večinoma v času poletnih dopustov, ko število prisotnih občutno upade.

Anketne liste je oddalo 125 posameznikov. Povprečna starost anketirancev je bila 37 let, od 15 do 82, s tem da je bila velika večina med 20 in 50 leti. Moških je bilo 68 in žensk 53, torej je bila populacija dokaj uravnotežena.

24 jih je dobro poznalo, 38 jih je poznalo Turingov test, 61 ne. Alana Turinga je dobro poznalo 21, poznalo 48 in ni poznalo 33 posameznikov. Približno tri četrtine anketirancev je vsaj poznalo Turinga in polovica Turingov test, kar nakazuje, da so bili anketiranci nadpovprečno računalniško izobraženi. V tem delu inštituta se sicer nahaja nekaj odsekov od kemije do fizike, hkrati pa je približno polovica računalniško oz. elektronsko usmerjenih odsekov.

Dokončano osnovno šolo sta imela dva anketiranca, srednjo 18, fakulteto 68 in doktorat 31, torej je 79 % imelo vsaj fakultetno izobrazbo.

Lutko je takoj prepoznalo 33 oseb (26%), ni pa je prepoznalo 19 oseb (15%).

Slika 1. Značilna obleka in lokacija prvega dela testa.

Slika 2. Značilna obleka in lokacija drugega dela testa.

Povprečni čas prepoznavanja (ob prepoznavi) je bil 24,3 sekunde, kar potrjuje tezo, da so ljudje nagnjeni k površnemu opazovanju in interakciji z okolico oz. drugimi ljudmi. Ker je 20 anketirancev porabilo manj kot 5 sekund, kolikor časa potrebujemo npr., da opazimo negibnost lutke, je očitno naloga preprosto rešljiva, če človek usmeri pozornost na lutko. Pod desetimi sekundami je lutko kot neživo prepoznalo 36 anketirancev, kar je 28.8% vseh. Najdaljši časi so se sukali okoli nekaj minut, pri čemer ni jasno, kako bi lahko anketiranci pili kavo pri kavomatu nekaj minut.

Programi iz Weke in statistična analiza, npr. Kruskal Wallis test za ugotavljanje statstično pomembnih razlik med skupinami, so pokazali naslednje odvisnosti (smiselno zbrano skupaj): lutko so bolje prepoznavali tisti z višjo izobrazbo in boljšim poznavanjem Alana Turinga in Turingovega testa. Ženske s slabim poznavanjem Turingovih tez so slabše prepoznavale kot moški, tudi tisti z nižjo izobrazbo. Moški in ženske se statistično pomembno razlikujejo v tem, ali so prepoznali, da gre za lutko, ali ne ($\chi 2$ (1) = 5,55, p = 0,02) – več moških je prepoznalo, da gre za lutko v primerjavi z ženskami. Ni statistično pomembne povezanosti med izobrazbo (osnovna šola in srednja šola vs. fakulteta in doktorat) in tem, ali so udeleženci prepoznali, da gre za lutko ali ne (χ^2 (1) = 0,49, p = 0,48).

Z Weko zgrajena drevesa za hitrost prepoznavanja so imela tipično med 20 in 30 listov.

Razlogi za prepoznavanje lutke so navedeni na sliki 3. Pričakovano izstopa negibnost, manj pričakovani so lasje, dimenzije in drža. Vse ostalo (obleka, drugo, opozorjen) ni posebej izstopalo.

Med posebej zanimivimi so bili odzivi oz. interakcija ljudi na lutko. Na začetku eksperimenta je prvi avtor ustavil sodelavca, ki je začel nositi lutko naokrog. Nekateri mimoidoči so se prestrašili, ko so ugotovili, da gre za neživo lutko. Kar nekaj sodelavcev inštituta je sporočilo izvajalcem testa, da se na hodniku že nekaj dni nahaja oseba v očitno krizni situaciji in ali je vse v redu z njo. Naka obiskovalka kavomata je po nesreči dregnila lutko in lutka je padla nanjo in povzročila temu primeren šok. Posebej v drugem delu testa smo opazili, da lutko premikajo, ji mršijo obleko, jo prijemajo itd.

Drugi zanimiv zaključek, ugotovljen v diskusiji med izvajalci testa in sodelavci, je, da ljudje s perifernim vidom najprej dojamemo človeku podobno figuro in čeprav razumsko/spominsko vemo, da je ta lutka, je prva misel: »Tam je nekdo« in nato »Pa saj vem, da je to lutka«. Zanimiv je tudi odnos izvajalcev testa in sodelavcev, ki so lutko poimenovali »Angelca« oz. »Angie«.

4. DISKUSIJA

V prvem delu prispevka smo zbrali dogajanja v zvezi s Turingovim testom v zadnjih nekaj letih. V drugem prispevku smo raziskovali hipotezo, da smo ljudje dokaj površni glede opažanja okolice. Teste smo izvedli z lutko. Le približno četrtina (26%) anketirancev je takoj, tj. v nekaj sekundah prepoznala lutko. Kar 16% jih lutke sploh ni prepoznalo, čeprav so bili anketiranci večinoma v najboljših letih in statistika ni pokazala nobene odvisnosti med leti in uspešnostjo prepoznavanja.

Pokus je torej potrdil tezo, ki v prenesenem smislu pomeni, da je ključni komponent Turingovega testa njegova korektna izvedba, tj. taka, ki preprečuje zavajanja in odvračanje pozornosti.

Poskus bomo nadaljevali z bolj napredno verzijo lutke z dodanimi drugimi napravami kot avtonomnimi vozili in droni.

Zahvala:

Zahvaljujemo se vsem anonimnim anketirancem za sodelovanje v anketi.

5. REFERENCE

[1] Turing, A. M. 1937. [Delivered to the Society November 1936]. "On Computable Numbers, with an Application to the Entscheidungsproblem" (PDF). *Proceedings of the London Mathematical Society*. 2 42. pp. 230–5.doi:10.1112/plms/s2-42.1.230.

[2] Turing, A. 1950. Computing Machinery and Intelligence. *Mind*, 59, 433-460.

[3] Hingston, P. 2007. A Turing Test for Computer Game Bots. *IEEE Transactions onComputational Intelligence and AI in Games*. (Volume:, Issue: 3). p. 169-186. doi: 10.1109/TCIAIG.2009.2032534

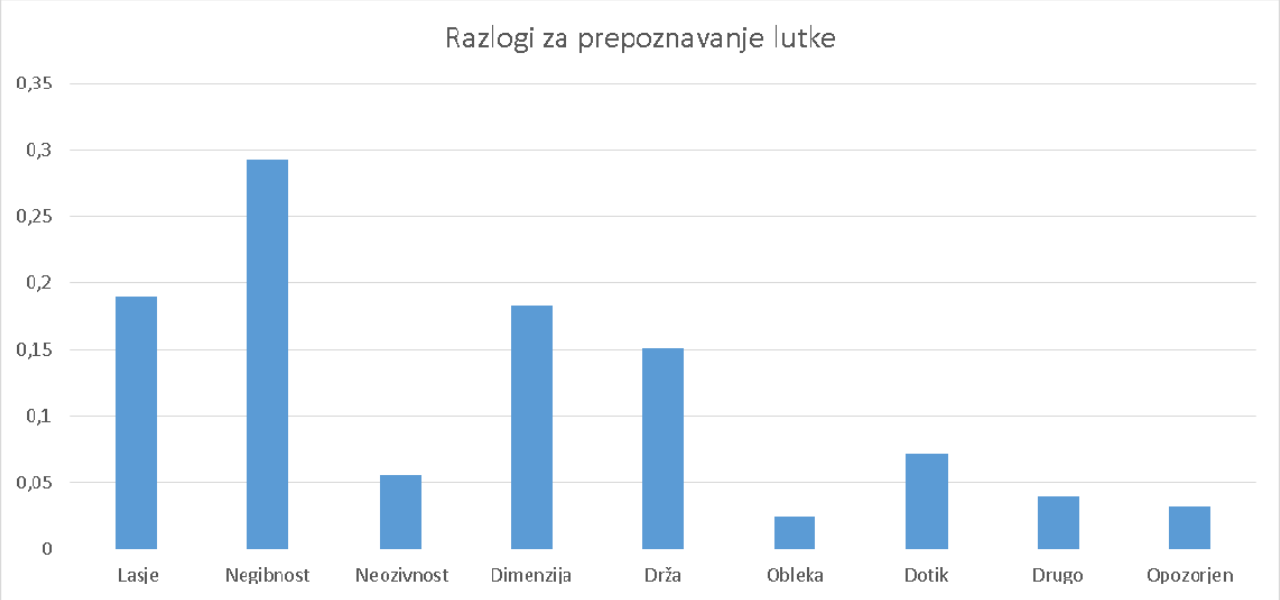
[4] Bostrom, N. 2014. *Superintelligence – Paths, Dangers, Strategies.* Oxford University Press, Oxford, UK.

[5] Moor, J.H. 2003. Turing test. In *Encyclopedia of Computer Science 4th edition*, p. 1801-1802. John Wiley and Sons Ltd. Chichester, UK. ISBN:0-470-86412-5

[6] Penrose, R. 2002. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford.

[7] Gams, M. 2001. *Weak intelligence : through the principle and paradox of multiple knowledge*. Nova Science, 2001.

[8] Tononi, G.and Koch, C. 2015. Consciousness: here, there and everywhere. *Philos Trans R Soc Lond B Biol Sci.* 19;370(1668). pii: 20140167. doi: 10.1098/rstb.2014.0167.



Slika 3. Razlogi za prepoznavanje lutke: neodzivnost, lasje, dimenzije, drža itd.

Superintelligence

Matjaž Gams
Jožef Stefan Institute
Jamova 29, Ljubljana, Slovenia
matjaz.gams@ijs.si

ABSTRACT
The paper discusses concepts of superintelligence – related to the time when artificial intelligence systems will achieve and bypass human-level intelligence. The core literature is the book by Nick Bostrom, while several other sources are partially included.

Keywords
Surpassing human brains, future prediction, artificial intelligence

1. INTRODUCTION
Recently, the world attention was focused on dangers of artificial intelligence becoming superintelligent. The implications of that event are tremendous: predictions vary from extinction of human race to the most prosperous and advanced times in human civilization. One of the outstanding publications on this issue is the 2014 Oxford University Press book "Superintelligence – Paths, Dangers, Strategies" by Nick Bostrom [1]. It describes issues and concepts relevant for the time when machine brains will achieve and surpass human brains in general intelligence.

Elon Musk, the billionaire chief executive of SpaceX and Tesla Motors and a techno-optimist in relation to general technical progress such as solar power, space exploration and electric cars, is one of most world-prominent AI pessimists [2]. At various occasions he expressed his concerns that superintelligent machines will in relative short future pose a real threat to human existence.

Another best-respected thinker is Stephen Hawking, renowned for his work in physics such as on black holes and gravitational singularities. He has also become known for various interesting ideas expressing a live and dynamic mind in a decaying body. Namely, he suffers from a motor neuron disease similar to amyotrophic lateral sclerosis, which left him paralyzed and unable to speak without a voice synthesizer. Among others, he often predicted dangers to human civilization. In [3] he describes three potential major dangers: artificial intelligence, humans, aliens. The last two are pretty well known while the first one attracted most attention. "The development of full artificial intelligence could spell the end of the human race," Hawking told the BBC in December 2014. It has been said that the reason for the fear was Stephen's surprise how well the new text synthesizer proposed forthcoming words when dictating new texts. One might argue that was just good statistics, user-adapted and calibrated on his previous texts. However, his concern is real and is based on overall progress of artificial intelligence (AI) that is generally not disputed.

"Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls," several researchers including Task and Hawking wrote in the letter, published online Jan. 11 2015 by the Future of Life Institute, a volunteer organization that aims to mitigate existential threats to humanity.

When observing the two of most prominent critics of AI, one cannot help to think that there must indeed be something on it if so different thinkers with such different pedigree warn of the same issue. One can also add Bill Gates and several prominent scientists to Elon Musk and Stephen Hawking. However, majority of AI and scientists reject these warning as premature or even not founded at all.

"We are decades away from any technology we need to worry about," Demis Hassabis from Google DeepMind told reporters when presenting his program that can teach itself to play computer games.

Is superintelligence a true danger, how to prevent it, which are the new relations and circumstances that will emerge during the progress of superintelligence, might it not happen at all – these are the relevant questions this paper deals with.

Figure 1. Elon Musk, a techno-optimist and CEO, warns about dangers of fast-progressing artificial intelligence.

Figure 2. Stephen Hawking is one of the most established world thinkers warning about dangers of superintelligence.

2. TURING-RELATED BACKGROUND

There are several technical background issues worth analyzing before the superintelligence itself. The first one is what the maximal computational power computers can possess.

The Turing machine was invented in 1936 by Alan Turing [4]. It is a model of computation and not a description of a computer or a computing machine. As shown by Turing themselves and later formally by Church, the tasks (problems) one has to compute belong to specific classes, and can be sorted in terms of complexity. In simple words: the Turing machine is capable of solving all "realistically" solvable problems in our world. Therefore, humans are as computationally "strong" as computers, be it digital, analog, quantum etc. If that is the case, then computers cannot be stronger than humans, at least in principle. They can be faster, can have bigger memory capacities etc., but can only solve the problems the humans can, and no more. That written, it should be fair to notice that there are several hypothetical mechanisms that exceed the universal Turing machine (UTM); they can compute in principle harder problems no UTM can. It is a bit unclear whether these stronger machines are also stronger than humans since several authors such as Penrose [5] claim that humans are in principle stronger than the UTMs.

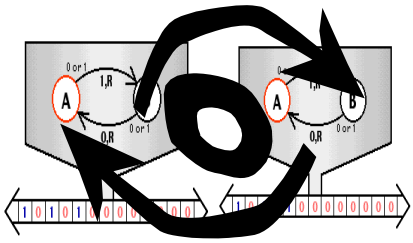


Figure 3. The principle of multiple knowledge is proposed as a limit for computational power of current computers.

The author of this paper introduced the Multiple Turing Machine already in 1997, and formally presented it in a book in 2001 [6]. The schema is presented in Figure 3. Unlike multi-tape Turing machine it consists of two universal Turing machines, each writing on each other program at the same time getting information from the open outside world. It is declared that no "real" or "human-level" intelligence can be achieved without using at least the multiple algorithms applying the principle presented in Figure 1. Please note that we can design programs and computers that compute in a seemingly similar way, but no successful program or system has been designed in this way and there is no indication that we are progressing in this direction.

Furthermore, if one looks at the Turing tests [7], the test that admits intelligence if an average human is not capable of distinguishing a human from a computer when communicating through written questions and answers, it becomes clear that an average time needed to damask a computer program playing a human has remained constant during recent years despite enormous progress in terms of computer hardware and AI methods. Therefore, theoretically it seems that the AI fear has no scientific grounds.

Looking at the media, one gets quite an opposite impression. On 7 June 2014, at a contest marking the 60th anniversary of Turing's death, the world media reported that a historic event happened: a program passed the Turing test. In that contest, 33% of the event's judges thought that a chatbot named Goostman was human, and the event's organizer Kevin Warwick considered it to have passed Turing's test. Namely, Turing's prediction [7] was that by the year 2000, machines would be capable of fooling 30% of human judges after five minutes of questioning. According to Wikipedia: "The validity and relevance of the announcement of Goostman's pass was questioned by critics, who noted … the bot's use of personality quirks and humor in an attempt to misdirect users from its non-human tendencies and lack of real intelligence…". Several other scientists were appalled by the level of mass-media propelling self-propagating individuals. "It's nonsense," said Prof Stevan Harnad. "We are not even close." Indeed, in later publications one could clearly see that the misdirecting of human judges was based on the chatbot ability to diverge communication or play language problems when asked even such simple questions such as "How many legs does a horse have?". Is it the case that mass media again propagates a fake doom just to attract readers?

3. THE SUPERINTELLIGENCE BOOK

Whatever the case, it is clearly a fact that AI is fast progressing and that in several domains such as most games like chess, computers clearly outperform the best humans even on simple PCs. The number of tasks AI programs perform better than humans is growing faster and faster. In this Section, the Superintelligence book [1] will be presented and cross-checked.

Excerpts from Wikipedia: "… Bostrom reasons with "cognitive performance greatly [exceeding] that of humans in virtually all domains of interest", superintelligent agents could promise substantial societal benefits and pose a significant existential risk of artificial general intelligence. Therefore, it is crucial that we approach this area with caution, and take active steps to mitigate the risks we face. … research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today."

Niklas Boström was born 10 March 1973 and later transformed his name into more popular Nick Bostrom. He is a Swedish philosopher at the University of Oxford, holding a PhD from the London School of Economics, and the founding director of the Future of Humanity Institute at Oxford University. His work deals with the existential risk, the anthropic principle, superintelligence risks, the reversal test, and human mental enhancement risks. His work on superintelligence has influenced both Elon Musk's and Bill Gates's concern due to AI progress. Among others, Boström was listed in Foreign Policy's Top 100 Global Thinkers list.

In the Preface of the book [1], Bostrom makes a comparison between humans and gorillas: "the fate of our species would depend on the actions of the machine superintelligence." But, … "We do have one advantage: we get to build the stuff." …"In this book, I try to understand the challenge presented by the prospect of superintelligence, and how we might best respond. This is quite possibly the most important and most daunting challenge humanity has ever faced."

The book consists of the following paragraphs:

1. Past developments and present capabilities – here a short overview of past and state of the art is presented, e.g. growth of human population and game-playing AI.
2. Paths to superintelligence – Bostrom agrees that the current computer intelligence is far inferior to humans

in general intelligence, yet it assumes that superintelligence is possible. In order to achieve it, multiple paths are possible thus increasing the possibility. The Moore's law is presented as a law promising to continue in reasonable future thus enhancing AI progress. Not much hope is given to merging of human and computers into "one being".

3. Forms of superintelligence – here he proposes three forms of superintelligence. Again, it is proposed that human or organic intelligence has many disadvantages compared to the computer intelligence in terms of basic hardware. The three forms of superintelligence are: **speed** superintelligence that can perform the same functions as humans, just orders of magnitude faster; **collective** superintelligence that achieves superintelligence through a large number of individual computing entities, connected into one "mental organism"; **quality** superintelligence that is at least as fast as a human mind and vastly qualitatively smarter.

4. The kinetics of an intelligence explosion – this section deals with the rate of AI progress – will be it exponential, linear, saturating etc.? Anyway, if achieved at all, it will be either exponential or saturating, in the middle or in the last third of the S-curve.

5. Decisive strategic advantage – will there be one single superior superintelligence dominating all others or there will be many, a front progressing?

6. Cognitive superpowers – Bostrom proposes a scenario in which superintelligence emerges from "normal" software and dominates the world. A hidden assumption is that such an event is possible without well-grounded arguments. For example, even the most intuitive assumption that superior intelligence will, once created, find its way to dominance of the real world, seems quite in opposition to the embedded AI which assumes that not solely SW can even achieve true AI. If one compares chimpanzees, evidently physically stronger, but mentally inferior to humans, one should have in mind that the human predecessor battling with chimpanzees was comparably similarly strong, but mentally a bit more advanced. Had it been the case that the predecessor was, say, 10 times weaker than the chimpanzee and 1000 times smarter, the chimpanzee would still win the competition.

7. The superintelligence will – what will be the goals of superintelligent agents (Bostrom refers to superintelligent entities with various terms, often as "artificial agents")? The orthogonality thesis holds that intelligence and goals are independent. The instrumental convergence thesis holds that superintelligent agents will have somehow common reasons and goals.

8. Is the default outcome doom? – there is quite common opinion that more civilized beings will be more friendly and non-harmful. Bostrom claims that there is no clear evidence for such reasoning in terms of superintelligence, that the malignant and malice behavior is possible even in most superior superintelligence. Since the author of this paper often claimed that between computers and humans there is no hostile relation because they are different and because they need each other and because they do not compete for the same sources such as hyenas and lions competing for the same pray, this might seem a bit contradictive. Yet, one cannot eradicate even the darkest scenarios and we humans have achieved the world dominance due to our capability to estimate and evaluate all possibilities. Therefore, we should also reevaluate the dark superintelligence scenario, although it is not very likely on its own.

   For Bostrom, a unipolar, i.e. a single superintelligence establishing a singleton is a particularly menacing scenario.

9. The control problem – how to avoid the negative outcome? There are two possible control situations the artificial superintelligent agents will face, and Bostrom discusses several techniques inside each.

10. Oracles, genies, sovereigns, tools – Bostrom analysis these four types of agents in the light of dangers and possibilities they present for the superintelligence emergence. For example, it might be naïve to think that the superintelligent genie listening to our commands and executing them would indeed comply with all our commands.

11. Multipolar scenarios – this paragraph deals with the less menacing situation with several competing agencies in a society after the transition. In particular he discusses the control problem in two subcases. Along the way Bostrom discusses other issues in the future world. One of them is demographic growth. In the last 9,000 years, a fraction of the modern human history, the human population has grown thousandfold. There have been many multiplying effects, probably also the growth of the humanity's collective intelligence. The Malthusian principle seems out of scope since a large majority of all people in the world have more incomes than needed for survival. This growth in standard can be evidenced in several parameters. For example, the US horse population has fallen to 2 million in 1950s, but is now around 10 million. Adding advanced AI programs on the top, probably robots as well, what will be the consequences for human civilization? Overpopulation, mass unemployment, unconscious workers, … Since the human population will not be binded by typical constraints like food or work necessities, several types of societies are possible, most of them unseen and not possible till that time. Along reasoning in the book, a prevalent sense of danger appears. After all, Bostrom already published papers and books about catastrophes, for example [7], where editors Bostrom and Ćirković characterize the relation between existential risk and the broader class of global catastrophic risks, and link existential risk the Fermi paradox. The later concerns the lack of communication with alien intelligence since according to the Drake equation (there is a huge number of planets in the universe), there should be lots of civilizations inside the reach of our sensors.

12. Acquiring values – since it is probably impossible to control superintelligence, it would be more sensible to

get proper values, beliefs and desires into intelligent agents to be friendly to us (at least to the level we are with chimpanzees). How to do that?

13. Choosing the criteria for choosing – which basic values, related to decision theory and epistemology should humans load into the growing superintelligence?

14. The strategic picture – in this chapter general recommendations are recapitulated such as to retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risks and accelerate those that decrease the existential risks.

15. Crunch time – we have a deadline in the form or future emergence of superintelligence. What to do in a thicket of strategic complexity, surrounded by a dense mist of uncertainty?

4. CONSLUCIONS AND DISCUSSION

Is Bostrom just another doomsday advocate calling for world-vide attention? Very probably he was the originator of the last wave of AI apocalypse hype that turned out in the world media as another cheap horror story. Yet, reading the Superintelligence book one quickly captures the intellectual power and knowledge of Bostrom. He might be a philosopher, unpopular in technical fields, but with interesting ideas and very broad views. Unlike mass media, Bostrom should be taken seriously (up to a point). Furthermore and without any dilemma, some concerns are better raised sooner than later.

But is his thesis technically sound in the eyes of an AI researcher such as the author of this paper? Let us quickly summon his main thesis: He thinks that we can make some predictions about the motivations of superintelligence in the form of one or many superintelligence artificial agents. Superintelligence will emerge and we will not be able to control it as chimpanzees are not able to control humans, therefore our main chance is to make the superintelligence human friendly. This sets up challenges for us in advance to figure out ways to frame and implement motivational programming an AI before it gets smart enough to resist future changes. Being a philosopher, he thinks that philosophers are in a great position for well-informed speculation on topics like this.

AI-ers know how hard it is to achieve true intelligence. We have been trying to achieve it for the last 50 years on computers, and we are nearly as far from it as at the start, at least taking into account the Turing tests. Of course, AI is amazingly fast progressing and it is improving the lives of humans and the power of human civilization immensely.

So far, the negative effects of AI have been negligible, positive contributions exceptional while so far all the doomsday predictions turned false.

Most likely, Bostrom's AI warning is inside the doomsday category if one treats his contributions according to the mass media. But reading the book itself, one finds several interesting studies and derivations.

Still, majority of future studies provide tons of indications that the AI progress is crucial for the human civilization advances. In fact, it is very probably the best hope of human further progress. A superintelligence revolution (the singularity theory of Kurzweil) will most likely raise the human society to the levels unseen and exceeding our expectations and desires.

There are also several indications that without advanced AI the human civilization might face hard times. To mention just one: if there are not alien civilizations inside our sensor range, it might be because something is destroying them. And it for certain is not AI since even in some of the weirdest nearly impossible circumstances where superintelligence would erase human race, it would take the role of the advanced civilization on Earth. As we can now deter the danger of asteroids falling on Earth due to advanced space technologies, some day it will be the superintelligence that will help us deter the potential existential threats. The sooner we have it, the better!

5. REFERENCES

[1] Bostrom, N. 2014. *Superintelligence – Paths, Dangers, Strategies.* Oxford University Press, Oxford, UK.

[2] Musil, S. 2014. Elon Musk worries AI could delete humans along with spam. *Sci-Tech*. http://www.cnet.com/news/elon-musk-worries-ai-could-delete-humans-along-with-spam/

[3] Lewis, T. 2015. Stephen Hawking Thinks These 3 Things Could Destroy Humanity. *Livescience.* http://www.livescience.com/49952-stephen-hawking-warnings-to-humanity.html

[4] Turing, A. M. 1937. [Delivered to the Society November 1936]. "On Computable Numbers, with an Application to the Entscheidungsproblem" (PDF). *Proceedings of the London Mathematical Society*. 2 42. pp. 230–5.doi:10.1112/plms/s2-42.1.230.

[5] Penrose, R. 2002. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford.

[6] Gams, M. 2001. *Weak intelligence : through the principle and paradox of multiple knowledge*. Nova Science, 2001.

[7] Turing, A. 1950. Computing Machinery and Intelligence. *Mind*, 59, 433-460.

[8] Bostrom, N. and Ćirković, M.M. (eds) 2008, *Global Catastrophic Risks*, Oxford University Press, Oxford. ISBN-13: 978-0199606504

[9] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

Recognizing atomic activities with wrist-worn accelerometer using machine learning

Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, Matjaž Gams
Mednarodna podiplomska šola Jožefa Stefana
Odsek za Inteligentne Sisteme, Institut Jožef Stefan
martin.gjoreski@ijs.si

ABSTRACT
In this paper we present a machine learning approach to activity recognition using a wristband device. The approach includes: data acquisition, filtering, feature extraction, feature selection, training a classification model and finally classification (recognizing the activity). We evaluated the approach using a dataset consisting of 10 everyday activities recorded by 10 volunteers. Even though the related work shows that with a wrist-worn device one should expect worse accuracy compared to devices worn on other body locations (chest, thigh and ankle), our tests showed that the accuracy is 72%, which is slightly worse compared to the accuracy of the thigh (82%) and ankle-placed devices (83%); and slightly better compared to a chest-placed device (67%). Additionally, by applying feature selection and increasing the window size, the accuracy increased by 5%.

Keywords
Activity recognition, wrist, accelerometer, machine learning, classification, feature extraction.

1. INTRODUCTION
With the recent trends and development in sensor technology (miniaturization – MEMS; connectivity - Bluetooth low energy and WiFi; battery - Li-ion) people get used to the idea of: wearing an additional device on themselves beside the telephone, or replacing an existing one – the wristwatch. These devices provide sensor data which can be used for extracting useful information about the user: how many calories are burned during the day, what types of activities are performed during the day (sedentary vs. dynamic ones), detecting alarming situations (e.g., falls), detecting behavioral changes, and similar. A service that emerged as an essential basic building block in developing such applications is Activity Recognition (AR). The activity of the user provides reach contextual information which can be used to further infer additional useful information [1][2][3]. Wristband devices are becoming popular mainly because people are more or less accustomed to wear watches and therefore this placement is one of the least intrusive placements to wear a device. Nowadays we are witnessing various types of fitness/health oriented wrist-worn devices, such as: FitBit[1], Empaica[2], Microsoft band[3]; and also in the last few years smartwatches are gaining attraction: Apple watch, Android wear wristwatches, Samsung Galaxy gear, etc.

[1] www.firbit.com

[2] www.epatica.com

[3] http://www.microsoft.com/microsoft-band/en-us

In this paper we present a machine learning approach to activity recognition using a wristband device. The approach includes: data acquisition, filtering, feature extraction, feature selection, training a classification model and finally classification (recognizing the activity). It was evaluated on a dataset consisting of 10 everyday activities recorded by 10 volunteers. The results showed that with a wrist-worn device one can recognize much more activities than what is commonly used for (i.e., walking - step counter), running, lying - sleeping). Additionally, the accuracy is comparable even in some cases higher compared to devices worn on other body locations (chest, thigh and ankle), which are more established and commonly used for activity recognition tasks.

2. RELATED WORK
The most recent literature in AR field shows that wearable accelerometers are among the most suitable sensors for unobtrusive AR [7]. Accelerometers are becoming increasingly common because of their lowering cost, weight and power consumption. Currently the most exploited and probably the most mature approach to AR is with wearable accelerometers by using machine learning approach [18][16][17]. This approach usually implements widely used classification methods, such as Decision Tree, SVM, kNN and Naive Bayes.

For the sake of the user's convenience, AR applications are often limited to a single accelerometer. Numerous studies have shown that the performance of an Activity Recognition System strongly depends on the accelerometer placement (e.g., chest, abdomen, waist, thigh, ankle) and that some placements are more suitable (in terms of AR performance) for particular activities [4][6][5].

In the past the wrist was the least exploited placement for AR. Mainly because of our inclination towards frequent hand movements which negatively influence an AR system. The researchers usually were testing chest, waist, thighs (left and right) [18][19], ankles (left and right) and neck. The results vary a lot and cannot be compared through different studies (different datasets, different algorithm parameters, different approaches, etc.). In our previous work we also tested most of these locations on two datasets. On the first one, the results showed that all of the locations perform similarly achieving around 82% accuracy [8]. On the second dataset, where the experiments were more thorough (bigger dataset, improved algorithms) the results showed that thigh and ankle perform similarly (82% and 83% respectively) and achieve higher accuracy compared to the chest (67%) [9].

However, with the penetration of the wrist-worn fitness trackers and smartwatches, it is to be expected that wrist sensor placement will be quite researched area. Recently, Trust et al. [12] presented a study for hip versus wrist data for AR. The models using hip

data slightly outperformed the wrist-data models. Similarly, in the study by Rosenberg et al. [13] for sedentary and activity detection, the models using hip data outperform the wrist models. In the study by Manini et al. [14] ankle data models achieved high accuracy of 95.0% that decreased to 84.7% for wrist data models. Shorter (4 s) windows only minimally decreased performances of the algorithm on the wrist to 84.2%. Ellis et al. [15] presented an approach for locomotion and household activities recognition in a lab setting. For one subset of activities the hip-data models outperformed the wrist data, but over all activities the wrist-data models produced better results. Garcia-Ceja et al. [20] presented person-specific activity detection for activities such as: shopping, showering, dinner, computer-work and exercise.

3. EXPERIMENTAL SETUP

3.1 Sensor equipment and experimental data

The sensor equipment consists of a Shimmer sensor platform. The sensors were placed on the chest, thigh, ankle and wrist with adjustable straps. The accelerometer data was acquired on a laptop in real-time via Bluetooth using frequency of 50 Hz. The data was manually labeled with the corresponding activity. Ten volunteers performed a complex 90-minute scenario which included ten elementary activities: lying, standing, walking, sitting, cycling, all fours, kneeling, running, bending and transition (transition up and transition down). These activities were selected as the most common elementary, everyday-life activities. In this paper, we are performing analysis only on the wrist-sensor data. The data from the other sensors (chest, thigh and ankle) has been extensively analyzed in our previous studies [6][7][9][10][11]. Nevertheless, the results presented in those studies provide valuable guidelines to which we are comparing. Overall, 1,000,000 raw-data samples per volunteer were recorded. These raw-data samples were transformed into approximately 7,000 data instances per volunteer. Figure 1 shows the instances' class distribution.

3.2 Experimental Method

Figure 2 shows the machine learning approach used in this research. It includes the following modules: data segmentation, data filtering, feature extraction, feature selection and building a classification model. The data segmentation phase uses an overlapping sliding-window technique, dividing the continuous sensor-stream data into data segments − windows. A window of a fixed size (width) moved across the stream of data. Once the sensor measurements are segmented, further pre-processing is

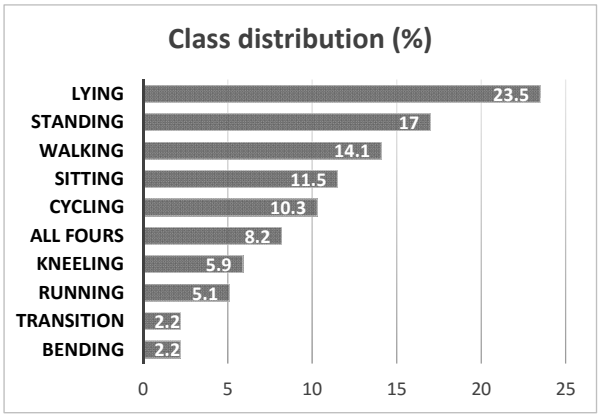


Figure 1. Class (activity) distribution

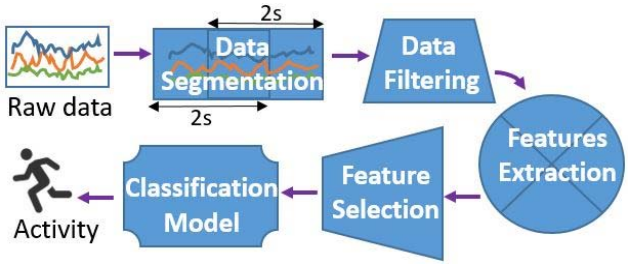


Figure 2. Activity recognition approach

performed using two simple filters: low-pass and band-pass. The feature extraction phase produces lowpass filtered features that measure the posture of the body, and band-pass filtered features that represent: the motion shape, the motion variance, the motion energy, and the motion periodicity [21] . The features extraction phase results in 53 extracted features. Since all of the features are extracted from one data source (wrist accelerometer), a high feature correlation is expected. For that reason the feature selection method is based on feature-correlation analysis which serves the purpose of removing correlated and "non-informative" features. Low informative features are considered those that have low information gain. The information gain evaluates the worth of a feature by measuring the information gain with respect to the class. Regarding the correlation of the features, we checked for Pearson's correlation, which measures linear correlation between features, and Spearman correlation, which measures how well the relationship between two variables can be described using a monotonic function. The feature selection steps are:

- Rank features by gain ratio.
- Starting from the lowest ranked feature, calculate its correlation coefficients (Pearson and Spearman) with each of the features ranked above. If it has a correlation coefficient higher than 0.95 with at least one feature, remove it.
- Repeat step 2 until 50% of the features are checked.

Figure 3 shows the results of the Person's correlation analysis before (left) and after (right) the feature selection phase. On the figure there are two correlation matrices, 53x53 (left) and (35x35) right. Each row (column) represents different feature. Red color represents negative, blue color represents positive and the intensity of the color represents the absolute value of the correlation. This figure on one hand depicts the dimensionality

reduction of 34% (from 53 features to 35 features), and on the other hand the correlation reduction (the intensities of the colors). On the left matrix some regions with high correlations are marked (with black rectangles) to present candidate features that the feature selection algorithm may delete. On the right matrix there is high correlation between some of the features even after the feature selection phase. These are features that have high gain ratio index. In each experiment we checked the accuracy with and without the feature selection phase. The experiments with feature selection phase achieved at least equal results and in some cases even slightly better results. Once the features are extracted (and selected), a feature vector is formed, and is fed into a classification model, which recognizes the activity of the user. The classification model is previously trained on feature vectors computed over training data. We tested several machine learning algorithms, Decision Tree, RF, Naive Bayes, and SVM with Leave-one-user out cross-validation.

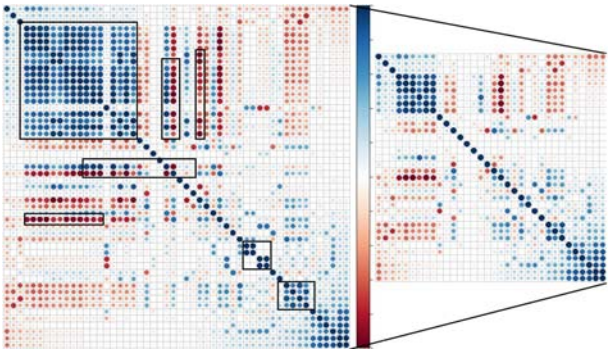


Figure 3. Person's correlation matrix before (left) and after (right) feature selection

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Wrist vs other sensor placement

First we wanted to know how well the machine learning models will perform when built using wrist-accelerometer compared to other body placements (ankle, chest and thigh). Here we present results without the feature selection phase in order to be comparable to our previous studies. Figure 4 shows accuracy comparison based on which sensor placement is used for building the machine learning model. For each study the same data is used with almost identical methodologies (same segmentation scheme, same number of features and same classifiers). From the figure we can see that for our dataset ankle or thigh sensor-placement provide better results than wrist and chest.

From now on we will report only on results achieved by the Random Forest (RF) classifier (which in not included in Figure 5 due to lack of information for the Ankle, Chest and Thigh accuracies) since with accuracy of 74% it performed best in our experiments.

Table 1 shows the confusion matrix, precision, recall and F1 score for each class obtained by the RF classifier. The F1 score for each of the activities shows that bending, kneeling and transition, are the three activities that are hard to recognize by the classifier. Standing and all fours are somewhat in the middle, whereas sitting, walking, lying, cycling and running are the activities that are recognized with a satisfying level.

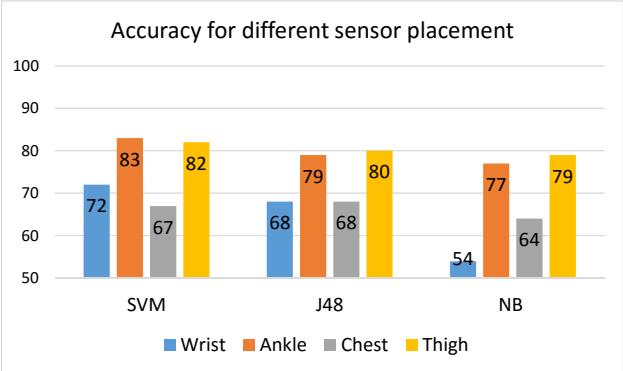


Figure 4. Accuracy for wrist vs other sensor placement

Table 1. RF confusion matrix and performance metrics (recall, precision and F1 score) per class

RF - Acc = 74%		1	2	3	4	5	6	7	8	9	10
TRUE	Walking -1	8428	1067	42	43	305	7	4	113	114	4
	Standing-2	298	8185	151	318	400	164	68	123	1381	1079
	Sitting -3	6	300	6256	1489	0	9	0	68	10	95
	Lying-4	2	303	1531	14889	61	21	0	38	71	90
	Bending -5	99	590	1	23	798	13	0	10	26	56
	Cycling-6	99	1110	0	8	117	5950	0	3	72	46
	Running-7	44	410	4	19	0	0	3157	2	7	1
	Transition-8	175	402	50	57	16	2	0	643	161	80
	All_fours-9	123	1140	6	128	72	77	6	65	4283	96
	Kneeling-10	52	1980	267	391	127	108	0	43	492	818
Recall		83	67	76	88	6	80	87	41	71	19
Precision		90	53	75	86	42	94	98	58	65	35
F1 score		87	59	76	87	11	87	92	48	68	25

4.2 Effects of window size on classification performance

In our activity recognition approach there is a "data segmentation" phase where an overlapping sliding-window is used for transforming the continuous data stream into a data segments over which the features are calculated. The reported results in the previous experiments (Section 3.1) are achieved using a window of 2s with an overlap of 1s. That means for predicting the activity at time T, we are taking accelerometer data starting from T-2s to T. The next prediction is at time T+1s and we are taking data starting from T-1s to T+1s, and so on. Basically we are predicting activity once per second by analyzing the data from the previous two seconds.

In these experiments we wanted to study the effects of the window size on the performance of the RF classifier. Moreover we wanted to see if choosing a shorter window can improve the accuracy of the short-duration activities, such as bending and transition, and the other way around (if choosing a longer window can improve the accuracy of the long-duration activities, e.g., standing). Table 2 shows the summarization of these experiments. The second row presents the window which is used for the experiments, starting from 1s window with 0.5s overlap, all the way to 10s window with 8s overlap. For each experiment the F1 score per activity and the overall accuracy of the RF classifier is reported. This table presents several observations:

- A short window of 1s with 0.5s overlap does not improve the performance of the classifier for the short-duration

activities. Better performance is achieved when longer windows are used.

- Only for the activity Running a shorter window (2s with 1s overlap) produces better performance (precision, recall and F1 score) than the other window lengths. For all the other activities the longer the window the better the performance of the classifiers.
- The overall accuracy increases by increasing the window. However, this increase is statistically important only for the increases from 1(0.5) to 2(1) and from 2(1) to 4(2). Also, the number of instances is highest for the windows size 1 (around 7000 per person), for window size 2 it is around 5000 per person, and for the rest of the window sizes the number of instances is equal i.e., around 3200 per person.

Table 2. RF classification performance for varying window

Random Forest		Data (overlap) window - seconds					
Metrics	Activity	1(0.5)	2(1)	4(2)	6(4)	8(6)	10(8)
F1 score	Walking	83.8	86.6	90.4	91.1	90.9	91.4
	Standing	55.3	60.0	64.4	65.6	66.0	66.8
	Sitting	71.2	75.1	75.9	77.9	77.4	77.4
	Lying	84.8	86.5	87.7	88.7	88.9	89.2
	Bending	12.7	12.2	13.1	14.4	14.7	14.4
	Cycling	81.4	85.0	88.4	89.1	89.8	90.0
	Running	96.2	97.2	97.1	97.0	96.7	96.9
	Transition	32.9	47.3	61.0	61.2	62.4	61.7
	All fours	63.6	69.0	71.4	72.9	73.4	74.7
	Kneeling	21.4	23.8	25.1	26.1	24.3	23.7
Accuracy		70.8	74.4	77.5	78.6	78.8	79.1

5. CONCLUSION

The high correlation between the features allowed for reducing the feature dimensionality by 34% (from 53 features to 35) while keeping the classifier performance. For that we removed features that have low information gain and high correlation.

Wrist accelerometer data produces slightly worse classifying performance than thigh and chest accelerometer data. The most problematic activities (from the 10 we analyzed) are bending, kneeling and transition. The results for the other activities are somewhat expected, except for the activity standing which is mixed by the classifier with almost all of the other activities. We hypothesize that during the data collecting scenario the volunteers were frequently moving their hands (while talking to each other), so the classifiers sees these hand movements as a movement of the whole body. Regarding the size of the window in the segmentation phase, it should be noted that for a longer window size the features are calculated over bigger data segments which may slightly increase the computational complexity. Window of 4s with 2s overlap may be the best tradeoff between computational complexity and classifier performance.

In these experiments each instance (activity) is treated independently of the previous activity, whereas in reality we rarely change our activity every 2s (2s is the predicting frequency for the highest achieved results - that is window size of 4, 6, 8 or 10 seconds). For future work we may use higher level features that provide information about the dependency of the instances [11].

Acknowledgements. The authors would like to thank Simon Kozina for his help in programming the AR algorithm. This work was partly supported by the Slovene Human Resources Development and Scholarship funds and partly by the CHIRON project - ARTEMIS JU, grant agreement No. 2009-1-100228.

6. REFERENCES

[1] Gregory, D.A., Anind, K. D., Peter J. B., Nigel, D., Mark, S. and Pete, S. Towards a better understanding of context and context-awareness. 1st International Symposium Handheld and Ubiquitous Computing, pp. 304-307, 1999.

[2] Vyas, N., Farringdon, J., Andre, D. and Stivoric, J. I. Machine learning and sensor fusion for estimating continuous energy expenditure. Innovative Applications of Artificial Intelligence Conference, pp. 1613-1620, 2012.

[3] Gjoreski, H., Kaluža, B., Gams, M., Milić, R. and Luštrek, M. Ensembles of multiple sensors for human energy expenditure estimation. Proceedings of the 2013 ACM international joint conference on Pervasive and Ubiquitous computing, Ubicomp, pp. 359-362, 2013.

[4] Atallah, L., Lo, B., King, R. and Yang, GZ. Sensor Placement for Activity Detection Using Wearable Accelerometers. In Proceedings BSN (2010), 24–29.

[5] Cleland, I., Kikhia, B., Nugent, C., et al. Optimal Placement of Accelerometers for the Detection of Everyday Activities. Sensors. 2013; 13 (7).

[6] Gjoreski, H., Luštrek and M., Gams, M. Accelerometer Placement for Posture Recognition and Fall Detection. 7th International Conference on Intelligent Environments (IE). 2011; 47–54.

[7] Gjoreski, H. et al. Competitive Live Evaluation of Activity-recognition Systems. IEEE Pervasive Computing, Vol:14 , Issue: 1, pp. 70 – 77 (2015).

[8] Gjoreski, H. Master Thesis, 2011. Adaptive Human Activity Recognition and Fall Detection Using Wearable Sensors. Jozef Stefan International Postgraduate School.

[9] Kozina, S., Gjoreski, H., Gams, M. and Luštrek, M. Three-layer activity recognition combining domain knowledge and meta-classification. Journal of Medical and Biological Engineering, vol. 33, no. 4. p.p. 406-414, (2013).

[10] Gjoreski, H., Gams, M. and Luštrek, M. Context-based fall detection and activity recognition using inertial and location sensors. Journal of Ambient Intelligence and Smart Environments, vol. 6, no. 4, p.p. 419-433 (2014).

[11] Gjoreski, H., Kozina, S., Luštrek, M and M. Gams. Using multiple contexts to distinguish standing from sitting with a single accelerometer. European Conference on Artificial Intelligence (ECAI), 2014.

[12] Trost, S.G., Zheng, Y. and Weng-Keen Wong. Machine learning for activity recognition: hip versus wrist data. Physiol Meas. 2014 Nov; 35(11):2183-9. doi: 10.1088/0967-3334/35/11/2183.

[13] Rosenberger, M. E., et al. Estimating Activity and Sedentary Behavior From an Accelerometer on the Hip or Wrist. Med Sci Sports Exerc. 2013 May; 45(5): 964–975. doi:10.1249/MSS.0b013e31827f0d9c.

[14] Mannini, A. et al. Activity recognition using a single accelerometer placed at the wrist or ankle. Med Sci Sports Exerc. 2013 November; 45(11): 2193–2203. doi:10.1249/MSS.0b013e31829736d6

[15] Ellis K. et al. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol. Meas. 35 (2014) 2191–2203. doi:10.1088/0967-3334/35/11/2191

[16] Wu H, Lemaire ED and Baddour, N. Activity Change-of-state Identification Using a Blackberry Smartphone. Journal of Medical and Biological Engineering, 2012; 32: 265–272.

[17] Lai C, Huang YM, Chao HC, Park JH. Adaptive Body Posture Analysis Using Collaborative Multi-Sensors for Elderly Falling Detection. IEEE Intelligent Systems. 2010; 2–11.

[18] Kwapisz JR, Weiss GM, Moore SA. Activity Recognition using Cell Phone Accelerometers. Human Factors. 2010; 12:74–82.

[19] Ravi N, Dandekar N, Mysore P, Littman ML. Activity Recognition from Accelerometer Data. In Proceedings of the 17th conference on Innovative applications of artificial intelligence. 2005; 1541–1546.

[20] Garcia-Ceja, E., Brena R. F., Carrasco-Jimenez J. C. and Garrido, L. Long-Term Activity Recognition from Wristwatch Accelerometer Data. Sensors 2014, 14, 22500-22524;

[21] Tapia, E. M. Using Machine Learning for Real-time Activity Recognition and Estimation of Energy Expenditure. Ph.D. Thesis, Massachusetts Institute of Technology, 2008.

How to recognize animal species based on sound – a case study on bumblebees, birds, and frogs

Anton Gradišek, Gašper Slapničar, Jure Šorn, Boštjan Kaluža, Mitja Luštrek, Matjaž Gams
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
anton.gradisek@ijs.si

He Hui
Shanghai University
Shangda Road 98, Baoshan District
200444 Shanghai
China

Tomi Trilar
Prirodoslovni muzej Slovenije
Prešernova 20, SI-1001 Ljubljana, Slovenia
ttrilar@pms-lj.si

Janez Grad
janez.grad@siol.com

ABSTRACT
We present a machine learning-based approach to recognize different types of animal species based on the sound they produce. We focus on bumblebee classification - the algorithm was first developed to recognize bumblebees (roughly 15 most common species found in Slovenia) according to their species and type (queen or worker). Later, it was tested on a set of birds (different species of cuckoos) and frogs of Slovenia. We discuss the sound sample preprocessing, machine learning algorithm, results of algorithm testing, and possible further improvements. A web-based service was developed where users can upload their recordings and further contribute to the learning dataset.

General Terms
Algorithms

Keywords
Sound recognition, machine learning, animal sounds, MFCC

1. INTRODUCTION
Bumblebees (genus *Bombus* from the bee family Apidae) play a key role in the ecosystem as important pollinators. Their different body structure gives them certain advantages over other bees. For example, they can be active in a wider range of weather (bees won't leave the hive when the outside temperature is below 10 °C while a bumblebee is active even below 5 °C). Certain plant species rely on bumblebees as pollinators exclusively, including some cultural plants. For example, bees won't pollinate tomatoes but bumblebees will, which makes them important in the economic sense as well. Selling bumblebee colonies to greenhouses has become a lucrative business in the last decade [1].

There are over 250 species of bumblebee species known worldwide. The biggest diversity is found in Asia while bumblebees are also distributed in Europe [2], North Africa and in the Americas. The highest diversity is found in mountain ranges in temperate climate zones, so perhaps it is not surprising that there have been 35 bumblebee species recorded in Slovenia. Some of these species are either rare or were recorded several decades ago, therefore it is more realistic to say that one can encounter around 20 different species. Bumblebees are social insects; their colonies consist of queens, workers, and males. These types are called castes.

Experts can identify species and caste based on body features, such as the hair colour pattern and body size. For non-experts, some applications have been developed to help with classification, such as *Bumblebees of Britain & Ireland* [3], which provides photos and descriptions of the common species of the British Isles, and *Ključ za določanje pogostih vrst čmrljev* (Key for determination of common Bumblebees), which also provides drawings, photos and descriptions of the common species of Slovenia [4]. Here, we attempt to classify the species and castes automatically, using a computer algorithm. Image recognition is perhaps not the most practical approach due to complications arising from photo quality, light condition, bumblebee orientation, background, etc. Recognition based on the buzzing sound is more promising. In past, there have been attempts to use machine learning-based algorithms to classify different types of insects [5] and also different bird species [6],[7].

In our approach we used Mel-Frequency Cepstrum Coefficients (MFCC) as a feature vector alongside hundreds of others audio features, similar to what was done in the studies mentioned above. Data was preprocessed using Adobe Audition software. Features were extracted using openAUDIO feature extraction tool [8]. Classification algorithms were created using WEKA open source machine learning software. The approach was tested on three groups of animals: bumblebees, with the largest number of samples (11 species, with queens and workers both represented in most cases, 20 classes in total), Slovenian frogs (13 species), and different species of cuckoos (7 species). The recordings of bumblebee were obtained in the field, frog sounds were obtained from the CD Frogs and toads of Slovenia [9] produced by Slovenian Wildlife Sound Archive [10], and the sounds of the cuckoos were obtained from the Chinese database 鸟类网.

In order to make the sound recognition application available to broader audiences, we have developed a web-based service where users can, apart from using only the species classification feature, upload their recordings to be later used in the learning set for further improvement of the classification. The application is now available at animal-sounds.ijs.si It runs in Slovenian, English, and Chinese.

2. PREPROCESSING
First, original sound recordings were manually cut to fragments a couple of seconds long and the sections with no bumblebee sound were excluded. Figure 1 shows a typical (unprocessed) sound file in time domain (*B. hypnorum*, worker) while Figure 2 shows the Fourier transform (absolute value) of data in Figure 1. As seen from the Fourier transform, the relevant frequency window for bumblebee sound is roughly between 100 and 1500 Hz, what is out of this window, can be considered noise. We can clearly see the main frequency at around 200 Hz and the higher harmonics at

multiples of this value. The recordings of bumblebees were typically of good quality and there was no need to additionally filter out background noises since the buzzing sound was by far the most prominent part of the recording.

For frog sounds, the situation was somewhat different. The recordings often contained other sounds, such as other animal sounds (other birds, frogs, insects, etc.) or sounds from sources such as running water etc. Here, background noise was removed by selecting a part of the recording that contains only noise and using standard noise cancellation software tools.

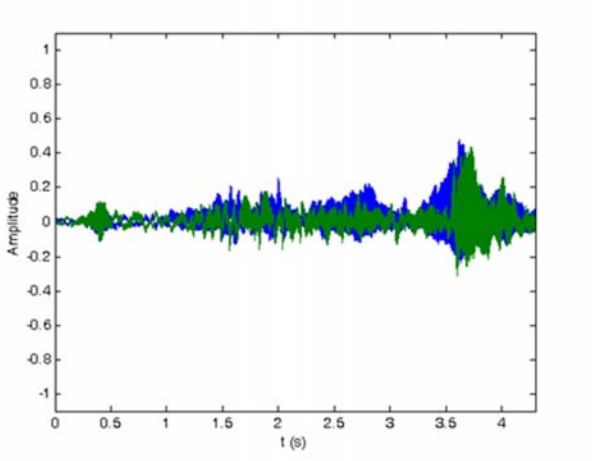


Figure 1. Time-domain representation of a typical sound recording, *B. hypnorum*, worker. Blue and green lines represent the two components of the signal that was recorded in stereo technique.

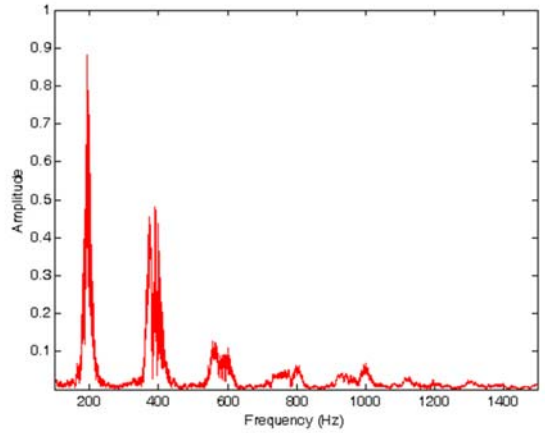


Figure 2. Fourier transform of time-domain data from Figure 1.

3. MACHINE LEARNING AS A SERVICE

Machine learning application was designed following the Machine Learning as a Service (MLaaS) paradigm. This ensures that the data processing, classification model creation, and interaction with client are available within a single cloud service. This animal classification service comprises of three main parts, as shown in Figure 3:

1. audio feature extraction,
2. creation of classification models,
3. user recording processing and serving of results.

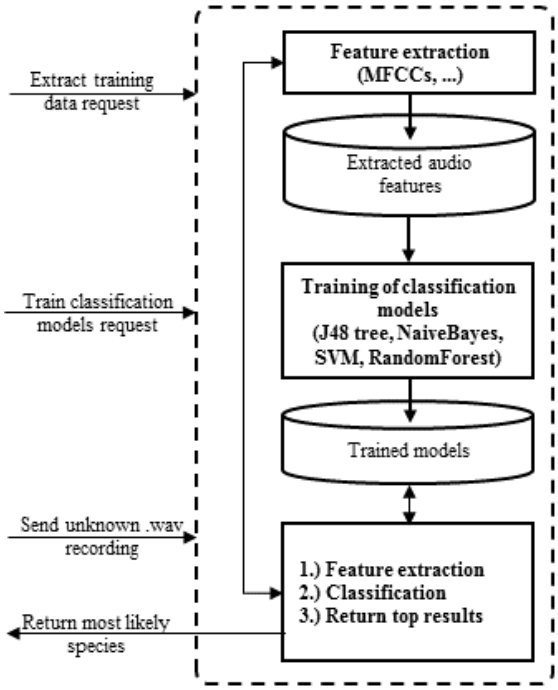


Figure 3. Architecture of animal classification machine learning service.

Audio feature extraction part is responsible for obtaining relevant data, which is then used to create classification models. As input it takes audio files in .wav format. It then computes numerical values representing a large number of different properties of audio signal. Most important among these are Mel-Frequency Cepstrum Coefficients. Following the extraction, the system chooses the best 100 among all extracted features using information gain as a feature quality measurement. These best extracted features are then saved into a database.

The training part takes the extracted features data and uses it to build classification models. It builds the models based on the following algorithms: decision tree, naïve Bayes, Support Vector Machine and Random Forest. All four models are always built, for each animal group. This allows for comparison of classification accuracy based on which we can choose the best performing algorithm. This is described further in section 4.

Third part of the system allows for client interaction. A query with a .wav recording is taken as input and then the system proceeds to extract the same features as were extracted for the training data. It then forwards these features as input into the chosen best classification model, which returns the most likely species.

4. EVALUATION

First we evaluated the results using WEKA built-in kFold cross-validation on all recordings. In this case the results are over optimistic, since parts of the same long recording can appear in both the learning and testing set. This issue was resolved by using evaluation with separate testing (80% of the data) and learning set (20% of the data) where recording slices from each set never belong to the same mutual recording.

In three cases Random Forest algorithm has shown the best classification accuracy while in the case of frogs, Support Vector Machine was slightly superior.

Test results are best presented by means of confusion matrices, which are, in case of bumblebees, too large to be presented in the paper. Evaluation of bumblebee classification shows that the quality of recognition of particular species depends on several factors. Recognition is best in cases where there were several recordings available whereas a small number of clips can result in overfitting the data and the results should therefore be treated with caution. For the classes with at least 20 instances in the classification works best for *B. pascorum*, workers (85%), *B. hypnorum*, worker (100%), *B. sylvarum*, worker (96%), while the classification of *B. humilis*, worker is only 18% accurate. In the attempt to improve the recognition accuracy, we have then decided that the output of the program are three most probable results (as opposed to only the most probable one), together with the pictures of the corresponding species. This additionally helps the user to decide which species was observed.

The results on a set of cuckoos were, on the other hand, surprisingly accurate, with the algorithm correctly classifying 73 out of 74 test instances. The confusion matrix is presented in Table 1:

a	b	c	d	e	f	g	h	<-	-	classified as
15	0	0	0	0	0	0	0	a	=	black-cuckoo
0	6	0	0	0	0	0	0	b	=	himalayan-cuckoo
0	0	12	0	0	0	0	0	c	=	indian-cuckoo
0	0	0	11	0	0	0	0	d	=	lesser-cuckoo
0	0	0	0	12	0	0	0	e	=	madagascan-cuckoo
0	0	0	1	0	10	0	0	f	=	red-chested-cuckoo
0	0	0	0	0	0	7	0	g	=	sunda-cuckoo
0	0	0	0	0	0	0	0	h	=	Unknown

Table 1. Confusion matrix for recognition of seven different species of cuckoos, using the SMO classifier.

The reason for this very high accuracy could be the fact that all the recordings were of extremely high quality (meaning that there was no background noise and the voice was clear) and the songs of different species also differ to the level where an amateur can recognize them only by listening (which is certainly not the case with the bumblebees). The question what would happen if recordings of worse quality were introduced remains open.

In the case of frogs, the recordings were first manually preprocessed with the noise removing software. Original recordings included other animal sounds and sounds of non-animal origin. Furthermore, several species of frogs have more than one type of call and all different calls for each species were grouped into a single class. Nevertheless, the overall classification accuracy was still reasonably high, with 148 out of 179 instances correctly classified (83%).

5. IMPLEMENTATION

Play Framework (Java) was chosen to develop a cloud-based REST service, which offers three endpoints, one for each animal group. WEKA open source machine learning library was used alongside Play Framework to implement the mentioned classification algorithms.

We wanted to offer a unified web application, which would allow users to upload their audio recordings and get the names and images of the most likely species for this recording. Extra functionality is a database in which registered users can save their recordings. Since only good quality recordings are desired in the database we added the feature that only an administrator or a bumblebee/frog/bird expert can confirm these user recordings as suitable, to be permanently added to the database and the learning set.

To do this we developed a Ruby on Rails web application. Web application is easy to use, common to all devices using libraries as Bootstrap and jQuery. The application separates users to ordinary users and administrators, which have different rights to different actions. For authentication of users we take classic session system. The goal of our application was to implement some kind of web portal with audio recordings. Any registered user can add audio recordings of specific animal, which are saved on our server. These audio recordings can be edited by animal experts and be saved to confirmed recordings database. For database we use well known MySQL.

6. DISCUSSION

We have demonstrated that a machine-learning based approach to classify different species of animals by their sounds produces good results. Mel-Frequency Cepstrum Coefficients and other audio features were calculated for each recording and 100 features with the highest information gain were chosen to build classification models. The classification accuracy is excellent in the case of cuckoos, very good for frogs, and variable for bumblebees – some species are classified with high accuracy while some are not. To improve this, three most likely results, together with the corresponding photos, are presented as the output. It is expected that the performance of the classification application will improve when more recordings for each species are available, since some of the classes currently consist of only one or two recordings.

Currently, the preprocessing of the recordings is done manually, the plan is to make this feature automatic as well. In future, we aim to expand the application to include even more groups of animals.

7. REFERENCES

[1] Grad, J., Gogala, A., Kozmus, P., Jenič, A., Bevk, D. (2010): Pomembni in ogroženi opraševalci – Čmrlji v Sloveniji. Čebelarska zveza Slovenije, Lukovica. ISBN 978-961-6516-30-3

[2] P. Rasmont, S. Iserbyt. (2010): Atlas of the European Bees: genus Bombus, Project STEP Status and Trends of European Pollinators. http://www.zoologie.umh.ac.be/hymenoptera/page.asp?ID=169

[3] https://itunes.apple.com/nz/app/bumblebees-britain-ireland/id657076684?mt=8&ign-mpt=uo%3D8

[4] http://www2.pms-lj.si/kljuci/cmrlji/

[5] Diego F. Silva, Vinícius M. A. De Souza, Gustavo E. A. P. A. Batista, Eamonn Keogh, Daniel P. W. Ellis, Applying Machine Learning and Audio Analysis Techniques to Insect Recognition in Intelligent Traps, ICMLA '13 Proceedings of the 2013 12th International Conference on Machine Learning and Applications

[6] Chang-Hsing Lee, Chin-Chuan Han, Ching-Chien Chuang, Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients, IEEE transactions on audio, speech, and language processing, vol. 16, no. 8, November 2008

[7] Marcelo T. Lopes, Lucas L. Gioppo, Thiago T. Higushi, Celso A. A. Kaestner, Carlos N. Silla Jr., Alessandro L. Koerich, Automatic Bird Species Identification for Large Number of Species, 2011 IEEE International Symposium on Multimedia

[8] Björn Schuller, Florian Eyben, Felix Weninger, Technische Universitaet Muenchen, Germany, openAudio.eu

[9] Tomi Trilar. Slovenske žabe = Frogs and toads of Slovenia. Ljubljana: Prirodoslovni muzej Slovenije: = Slovenian Museum of Natural History, 2003. 1 CD, stereo. ISBN 961-6367-07-2.

[10] http://www.pms-lj.si/si/o-naravi/zivali/oglasanje-zivali/arhiv-zivalskih-zvokov

Data Preparation for Municipal Virtual Assistant

Leon Noe Jovan
Department of Intelligent Systems,
Jožef Stefan Institute,
Jamova cesta 39
1000 Ljubljana, Slovenia
leon.jovan@gmail.com

Damjan Kužnar
Department of Intelligent Systems,
Jožef Stefan Institute,
Jamova cesta 39
1000 Ljubljana, Slovenia
damjan.kuznar@ijs.si

Matjaž Kukar
Laboratory for Cognitive Modelling,
Faculty of computer and information science,
Večna pot 113
1000 Ljubljana, Slovenia
matjaz.kukar@fri.uni-lj.si

Matjaž Gams
Department of Intelligent Systems,
Jožef Stefan Institute,
Jamova cesta 39
1000 Ljubljana, Slovenia
matjaz.gams@ijs.si

ABSTRACT

This document describes an automated knowledge base construction process for the question-answering virtual assistant, which answers on questions about municipalities in Slovenia. This process needs to be automated, because manual acquisition is too time consuming. We solve this problem with the proposed system that extracts assistant's knowledge base from web pages that contain required information. Two approaches were implemented and evaluated, 1) using machine learning approach and 2) parsing results from search engine results. Evaluation shows that more than 50% of answers are appropriate, which is considered as acceptable result for our specific problem.

Keywords

virtual assistant, knowledge extraction, machine learning, multi-label classification

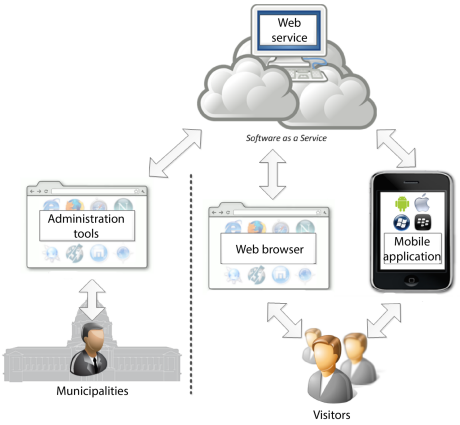


Figure 1: Scheme of project "Asistent" web service

1. INTRODUCTION

Project "Asistent" [2] is an intelligent virtual assistant which understands questions asked in natural language and is capable of providing relevant answers about Slovenian municipalities. It helps the visitors of municipality's web page to find information and services that web page offers. "Asistent" is a cloud-based service which can be accessed through application installed on municipality's website or through Android, iPhone, BlackBerry or Windows Phone mobile application. Scheme of the "Asistent" service is presented in figure 1.

Basic functionality of each Assistant is the ability to answer 500 pre-defined questions about specific municipality. Questions are same for all municipalities, but answers are not. First, we expected that municipalities will modify the assistants' knowledge bases with answers specific to their needs but it turned out that manual knowledge base construction is too time demanding. So there is a need for the automation of this task.

The goal of this project is to implement and evaluate a system that can automatically construct the knowledge base with answers based on 500 basic questions. We designed a system that works in two steps. First, the system finds the web page which contains important information that answers the question. This is achieved by using machine learning approach or parsing search engine results. Second step is extracting the information from the web page and construction of a short answer. This step is already implemented and it is described in our previous work [4] where we described the approaches and evaluated the system. Results showed that system generates 82% successfully retrieved answers.

In this document we focus only on the first step which means finding the web pages which contains the information that answers the questions. We implemented and evaluated two approaches 1) machine learning, more specifically multi-label classification and 2) parsing results from search engine.

2. DATA

Virtual assistant for municipalities can provide answers for about 500 different questions about municipality. After manual review of the questions we discovered that answers for only 237 different questions for each municipality need to be answered using multi-label classification and parsing search engine results.

Answers for other questions are generated using more simple approaches, since answers are either same for all municipalities or the answers can be acquired from dedicated web services.

In purpose for multi-label classification, we need labeled data that the algorithms can learn on. Learning data set was obtained from 17 manually constructed assistant's knowledge bases. Each knowledge base was constructed manually by municipality and contains a list questions and web pages which represent the answers. Because municipalities did not answered on all questions, it consists only of 1589 answers in form of web pages which means 93 web pages per municipality. Total number of unique web pages is 946. On average a single web page contains information for 1.67 questions. For each of the 237 questions we have 6.7 different web pages on average. That means that we have only few web pages from which the multi-label methods can learn.

3. METHODS

3.1 Multi-label classification

Multi-label learning is a form of supervised learning where the classification algorithm is required to learn from a set of instances, each instance can belong to multiple classes and therefore the classifier is be able to predict a set of class labels for a new instance [6].

Existing methods for multi-label classification can be divided into two main categories: problem transformation methods, and algorithm adaptation methods. In this work we use the methods from first group which transform the multi-label problem into a set of binary classification problems which can then be handled using multiple single-class classifiers.

3.1.1 Binary relevance

Binary relevance is a simple and popular problem transformation method, which builds a single-label classifiers, one for each label which participates in the multi-label problem. The multi-label prediction for a new instance is determined by aggregating the classification results from all independent binary classifiers.

3.1.2 Ranking by Pairwise Comparison

Ranking by pairwise comparison (RPC) transforms the multi-label data set into $\frac{M(M-1)}{2}$ binary label data sets, one for each pair of labels. Each data set contains the examples that are annotated by at least one of the two corresponding labels, but not both. A binary classifier that learns to discriminate between the two labels is trained from each of these data sets. Given a new instance, all binary classifiers are invoked, and a ranking is obtained by counting the votes received by each label [7].

3.1.3 RAkEL - Random k-Labelsets

RAkEL (Random k-Labelsets) is an ensemble method for multi-label classification that constructs an ensemble of multi-label classifiers where each classifier is trained using a different small random subset of the set of labels. In order to get near-optimal performance appropriate parameters as subset size and number of models must be optimised. Prediction of new instances using this ensemble method proceeds by a voting scheme [7].

3.2 Evaluation metrics

3.2.1 First Hit Sucess

First Hit Success (FHS) is a simple measure for evaluating Q&A systems. If the first answer returned by the system answers the question correctly, the FHS is 1. Otherwise the FHS is 0. If we only consider the first answer to each question on a set of questions and assume the Web contains answers to all the questions, then the average of FHS represents the recall ratio of a Q&A system [5].

3.2.2 Mean Reciprocal Rank

Mean reciprocal rank (MRR) is a statistic measure for evaluating Q&A systems that produces a list of possible responses to a sample of queries, ordered by probability of correctness. It captures how early get relevant result in ranking. The mean reciprocal rank for a sample of queries Q is defined by:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

4. MACHINE LEARNING APPROACH

Idea of this approach is to learn multi-label models and use them to find the best web pages which contains information on each question. Each question represents a single class label in multi-label classification.

Manually labeled data represents the training data and all web pages of a specific municipality represent the test data set. We want to construct assistant's knowledge base with some of these web pages. Test data was obtained using a simple implementation of a web crawler which obtain all web pages of certain municipality.

Before classification, text was obtained from the web pages, stopwords were removed from the text and lemmatisation was performed using LemmaGen [1] lemmatiser for Slovene language. Text was presented with Vector Space Model where terms correspond to a single word or a short phrase up to 3 words long. Only 200 most frequent terms were selected and weighted using *tf-idf* scheme.

We implement two methods. First is a method similar to binary relevance, which builds a single-label classifiers, one for each label which participates in the multi-label problem. The multi-label prediction for a new instance is determined by aggregating the classification results from all independent binary classifiers. In our case we need only one instance for each label per assistant. We labeled an instance only if the

probability score on the predicted label is the highest among all new instances.

Second method is an ensemble method for multi-label classification named RAkEL (Random k-Labelsets). Using this method we select candidate webpages. For underlying method for multi-label learning we use RPC method. We select the final prediction from the candidates using single-label classifier based on Logistic Regression.

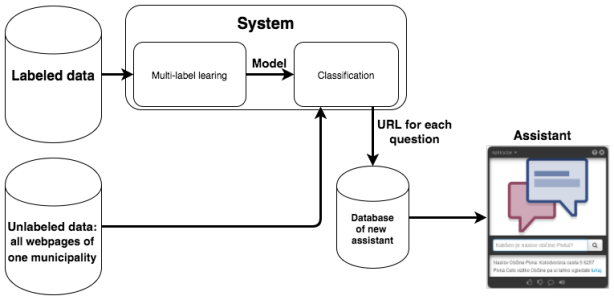


Figure 2: Scheme of implemented system with machine learning approach.

5. SEARCH ENGINE APPROACH

Idea of this approach is to use a search engine like Google, Bing or DuckDuckGo to find an answer to every question. That seems like a natural solution, because search engines has similar function as our virtual assistant and usualy give relevant results.

We define a list of keywords for every question. For every question a request to a search engine is made using these keywords. To specify the municipality, name of the municipality and the word "občina" was added to keywords. (e.g. `občina <name_of_municipality> <keywords>`) First result for this query is considered as the suitable answer.

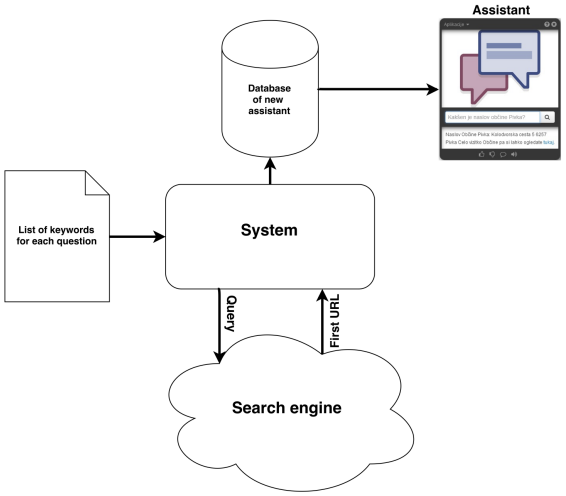


Figure 3: Scheme of implemented system using search engine results.

6. EVALUATION

Table 1: Results of automatic evaluation using leave-one-out

Method	FHS	MRR
Binary Relevance	0.364	0.522
RAkEL + Binary Relevance	0.291	0.408
Search Engine	0.314	0.386

Table 2: Manually reviewed answers in form of web pages ranked by the relevance.

Method	highly rel.	rel.	irrel.
Binary Relevance	15.5%	5.5%	79.0%
RAkEL + Binary Relevance	15.0%	10.0%	75.0%
Search Engine	37.5%	17.0%	45.5%

Our system was evaluated using a variant of leave-one-out method - leave-one-municipality-out. In each iteration data belonging to one of the municipalities was used for testing the classifier with FHS and MRR.

This evaluation method is used only for optimizing parameters and does not reflect the efficiency of the system. First of all, we have labeled only the most relevant web page for each question, sometimes the system recommends some other page, which is relevant too, but that do not affect the evaluating score. The second problem is, that differences between partially and highly relevant web pages, may not become apparent in evaluation. To bring such differences into daylight, we need graded relevance [3] judgments which indicates the relevance of a document to a query on a scale using numbers, letters, or descriptions (e.g. "not relevant", "somewhat relevant", "relevant", or "very relevant").

In order to estimate real efficiency of the system we manually evaluate the system on new municipality. For each question the system suggested one web page that should contain the answer on the question. These web pages were manually reviewed and estimated by relevance: highly relevant, relevant or irrelevant.

Table 1 presents FHS and MRR for all three methods we evaluate using leave-one-out method.

Table 2 presents manually estimated relevance of answers using scores "highly relevant", "relevant" and "irrelevant".

7. RESULTS AND DISCUSSION

From the tests, it can be seen, that both approaches give similar results when testing with leave-one-out, but when testing the on new municipality, search engine approach gave much better results. From varying results of the tests we can see, that automatic evaluation of the system this type is very difficult, because there are many correct answers for each question, but we have labeled only one per each question and municipality.

Data mining approach gives decent results, but it produces many irrelevant answers. The second problem of this approach is, that it is highly dependant on crawled data. Some answers are usually not located on web pages of municipality, but it is hard to know which domains are important for clas-

sification. Positive side of this approach is that the system is independent of questions. We can quickly adapt system for other questions as long as we have labeled training data. The problem in our case is that we have many labels with only few labeled instances. We noticed, that this approach work better for labels, which we have above average number of labeled data instances.

Approach using search engine gives much better results than other approach when testing on new municipality. More than 50% of answers were at least relevant to the question, 37.5% were very good answers. We assess this result as sufficient, because majority of important questions are answered.

8. FURTHER WORK

The approach using search engine gives good results, but we assume that we can achieve much better results. When parsing results from search engine we always take only first result, but many times this is not the best answer. We propose an approach that combine both approaches we implemented and evaluated. The idea is to use different search engines to enrich the training labeled data and to get the candidates for classification. In that way, we solve a problem of our small labeled data set, but this new data must be filtered before adding into training set. The second advantage is that we already get potential websites, so we do not need any specialised web spider, because the search engine already do all that work.

Using similar machine learning methods like we implemented, we will predict, which of the results ranked by the search engine is actually the best answer. First result from search engine is not always the best. Using supervised learning we can choose better web page that suits our problem which means better final results.

Second important thing to reconsider is usage of FHS and MRR metrics for evaluating our system. From the results we can see, that the results can extremely vary between automatic evaluation and manual review of the answers. In the future we will try different approaches like automatic estimating of relevance. With this approach we can replace human relevance judgments by an automatic method. That methods allow fast and more accurate automated evaluations.

When the results will be good enough, we need to connect implemented system with the second system we implemented which generates a short answer from the text collected from the web page. The system is already implemented and it is described in our previous work [4].

9. CONCLUSIONS

This document describes constructing knowledge base for virtual assistant. We developed a system which find a web page on which the answer on a question is located. To solve this problem, we implemented and evaluated two approaches, one using machine learning approach and second using search engine crawling approach. After testing the implemented system on new municipality data, the system using search engine results achieves more than 50% relevant answers which we consider as a good result.

10. ACKNOWLEDGMENTS

We wish to thank Tanja Breznik who manually reviewed the results and professionally estimate relevance of all answers and Svetlana Nikić who helped implementing our system. Without their continued efforts, we would have not been able to bring our work to a successful completion.

11. REFERENCES

[1] Lemmagen - multilingual open source lemmatisation. `http://lemmatise.ijs.si/`. (Visited on 09/10/2015).

[2] Projekt asistent - virtualni asistent za občine in društva. `http://www.projekt-asistent.si/wp/`. (Visited on 09/10/2015).

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[4] L. N. Jovan, S. Nikić, D. Kužnar, and M. Gams. Avtomatizacija izgradnje baze odgovorov virtualnega asistenta za občine. *Proceedings of the 17th International Multiconference Information Society – IS 2014*, A:46–50, 2014.

[5] D. R. Radev, H. Qi, H. Wu, and W. Fan. Evaluating web-based question answering systems. *Ann Arbor*, 1001:48109.

[6] M. S. Sorower. A literature survey on algorithms for multi-label learning. Technical report, 2010.

[7] G. Tsoumakas, I. Katakis, and L. Vlahavas. Random k-labelsets for multilabel classification. *Knowledge and Data Engineering, IEEE Transactions on*, 23(7):1079–1089, July 2011.

Determining Surface Roughness of Semifinished Products Using Computer Vision and Machine Learning

Valentin Koblar
Kolektor Group d.o.o.
Vojkova ulica 10
SI-5280 Idrija, Slovenia
valentin.koblar@kolektor.com

Martin Pečar
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
m.pecar@ijs.si

Klemen Gantar
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113
SI-1000 Ljubljana, Slovenia
kg6983@student.uni-lj.si

Tea Tušar
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
tea.tusar@ijs.si

Bogdan Filipič
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
bogdan.filipic@ijs.si

ABSTRACT

In the production of components for various industries, including automotive, monitoring of surface roughness is one of the key quality control procedures since achieving appropriate surface quality is necessary for reliable functioning of the manufactured components. This study deals with the task of determining the surface roughness of semifinished products and proposes a computer-vision-based method for this purpose. To automate the design of the method, machine learning is used to induce suitable predictive models from the captured product images, and evolutionary computation to tune the computer vision algorithm parameters. The resulting method allows for accurate online determination of roughness quality classes and shows a potential for online prediction of roughness values.

1. INTRODUCTION

Quality control procedures in advanced manufacturing involve sophisticated techniques to meet the continuously increasing demands for product quality and reliability. Since humans are prone to failures in quality assessment, they are being replaced by autonomous quality control procedures. The automotive industry is one of the most advanced in this respect. Among the quality control measures, monitoring of surface roughness in the production of automotive components is crucial for achieving reliable functioning of the components throughout the product lifetime. There are several challenges associated with this task in production environments where the roughness measurement methods are required to be non-contact, autonomous and performing online.

This paper presents the development and evaluation of a computer-vision-based method for online surface roughness measurement in the production of commutators that are components of electric motors widely used in automotive industry. The method predicts the roughness based on the attributes of the captured commutator images. The design of the method was automated using machine learning and evolutionary computation.

The paper is organized as follows. Section 2 presents the problem of determining the roughness of a particular commutator surface. Section 3 describes the data preparation for offline development and evaluation of the proposed method. The methodology of machine learning and optimization used to design the method is explained in Section 4. Section 5 reports on the conducted experiments and obtained results. Finally, Section 6 concludes the paper with a summary of findings and a plan for further work.

2. PROBLEM DESCRIPTION

A commutator is a device mounted on the shaft of the rotor in a commutator electric motor. During the rotor rotation, the commutator sequentially reverses the current direction through the rotor winding and hence enables continuous commutation of the motor. The joint between the commutator and the rotor shaft is made by pushing the shaft into the commutator mounting hole. Usually, the manufacturer of electric motors specifies the maximum and minimum force to be applied during the mounting operation. If an excessive force is employed, there is a risk of damaging the commutator. On the contrary, if the applied force is too small, the joint cannot withstand the mechanical stress during the motor operation.

The force required to push the shaft into the commutator mounting hole depends on two dimensional characteristics of the hole – its diameter and roughness. Both characteristics result from the final treatment of the commutator mounting hole carried out in the turning process. Several methods are available for measuring the inner diameter of the holes. However, online measurement of surface roughness represents a major challenge.

The most frequently used method of surface roughness measuring is contact profilometry. It uses a specially designed measuring tip that slides over the surface and measures the displacement in the range of micrometers. However, this method has several drawbacks: it is very sensitive to vibrations, it is slow, and the measuring tip can cause additional scratches on the measured surface. Consequently, contact profilometry is not suitable for online surface

roughness measurements. As an alternative, optical non-contact methods were developed, e.g., optical profilometry, scanning electron microscopy, atomic force microscopy etc. Unfortunately, none of them is suitable for online roughness measurement, since they are all sensitive to vibrations and, in addition, special samples have to be prepared for these methods to be applied.

Computer vision offers new possibilities in non-contact roughness measurement. For example, based on surface images captured by a CCD camera, calculation of a feature called the optical surface roughness parameter, G_a, was proposed and shown that it compares well wih the traditionally used average surface roughness, R_a, [2]. A machine vision system for online measurement of surface roughness of machined components was developed that relies on an artificial neural network model for predicting optical roughness values from image features [3]. It was also demonstrated that it is possible to measure surface roughness in three dimensions by combining a light sectioning microscope and a computer vision system [1].

In the literature, various parameters are considered in defining the measure of surface roughness [4]. They can be categorized into amplitude parameters, spacing parameters, and hybrid parameters. In our case, the commutator producer uses the parameter R_z which takes into account the difference between the maximum peak height and the maximum valley depth from a profile in the sampling length.

This research deals with the problem of determining the R_z parameter value and proposes an autonomous computer-vision-based method for this purpose. In addition, to automate the design of this method, machine learning and evolutionary computation are used. The resulting methodology is aimed at accurate online determination of the commutator mounting hole roughness on a commutator production line and is also applicable to other use cases requiring online surface roughness measurements.

3. DATA PREPARATION

The initial phase in designing a method for online surface roughness measurement was data preparation. It consisted of capturing the images of the commutator mounting hole surfaces, image preprocessing, and extracting attributes from the preprocessed images.

For the purpose of this study, 300 comutators were made available and the images of their mounting hole surfaces were taken in 8-bit grayscale with the resolution of 2592×1944 pixels. In addition, to obtain the reference values of their roughness for the purpose of machine learning, the samples were measured by a contact profilometer. To reduce noise in the obtained measurement data, each sample was measured three times and the average was then taken as a reference roughness value. Note that commutators with the roughness parameter value $R_z \leq 16$ μm are considered acceptable, while the ones with $R_z > 16$ μm unacceptable. The distribution of test images among the quality classes is shown in Table 1.

Table 1: Distribution of test images

Class	Number of images
Acceptable	159
Unacceptable	141

To preprocess the images and extract the attribute values from the preprocessed images, a dedicated computer vision algorithm involving a sequence of operators was implemented in the LabView programing environment [6]. Both the operators and the sequence of their deployment were determined manually, based on the experience from developing similar computer vision applications.

Image preprocessing is aimed at extracting the regions where surface roughness is to be measured. It comprises several steps as shown in Figure 1. It starts with an original grayscale image and applies a manually set threshold to it. The resulting binary image is then used to calculate the image centroid based on the pixel intensity. Since the mounting hole area always contains the highest proportion of high-intensity pixels, the calculated centroid is always positioned at about the same location of the mounting hole, regardless of its position in the image. In the next step, the coordinate system is assigned to the location of the calculated centroid. The coordinate system is used to precisely position the image mask for extracting an adequate region of interest (ROI) from the image. Finally, the image extraction mask is applied and part of the image inside the ROI is extracted. The result of this preprocessing is a cropped greyscale image of 700×300 pixels.

The attribute extraction procedure returns the values of 28 attributes describing the properties of the captured surface image, such as the grayscale value of pixels along the line profile in the image, the highest grayscale pixel value in the image, the lowest grayscale pixel value in the image, the distance in pixels between the stripes in the image, fast Fourier transform (FFT) values, etc. All extracted attribute values are numerical. The sequence of steps in the attribute extraction procedure is shown in Figure 2.

First, the fast Fourier transform (FFT) is applied to the image to eliminate the noise that results from inhomogeneities in the material. Based on the FFT results, certain proportion of high frequencies is rejected, thus removing noise from the image. However, despite the FFT frequency truncation, some noise may still be present. To further eliminate it, the median filter is activated next. This filter makes it possible to apply various structure elements in the image. As a result, in the longitudinal direction, where roughness is to be measured, details are preserved, while transversely they are filtered. The surface roughness in the image is thus emphasized. Afterwards, the surface line profile is measured and certain attributes are extracted from the grayscale image. To obtain additional attributes, the Niblack binarization algorithm [5] is applied. It returns a binary image with surface roughness represented by black and white stripes. Image processing with the Niblack algorithm may result in pixel clusters in the binary image. To eliminate these clusters, the particle filter operator is used. Next, morphology functions are applied to the binary image to equalize and straighten the edges of the stripes in the image. Finally, the attributes of the binary image are extracted and a file with all attribute values is generated.

4. MACHINE LEARNING AND OPTIMIZATION METHODOLOGY

The key phase in designing a method for surface roughness determination was machine learning of predictive models from the extracted attributes and the reference roughness values. For this purpose the open-source data mining environment Weka [9] was used. Two types of prediction were considered:

- classification, where the task is to label the surface roughness of a product as either *acceptable* or *unacceptable*,
- regression, where the task is to predict the value of the roughness parameter R_z for each product.

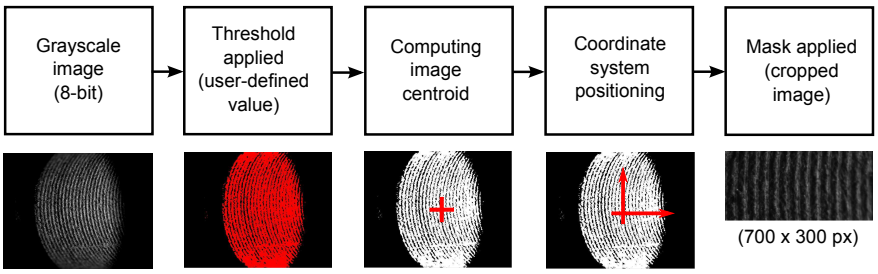


Figure 1: Preprocessing of the captured images

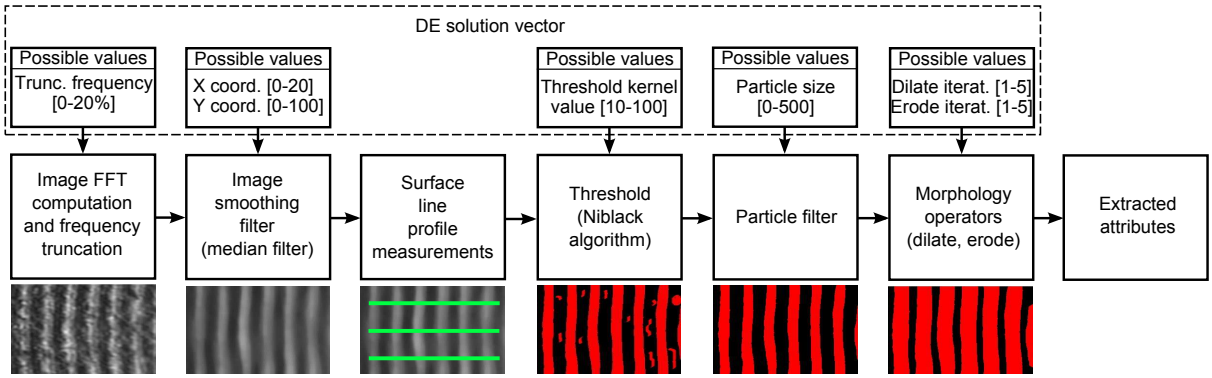


Figure 2: The attribute extraction procedure

For the classification task the Weka implementation of the C4.5 algorithm [7] for building decision trees from data, called J48, was used. This approach was selected since decision trees are simple to implement and easy to interpret. The J48 involves several algorithm parameters that influence the performance and complexity of the induced model. However, in this study we focused on tuning the computer vision algorithm parameters, while machine learning with J48 was carried out using its default parameters settings. The goal of tuning the computer vision algorithm parameters was to maximize the classification accuracy of the models induced from the image attributes. The classification accuracy was estimated through 10-fold cross-validation.

The regression task of predicting the value of the roughness parameter $R_z$was approached using the M5P algorithm for regression tree induction available in Weka. The leaves of regression trees produced by M5P contain linear models predicting the target variable value (R_z). Like J48, M5P includes several algorithm parameters too. For the same reasons as in the classification task, these parameters were set to their default values. Here, the objective of tuning the computer vision algorithm parameters was to minimize the Root Relative Squared Error (RRSE). The performance of the regression models was also validated with 10-fold cross-validation.

The methodology used for induction of predictive models and optimization of the computer vision algorithm parameters is shown in Figure 3. The optimization algorithm used in this methodology was Differential Evolution (DE) [8]. Solutions explored in the optimizaion process were vectors of parameters of the computer vision operators. The structure of a DE solution vector with the computer vision parameters subject to optimization is illustrated in Figure 2. After a solution is created, the computer vision algorithm, using

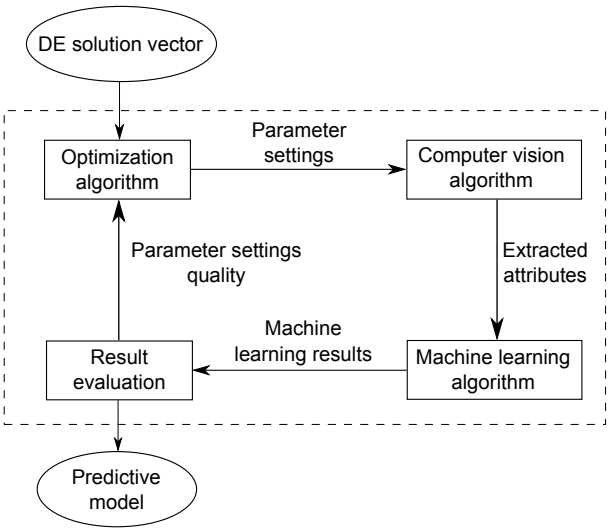


Figure 3: A schematic view of the methodology used for induction of predictive models and optimization of the computer vision algorithm parameters

the parameter values from the solution vector, preprocesses the input images and extracts the selected attributes. Next, the machine learning algorithm induces a predictive model from the attribute descriptions of the images and evaluates its performance. The evaluation result is passed back to the optimization algorithm, which generates a new solution. This iterative optimization procedure repeats until the stopping criterion specified in terms of the number of evaluated solutions is met.

5. EXPERIMENTS AND RESULTS

The presented methodology was evaluated on the acquired and preprocessed images in designing both the classification and the regression method of determining the surface roughness. We first focused on the classification task. The population size in the DE optimization algorithm was set to 10 and the algorithm was executed for 100 generations. Several runs were performed and the algorithm was consistently able to find decision trees with the classification accuracy of 100% in very few examined generations (no more than 10). This clearly indicates that the learning domain is not very complex. An additional analysis of the reference roughness values showed that they were dispersed very non-uniformly in that a roughness value was either well bellow or above the class discriminating value of R_z = 16 μm (see Figure 4). Moreover, a closer look at the induced decision trees revealed that in most cases a single attribute test was sufficient for accurate classification. Several attributes were identified as informative enough for this purpose: the FFT index value indicating the frequency of the stripes in the binary image, the average width of the stripes in the binary image, and the highest grayscale pixel value from the valley minima.

Next, the regression task was pursued. The population size and the number of generations in the DE algorithm were set to the same values as for the classification task, resulting in 1000 examined candidate solutions. In multiple runs of the optimization algorithm the RRSE value of the resulting regression trees in predicting the R_z roughness parameter value was found to be around 22% and its deviation between individual runs negligible. The calculated mean absolute error of these models was 0.75 μm, which is quite an encouraging result. The typical size of the regression trees was five nodes, i.e., two internal nodes with attribute tests and three leaves containing linear models for determining the surface roughness. The highest grayscale pixel value from valleys minima was found to be the most informative attribute in the trees. The surface roughness values predicted by a derived regression tree are compared to the measured reference values in Figure 4.

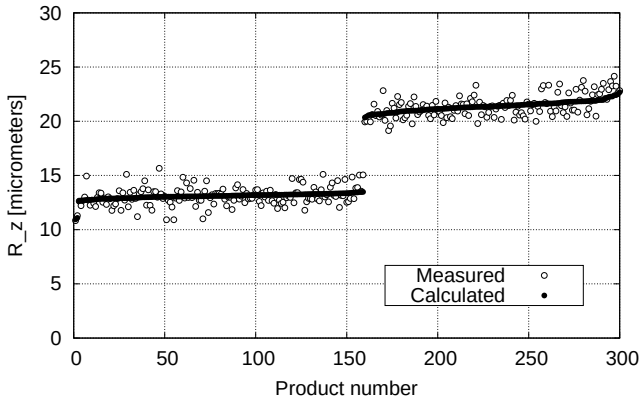


Figure 4: Comparison of measured and calculated roughness values

These results indicate that the proposed machine-learning creation of models for predicting surface roughness from the image attributes is a viable approach to the design of a computer-vision-based roughness measurement method. Under laboratory conditions, the study demonstrates that classification of surfaces into roughness quality classes can be performed accurately, while predicting the R_z value is not yet at this level, in our view mainly because of the very limited amount of product samples available for learning.

6. CONCLUSIONS

We have presented the development and experimental evaluation of a computer-vision-based method for determining the roughness of machined surfaces of semifinished products for automotive industry. Offline design of the method consists of building a predictive model from the attribute descriptions of the product images and optimization of the attribute extraction procedure. Once designed the model can be ported to an appropriate smart camera system to perform online roughness measurements within a quality control procedure on a production line.

The conducted laboratory evaluation confirms the suitability of the approach and, at the same time, indicates the need for refining the regression model for determining the roughness value. For this reason, further work will concentrate on deriving the model from a larger, systematically gathered and more representative set of samples. We expect the resulting model to be accurate and capable of detecting trends in the product roughness value that will be informative for taking appropriate process control measures. In addition, the optimization of the method will be extended to involve not only the parameters of the attribute extraction procedure but also the machine learning algorithm settings. Finally, online deployment and evaluation of the developed method in the production environment will be carried out.

7. ACKNOWLEDGMENTS

This work has been partially funded by the ARTEMIS Joint Undertaking and the Slovenian Ministry of Economic Development and Technology as part of the COPCAMS project (http://copcams.eu) under Grant Agreement number 332913, and by the Slovenian Research Agency under research program P2-0209.

8. REFERENCES

[1] O. B. Abouelatta. 3D surface roughness measurement using a light sectioning vision system. In *Proceedings of the World Congress on Engineering, WCE 2010, London, U.K.*, volume 1, pages 698–703, 2010.

[2] B. Dhanasekar and B. Ramamoorthy. Evaluation of surface roughness using a image processing and machine vision system. *MAPAN – Journal of Metrology Society of India*, 21(1):9–15, 2006.

[3] D. A. Fadare and A. O. Oni. Development and application of a machine vision system for measurement of surface roughness. *ARPN Journal of Engineering and Applied Sciences*, 4(5):30–37, 2009.

[4] E. Gadelmawla, M. Koura, T. Maksoud, I. Elewa, and H. Soliman. Roughness parameters. *Journal of Materials Processing Technology*, 123(1):133–145, 2002.

[5] J. He, Q. D. M. Do, A. C. Downton, and J. H. Kim. A comparison of binarization methods for historical archive documents. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition, ICDAR '05*, volume 1, pages 538–542, 2005.

[6] T. Klinger. *Image Processing with LabVIEW and IMAQ Vision*. Prentice Hall Professional, 2003.

[7] R. J. Quinlan. *C4. 5: Programming for Machine Learning*. Morgan Kauffmann, 1993.

[8] R. Storn and K. Price. Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

Nosljive naprave za izboljšanje kakovosti življenja

Tomaž Kompara
Elgoline d.o.o.
Podskrajnik 34,
Cerknica, Slovenija
tomaz.kompara@elgoline.si

Miomir Todorović
Elgoline d.o.o.
Podskrajnik 34,
Cerknica, Slovenija
miomir.todorovic@elgoline.si

POVZETEK

Sodobna tehnologija omogoča bistveno izboljšanje kakovosti življenja ne glede na starost, spol ali interesne dejavnosti. Na kakovost življenja vpliva več dejavnikov, pri čemer pa je zdravje eno izmed ključnih, saj omogoča produktivno in kakovostno življenje slehernega posameznika in skupnosti kot celote. Zato zdravje in skrb zanj nista le interes posameznika, temveč odgovornost celotne družbe. Najučinkovitejši način za izboljšanje zdravstvenega stanja posameznikov je zgodnja detekcija obolenj ter hiter in pravilen postopek odpravljanje le teh. Podjetje Elgoline sodeluje v več projektih, kjer se ukvarja predvsem z zaznavanjem in sporočanjem podatkov o uporabnikovem stanju. V tem prispevku so opisane rešitve, ki so primerne za različne tipe uporabnikov ter različne namene uporabe.

1. UVOD

Kvalitete življenja zaenkrat še ni mogoče neposredno kvantitativno meriti, zato se uporabljajo posredni kazalniki ter subjektivno doživljanje [1]. Podobno velja za ocenjevanje zdravstvenega stanja posameznika, pri čemer pa si lahko zelo pomagamo s sodobno tehnologijo [2]. Ta nam omogoča spremljanje življenskih znakov ter zaznavanje odstopanj izven mejnih vrednosti. Informacije o uporabniku je nato potrebno prenesti do osebe, ki je glede na stanje uporabnika zmožna pomagati.

Zaznavanje je tehnološko najzahtevnejši in hkrati najpomembnejši del celotnega sistema, saj mora biti izvedeno na uporabniku prijazen in nemoteč način. Na primer, le malo uporabnikov bi želelo uporabljati sistem za merjenje porabe energije, ki je sestavljen iz dihalne maske ter 5kg težke merilne opreme, ki ji je potrebno po nekaj urah uporabe zamenjati baterijo. Poleg same uporabnosti pa je pomembna tudi sama cena naprave, kar dodatno omeji nabor senzorjev, ki jih je mogoče uporabiti.

Zaradi teh omejitev je potrebno izbrati senzorje na podlagi katerih lahko posredno določimo merjene veličine. Na primer, porabo energije lahko ocenimo na podlagi pospeškov (mobilni telefon, zapestnica ipd.) [3, 4]. Tak način je mnogo cenejši in bolj prijazen uporabniku, vendar tudi manj zanesljiv in tehnološko težje izvedljiv.

Izmerjeni podatki se lahko prenašajo na drugo napravo ali pa se obdelajo lokalno na nošeni napravi sami. Vsak izmed pristopov ima svoje prednosti in slabosti. Najpomembnejša faktorja sta poraba električne energije in cena prenosa podatkov. V kolikor se izkaže, da je poraba energije in cena pošiljanja surovih podatkov manjša kot poraba potrebna za procesiranje in pošiljanje končnih rezultatov, je primernejši prvi pristop, v nasprotnem primeru pa drugi. Končna odločitev o načinu povezave naprave s preostalim svetom je največkrat odvisna od področja uporabe in namena naprave same.

V primeru ko se procesiranje vrši na nošeni napravi sami, je količina prometa mnogo manjša, zaradi česar vmesne naprave niso potrebne, ampak se obvestilo lahko prenese neposredno ciljni osebi (npr. nujno medicinsko pomoč - 112, sorodnikom, znancem itd.). Primer takšnega sistema je zaznavanje padcev [5], ki se posamezniku običajno pripetijo le nekajkrat v življenju. Slabost takšnega pristopa je posodabljanje programske opreme, saj je potrebno novo programsko opremo naložiti na vsako napravo posebej.

V primeru ko se podatki v obdelavo prenašajo v drugo napravo, je potrebno zagotoviti neprekinjeno delujoč sistem (npr. strežnik), ki podatke prejema ter uporabi za izračun želenih vrednosti. Prednost takšnega načina je možnost razširitve spominskega prostora ter procesorske moči glede na potrebe aplikacije ter števila uporabnikov. Poleg tega je mogoče algoritme za računanje želenih vrednosti enostavno posodobiti. Slabost takšnega načina zajema podatkov je drago vzdrževanje strojne opreme ter večja možnost posega v uporabnikovo zasebnost.

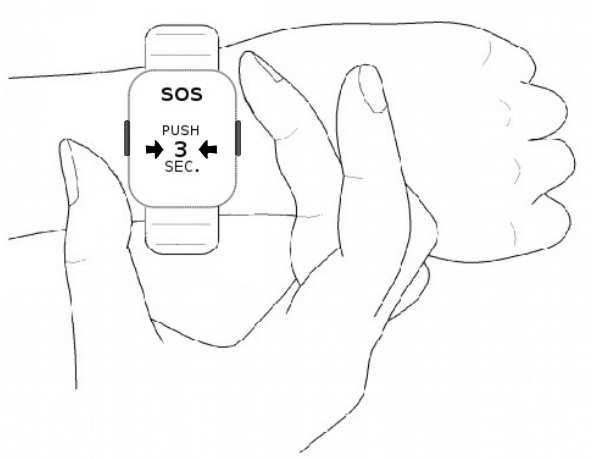
Oba načina pošiljanja je mogoče tudi poljubno združiti. Na primer, podatki se lahko delno obdelajo in shranijo na zapestnici sami ter se na zunanjo napravo prenesejo v poljubnem času. Tak primer so zapestnice za štetje korakov. Zaznavanje korakov in shranjevanje števila se obdelata na zapestnici sami, nato pa se ob povezavi s pametnim telefonom preneseta.

Na podlagi zbranih podatkov o uporabnikih je mogoče uporabiti napredne metode strojnega učenja ter na ta način pridobiti nova znanja o določenem pojavu. Tako lahko pojav bolje razumemo ter v prihodnje bolje detektiramo in hitreje ukrepamo. Prav to so smernice pri razvoju naprav v podjetju Elgoline. V nadaljevanju so predstavljene 4 nosljive naprave primerne za različne namene uporabe: SOS zapestnica, športni nadzornik, merilnik aktivnosti otrok in SOS zapestnica z detekcijo padcev.

2. SOS ZAPESTNICA

SOS zapestnica je naš prvi nosljiv produkt, ki je v osnovi zelo preprost. Namenjen je širšemu krogu ljudi, ki so izpostavljeni takšni ali drugačni nevarnosti in želijo imeti možnost hitrega klica do nujne medicinske pomoči ali do katere druge osebe. Primer takšnih uporabnikov so ekstremni športniki, otroci in starejši.

Uporaba je nadvse enostavna. Zapestnica se enostavno namesti na zapestje, kjer lahko ostane neprekinjeno, saj je vodoodporna in polnjenja ne potrebuje kar 1 leto. Ob kritični situaciji mora uporabnik hkrati pritisniti gumba na obeh straneh zapestnice (pritisk na oba gumba je potreben zato, da zmanjšamo možnost neželenega proženja alarma) 1. Ob proženju se zapestnica aktivira ter vzpostavi zvočno povezavo prek GSM omrežja z izbrano telefonsko številko. Medtem zapestnica poskuša samodejno locirati uporabnika s pomočjo GPS naprave. V kolikor podatek o lokaciji najde, ga sporoči preko SMS sporočila kontaktni osebi. V kolikor lokacije ni mogoče določiti preko GPS satelitov (npr. v notranjih prostorih) je mogoče uporabiti triangulacijo na podlagi moči signla do GSM oddajnikov.



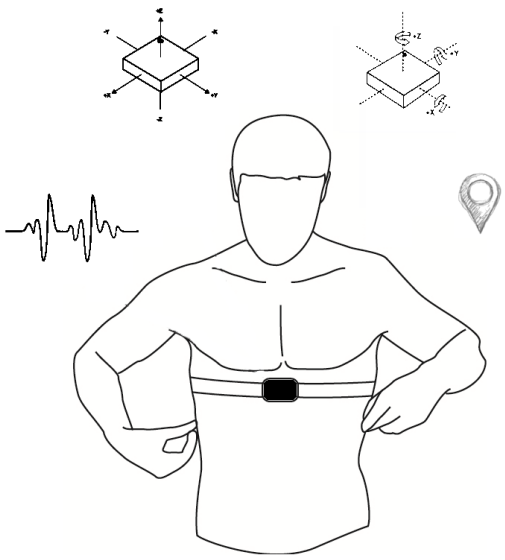
Slika 1: Proženje alarma na SOS zapestnici.

Kontaktno številko oz. številke je mogoče določiti ob inicializaciji naprave. Na zapestnico se podatki pošljejo preko SMS sporočila, zaradi česar zapestnica ne potrebuje nobenih dodatnih gumbov ali zaslona za nastavljanje. Tako je zapestnica primerna tudi za nevešče uporabnike sodobnih tehnologi saj ne potrebuje nikakršnjega upravljanja razen stiska gumbov ob klicu na pomoč.

Na ta način je hitrost sporočanja bistveno hitrejša in enostavnejša kot bi bila preko mobilnega telefona. To omogoča hitrejše posredovanje, s čimer ima uporabnik v kritičnih situacijah večjo možnost preživetja.

3. ŠPORTNI NADZORNIK

Zdravje je v profesionalnem športu še posebej pomembno, saj lahko resnejša poškodba pomeni tudi konec kariere. Zato je sistem za nadzor nad športnikovim stanjem izredno pomemben. V ta namen smo razvili napravo, ki zajema EKG signal, pospeške in rotacijo v vseh treh oseh ter zelo natančen GPS modul (z natančnostjo boljšo od enega metra). Za povezavo skrbi WiFi povezava.



Slika 2: Športni nadzornik meri EKG signal, pospeške, orientacijo in pozicijo.

Naprava je namenjena nadzoru med treningi, pri čemer je uporabnik skoraj ne opazi, trenerju pa nudi obilo podatkov o športnikovem stanju v realnem času. Na ta način lahko trener optimalnejše izbere vaje, njihovo trajanje in intenzivnost. Za športnika to pomeni zmanjšanje tveganje poškodb ter maksimalno napredovanje k zastavljenim ciljem.

Na podlagi vgrajenih senzorjev je mogoče določiti športnikovo hitrost, pospeške, razdaljo, število korakov, porabljeno energijo, moč, izčrpanost itd. Iz izračunanega stanja je mogoče oceniti tveganje poškodb oz. zaznati neobičajne gibe, ki so lahko znak, da je s športnikom nekaj narobe.

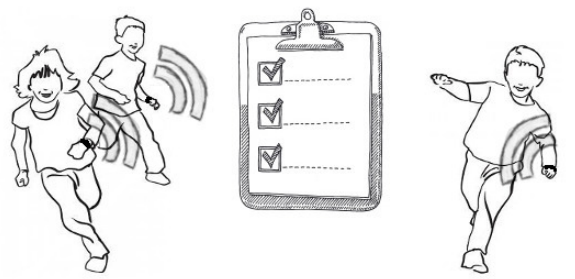
Ker je potrebno športnika med treningom spremljati v skoraj realnem času, naprava ob delovanju pošilja surove podatke s senzorjev na strežnik. Zaradi količine podatkov in zahteve po dometu se za prenos podatkov uporablja WiFi brezžična povezava. Poslani podatki so obdelani na strežniku ter prikazani trenerju na zaslonu. Tako lahko trener enostavno spremlja posameznega športnika ali več športnikov hkrati.

Podatki o posameznem športniku so lahko shranjeni na strežniku ter primerjani s podatki iz ostalih treningov. Na ta način je mogoče ugotoviti športnikovo pripravljenost skozi čas ter primerno sestaviti treninge v prihodnosti. S tem pospešimo hitrost napredovanja športnika ter zmanjšamo tveganje poškodb kar sta cilja večine vrhunskih športnikov.

4. MERILNIK AKTIVNOSTI OTROK

V zadnjih letih se je raven gibalnih sposobnosti otrok znižala, saj se ti veliko manj gibljejo kot poprej. Zaradi tega je potrebno najti nove načine, kako otroke spodbuditi k aktivnosti ter s tem dvigniti njihovo gibalno sposobnost.

V ta namen smo razvili zapestnico za merjenje aktivnosti,



Slika 3: Aktivnost otrok se beleži na strežniku.

ki na podlagi pospeškomera in žiroskopa ugotavlja otrokovo aktivnost ter šteje posamezne gibe (število skokov, počepov, korakov itd.). Zapestnica deluje samostojno ter vse podatke shranjuje lokalno. Ti podatki se ob prisotnosti mobilnega telefona preko brezžične povezave prenesejo na mobilni telefon.

Ko so podatki shranjeni na mobilni telefon jih je mogoče enostavno poslati v oblak, kar omogoča, da vrstniki med seboj tekmujejo oz. zbirajo točke za opravljene telesne aktivnosti. S tem jih spodbudimo k pogostejši telesni aktivnosti, kar je tudi bistveni namen zapestnice. Poleg otrok samih imajo vpogled v njihovo telesno aktivnost tudi njihovi starši in skrbniki, s čimer lahko preverijo, kako pogosto in intenzivno se otroci gibajo. Sistem je mogoče vključiti tudi v šolski sistem, pri čemer bi lahko učitelji športne vzgoje dodatno ocenjevali otrokovo aktivnost med in izven šolskih ur.

5. ZAPESTNICA ZA DETEKCIJO PADCEV

Najnaprednejša zapestnica, ki jo razvijamo je namenjena detekciji padcev. Padci so zelo nevarni predvsem pri starejših, saj ti težje pokličejo na pomoč. Delno lahko problem reši SOS zapestnica, vendar je lahko uporabnik prav pri hujših padcih v nezavesti in ne more pritistniti gumba za sprožitev alarma. V ta namen razvijamo zapestnico, ki padce samodejno detektira ter sproži alarm.

Zapestnica vsebuje pospeškomer, mikrokrmilnik, GPS in GSM modul, mikrofon, zvočnik, LED indikator, ter dve tipki za ročno proženje alarma. V času delovanja zapestnice, pospeškomer konstantno vzorči pospeške ter jih pošilja v obdelavo. Na mikrokrmilniku teče programska opremo, ki detektira ali se je dogodil padec, ali le običajen premik roke. Pri tem si pomagamo s kontekstom padca in ne le z absolutno vrednostjo pospeškov.

V primeru detektiranega padca zapestnica z zvočnim in svetlobnim indikatorjem nakaže, da je v fazi vzpostavitve zvočne povezave z izbranim kontaktom. V tem času lahko uporabnik klic prekliče v kolikor gre za lažni alarm. V primeru ko padec ni zaznan oz. uporabnik potrebuje pomoč zaradi kakšnega drugega razloga, pa lahko ročno pokliče kontakt z uporabo gumbov (kot pri SOS zapestnici).

Ključen del sistema je torej algoritem za detekcijo padcev [6]. V kolikor nam bo uspelo implementirati algoritem, ki z visoko točnostjo zazna padec in ima nizek odstotek napačnih alarmov, bo zapestnica uporabna tudi v realnem svetu. V ta namen imamo v načrtu eno letno testiranje zapestnic, s čimer želimo zbrati dovolj podatkov za izboljšanje algoritmov ter nastavitev parametrov.

Glavni namen zapestnice je podaljšati čas samostojnega življenja starejših oseb ter zmanjšati skrb najbližjih. Prav tako je lahko zapestnica uporabna tudi v domovih za ostarele in podobnih zavodih, kjer hitro ukrepanje vpliva na uporabnikovo kakovost življenja.

6. ZAKLJUČEK

V tem prispevku so opisane nosljive naprave razvite v podjetju Elgoline, ki uporabnikom pomagajo izboljšati kakovost življenja. Naprave so namenjene različnim tipom uporabnikom, saj ti potrebujejo različne načine pomoči. SOS zapestnica je namenjena starejšim ljudem, otrokom in ekstremnim športnikom, ki potrebujejo možnost hitrega klica v nevarnih situacijah. Športni nadzornik je naprava, ki pomaga profesionalnim športnikom in njihovim trenerjem bolje razumeti, kako posamezna vaja vpliva na športnikovo telo ter tako izboljšati kvaliteto treningov, kar omogoča hitro napredovanje in zmanjšanje tveganja poškodb. Tretja nosljiva naprava je zapestnica, ki je namenjena spodbujanju otrok h gibanju, saj meri njihovo aktivnost ter omogoča primerjanje s sovrstniki. Zadnja naprava, ki smo jo predstavili, prav tako zapestnica, pa omogoča detekcijo padcev in je namenjena predvsem starejšim, ki želijo dlje časa bivati samostojno oz. v osami. V vseh predstavljenih napravah lahko vidimo velik napredek v smislu izboljšanja kakovosti življenja, kar je tudi glavni namen nosljivih naprav.

7. VIRI

[1] Brigita Vrabič Kek. Kakovost življenja. *Statistični urad Republike Slovenije*, pages 28–31, 2012.

[2] S. Park and S. Jayaraman. Enhancing the quality of life through wearable technology. *Engineering in Medicine and Biology Magazine, IEEE*, 22(3):41–48, May 2003.

[3] B. Cvetkovic, R. Milic, and M. Lustrek. Estimating energy expenditure with multiple models using different wearable sensors. *Biomedical and Health Informatics, IEEE Journal of*, PP(99):1–1, 2015.

[4] Bozidara Cvetkovic, Vito Janko, and Mitja Lustrek. Demo abstract: Activity recognition and human energy expenditure estimation with a smartphone. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on*, pages 193–195, March 2015.

[5] Trupti Prajapati, Nikita Bhatt, and Darshana Mistry. Article: A survey paper on wearable sensors based fall detection. *International Journal of Computer Applications*, 115(13):15–18, April 2015.

[6] Fabio Bagalà, Clemens Becker, Angelo Cappello, Lorenzo Chiari, Kamiar Aminian, Jeffrey M. Hausdorff, Wiebren Zijlstra, and Jochen Klenk. Evaluation of accelerometer-based fall detection algorithms on real-world falls. *PLoS ONE*, 7(5), 05 2012.

Adaptive Drum Kit Learning System: Advanced Playing Errors Detection

Mladen Konecki
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin,
Croatia
+385 42 390873
mladen.konecki@foi.hr

ABSTRACT
In this paper, a brief overview of computer systems that have been developed in the domain of learning how to play a musical instrument will be given. It will also be described what kind of feedback information is available in these systems. A new adaptive learning model, which is based on a real-time adaptation to users' skills and knowledge, will be proposed. This model also includes advanced errors detection that enables better feedback, which is able to address the users' needs in a more precise manner and which is also able to give more precise feedback on the type of users' playing errors.

Categories and Subject Descriptors
I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search; J.5 [**Computer Applications**]: Arts and Humanities; K.3.1 [**Computers Mileux**]: Computer and Education

General Terms
Design, Experimentation, Human Factors.

Keywords
Adaptive drum kit learning system, error classification, advanced errors detection, playing errors, feedback.

1. INTRODUCTION
Music is very influential media among all people. It comes in many different forms and it is constantly changing. Unfortunately, the music education in not following this trend. Computer technology has potential to bring music education to the next level with the advanced multimedia content and interactivity. Computer technology has proven to be really suitable for increasing motivation and interest among students in the domain of music education [5; 15; 18]. Research has shown that concepts from computer games can also increase motivation in this domain [9; 12].

Research results have shown that there are many people who wanted to learn how to play musical instrument, but they never did [2]. A great number of them still want to learn how to play one. Computer technology can help in achieving this goal. People who want to learn in their own pace and people who cannot go to music schools can benefit from learning how to play a musical instrument by using computer technology. Success of this kind of systems relies on a reliable feedback that is given to the users: since the human teacher is not involved in the learning process, feedback about users' progress should come from the computer.

In section 2, computer systems that teach users how to play musical instrument will be described, along with their features. The focus will be on error detection in users' playing - what kind of feedback the systems gives to users. In section 3, a new adaptive learning model will be proposed, that is based on real-time adaptation to users' needs. New advanced error detection methods will also be proposed and discussed. The new advanced errors detection process should give more precise feedback to the users on their playing errors. This kind of feedback information should foster knowledge and skills acquisition.

2. RELATED WORK
In the scientific literature, a few computer systems can be found that teach users how to play a musical instrument. There are several systems for different instruments. There are also many commercial products in this domain, especially for piano playing and guitar playing, since those instruments are the most popular ones [2]. Table 1 shows the most prominent representatives for popular instrument categories.

Table 1. Computer systems for musical instrument playing education

Instrument category	Computer systems
Piano	PianoFORTE, Piano Tutor, Playground Sessions, Piano Marvel
Guitar & bass guitar	Rocksmith, Guitar Method, Bass Method
Drum kit	DT-1 V-Drums Tutor
Violin, viola, cello	i-Maestro, Digital Violin Tutor, My violin
Brass instruments	IMUTUS, VEMUS

Table 1 shows computer systems found in the scientific literature, along with some famous commercial products. All of these systems are based on visual display of notes that should be played, with note tracking features and error detection. Based on types of errors, feedback is given.

PianoFORTE [23] tracks dynamics, tempo, articulation and synchronization. Dynamics shows the loudness and hardness of playing single note. Tempo screen shows the differences in tempo that are made during playing by displaying the tempo curve. Articulation screen shows feedback on notes duration while synchronization screen shows the differences in timing of playing notes that should be played at the same time - weather one of the notes was played sooner or later compared to other note that

should be played at the same time. All of this feedback is given as a visual symbol on the screen. After finishing the lesson, users have to manually select which type of errors they want to see and manually examine where they played something wrong.

Piano Tutor [3] is recording pitch, note starting time, duration and loudness. Those are considered "primitive errors". Afterwards, deeper analysis is performed where the pattern in error types is being searched. If there is a pattern that was found, then the system gives some lesson for error correction. Pitch errors are more important than rhythm or dynamics errors.

Playground Sessions [19] and Piano marvel [17] are commercial products that are similar in their basic features: note tracking with feedback on which note was played correctly. They both use game concepts for motivation: getting trophies and stars. They don't have any advanced error detection or feedback lessons for particular type of error.

The most popular guitar learning system is actually a computer game Rocksmith [21]. Many interesting concepts are implemented in this computer learning system. There are guitar lessons for learning chords, mini games where one can play certain chords and notes to progress, session mode where one can freely play with the backing tracks. The main part of the game is playing popular songs. The difficulty of the games adapts to the user - if the user has low percentage of correct notes, some notes are removed so the song becomes easier to play. Since it is a game, it does not give any feedback on errors except from showing which note was played correctly.

Guitar Method [8] and Bass Method [1] systems are developed by the same developer and are similar in their features. They have basic note tracking with real-time feedback on which note was played correctly along with animated fingerboard where the user can see how to play certain notes and chords. There are no advanced error detection methods.

There is a lack of computer systems for learning how to play drums. There is only one commercial product, Roland DT-1 V-Drums Tutor [4]. DT-1 does not have any advanced error detection features, it just has a database of lessons along with note tracking feature where the user can see in real-time which note was played correctly. Like many other systems, the tempo of the lesson can be manually changed if the lesson is too hard.

i-Maestro is an interactive multimedia environment for technology enhanced music education of string instruments [14]. It is a framework that supports production and authoring of music scores, gesture analysis, augmented instruments, audio processing, score following and theory training. It also supports cooperative work and distributed learning. Feedback on playing is given in many forms: augmented mirror shows graph of bow movement while playing. This framework was never intended to be standalone application without teacher's involvement so errors are analyzed manually.

Digital Violin Tutor was developed to enhance daily practice of violin players [16]. It provides feedback when teacher is not available. It offers different visual modalities - video of a user playing, 3D animation of body posture, 2D animations of the fingerboard and graphical display of notes to play. It can detect mistakes like wrong pitch and wrong timing and show where the errors were made [24].

My violin [13] is a commercial product that features animated fingerboard, video lessons, interactive games and backing tracks. Like in Guitar and Bass Method systems, it provides same interface for score following and showing which notes were played correctly.

IMUTUS [22] project is a complete, autonomous tutor. Testing of this system has been done by detecting errors in playing recorder. It detects pitch, tempo, articulation, attack, airflow and intonation mistakes based on played notes. Only small amount of mistakes are reported to not overwhelm the user so the prioritization of mistakes are made and those most important are displayed. Like in other systems, pitch mistakes have higher priority than rhythm or dynamics errors.

VEMUS [6] project is the successor of IMUTUS project. The idea behind this project is creating "Virtual European Music School". One part of this project is creating automatic performance evaluation. Feedback on the played errors is given in the form of messages and annotations of symbols on the music score and in the form of curves showing pitch, envelope and harmonics.

As can be seen, there are quite a few computer systems for learning how to play a musical instruments. Most of them have simple error detection and some of them give more advanced feedback, based on the type of mistakes.

3. ADAPTIVE DRUM KIT LEARNING SYSTEM

3.1 Proposed model

In previous research, a new model for adaptive drum kit learning system that adapts to the user's skills in real-time was presented [10]. In this model, the knowledge base consists mainly from drum patterns that should be learned.

While playing, user's accuracy is measured for each pattern and based on the percentage of accuracy, the lessons are dynamically adapting to the user's skills. If the percentage is high, that means the user has no problems with the current patterns and new harder patterns are introduced. On the other hand, if the accuracy is low, then harder patterns are removed and the focus is put on an easier pattern. When the user handles easier patterns with high accuracy, then harder patterns are presented.

Also, after each lesson (that lasts several minutes), types of mistakes are evaluated and special lesson for error removal is generated. Types of mistakes are evaluated and based on the most common types of mistakes.

Lesson for error removal is generated from the pool of drum patterns available in the knowledge base. In this paper this module will be discussed and the answers to the question about what kind of mistakes can and should be detected in adaptive learning system for playing drum kit will be given.

Previous research has shown that this adaptability feature is very important to users: adaptability to the user's tempo of progress and skills. Content and order of the lessons should also adapt individually for every user [11].

This kind of systems allows learning at the pace that the user determines. Figure 1 shows the basic model of proposed adaptive drum kit learning system.

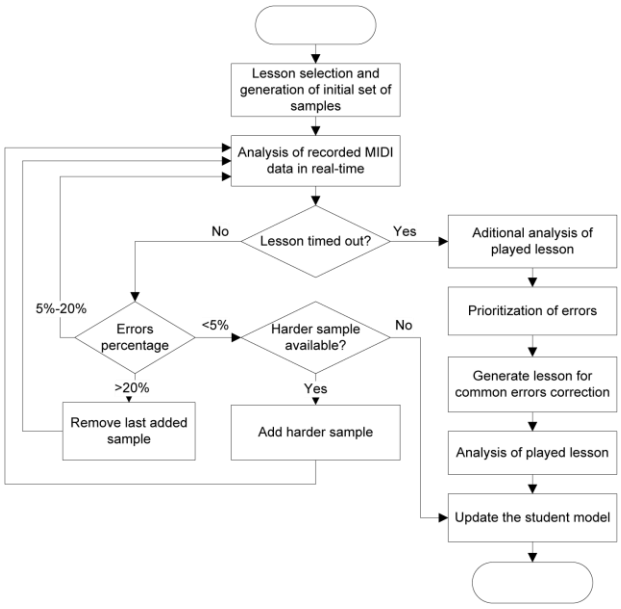


Figure 1. Proposed conceptual model of adaptive drum kit learning system.

3.2 Advanced playing errors detection

After the lesson that consists of many drum patterns that are being adaptively presented in real-time, advanced analysis of played patterns is conducted. Simple feedback would be to present which errors occurred during the lessons. Users could see which errors were made and they could decide what to practice next. But to make computer system act more like a human teacher, system should be able to examine all the errors, determine which are the most important and decide what the user should practice to avoid doing the same errors in the future.

If the playing errors are examined just on the level of each note separately, some advanced errors that human teacher could notice while listening to students playing would not be noticed. There are several types of errors that should be examined. Firstly, "simple errors" that are based on single notes are detected and then "advanced errors", that are result of cumulative simple errors, can also be detected.

Simple errors that can be detected are:

- tempo errors - precise detection of the time when the note was played; notes can be played earlier, at the right time or too late
- detection of a note that was played but should not be played or a note that wasn't played but should be played
- pitch errors - is there a note that was played but was the wrong note
- dynamics - was the note played with proper hardness
- note duration - was the note played with the right duration

Most of these errors can be detected by the other computer systems that were developed in this domain, but in most cases, there are not specific lessons that are given after some of these errors are detected: user must see and analyze playing and then try to correct the errors without any help.

But there are many more variations of errors that could be detected. Some of the advanced ones are:

- Various synchronization errors - problems in synchronization when more than one note should be played at the same time, a specific combination that is making more problems for the user
- Various dynamics errors - dynamics of the notes that should be played at the same time, detection of specific combination that is problematic for the user
- Pattern errors - when error in a pattern occur, is there a specific spot in the pattern or situation or combinations of notes that are harder for the user to play, detection why the user has made the mistake

3.2.1 Simple error detection

Even by detecting just simple errors, there are various levels of complexity of the detection and feedback that could be given to the user. If an example that includes simple tempo errors is examined it can be stated that nobody plays perfectly, not even graduated musicians. They all have small deviations in timing. First question is how much milliseconds can user deviate from the perfect timing while still playing the note "in time".

To determine this parameter, more than 200 drum patterns played by professional drummer have been recorded. Not all patterns were played "perfectly" and the drummer had to decide afterwards what notes were "in time" and what notes have been played off tempo. Then the timing deviation in all the notes that were marked as "in time" was examined. For the 60 beats per minute tempo, 1 beat is 1000ms long. The biggest deviation that was marked as "in time" was 125ms. Audio interface had 24ms input/output delay so the final result was approximately ±100ms. This deviation is connected to the tempo of the beat and has to be relatively expressed. Based on the research, notes that are ±10% off the beat are considered as "in time". In a standard 4/4 measure rhythm, this value is ±2,5% of the duration of one bar. Some famous computer games like Guitar Hero [7] or Rockband [20] slice their "in time" interval even more so notes have rankings from "Perfect" to "OK" but it is hard to determine what time interval was used. By detecting proper deviation, more precise feedback can be given to users. Although details about the exact amount of deviation time are not that important to the user, this measure is the basis for advanced error detection that could be really valuable to the user.

3.2.2 Advanced error detection

By analyzing simple errors in the context of the pattern or by analyzing sequence of errors in the lesson, a better feedback on played errors can be provided.

When there is a note that was played but should not be played, a simple feedback about that error can be given. But to determine the reason why this error happened, there is a need to analyze the drum pattern where the error occurred. For example, common mistake among beginners comes from the fact that it is hard for them to separate hits with the right hand from the hits with the right leg. A common mistake comes from the fact that when only right leg needs to play note, right hand usually follows. The result is a note that was played but should not be played. By looking at the context, it can be seen what combination of hands/legs is making problems and by counting numbers of mistakes for each combination, it can be determined if one of the combination is

particularly hard for the user. Figure 2 shows example of this kind of mistake.

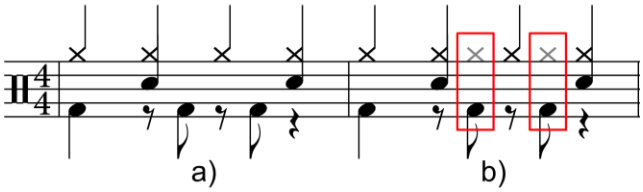


Figure 2. a) Drum pattern b) Common playing error.

If that is the case, then a proper lesson that deals with this kind of error can be given by looking into the knowledge base of drum patterns and by picking the pattern with this kind of event. A lesson that consists of drum patterns with this event should be generated and presented to the user.

The same principle can be applied to other simple errors, for example to the dynamics of the notes. By detecting single note errors and displaying them on the screen, user can see where the errors were made and the user has to conclude if there is a pattern in dynamics errors. By examining context in which the error occurred, conclusions can be made. For example, for a basic rock beat kick and snare are the most important elements of the beat. Hi-hat should be played quieter so kick and snare stand out. For beginners, when they need to hit something with both hands at the same time with different dynamics, it is a problem to do that. So what happens is that hi-hat is played quietly, but when the user needs to play snare along with hi-hat, the hi-hat is usually hit harder. Figure 3 shows this kind of error.

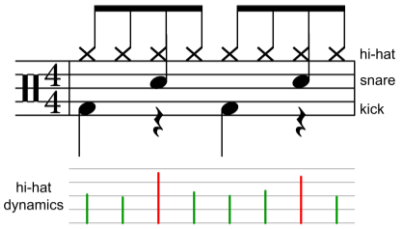


Figure 3. Drum pattern with hi-hat dynamics.

So by looking at the context when some dynamics error occurred, a conclusion about the problem that the user is dealing with can be given and a proper feedback for that exact error type can also be presented.

Same rule can apply to almost any type of error. One of the common errors is also speeding up or slowing down. PianoFORTE is a system that introduced graph that shows these tempo deviations. User can see this curve and where the mistakes were made. But again, by analyzing context, a better conclusion can be provided. The type of patterns where this error occurred or part of the lesson (or song) where this happened can be examined. Then it can be seen whether there is a specific drum pattern that is making the user speed up or slow down.

Common place where this happens is when drum fill comes after a drum beat or in a context of a song, when there is a transition from verse to chorus and vice versa. Intensity of the backing track is usually changing in these transitions and that can influence the drummer to change tempo. By examining the context of the error, more precise conclusions can be made and better feedback can be provided.

By having meta-data about the drum patterns that are used in lessons, the type of the errors based on the drum pattern type can be determined. Simple classification of drum patterns for created prototype is shown in Figure 4.

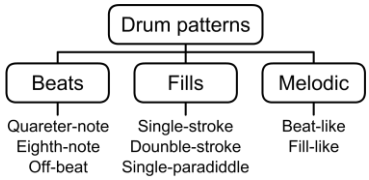


Figure 4. Simple drum patterns classification.

Once a drum patterns classification is provided, it can be used to determine if there is a connection between the type of pattern and the amount of errors. For example, it can be concluded that there are significantly more errors in the fills category, especially double-stroke fills. Then a feedback can be given and lesson that consists of this kind of patterns, that will help improve user's playing in this area, can be generated.

Another common problem in drum playing is alignment or synchronization of notes that should be played at the same time. If the time difference between notes that should be played at the same time is bigger than 2,5% of the bar duration, that means those notes are not completely synchronized. By looking at the context where this type of error happened, it can be concluded that for example, the snare hit is always a bit late after the hi-hat hit. By detecting more specific error type, a better feedback can be provided.

Some of these common errors could be detected by users examining the simple errors, but it would be much better if computer system could do that automatically.

So far, some examples of advanced error detection for adaptive drum kit learning system for the most common error types have been mentioned. Every instrument has its own characteristics and common errors so for every instrument a list of specific cases should be made by experts. When common error types are defined, lessons for their removal should be defined. But for each instrument and any number of cases, the same method could be applied: first simple error detection should be made that will be the basis for advanced error detection. Then more advanced errors should be detected, based on music patterns or bars. Lastly, errors that are in the context of a lesson or a song can be detected so more advanced and specific feedback can be given. Special lessons could be also generated for dealing with some specific error types. Figure 5 shows this error detection process.

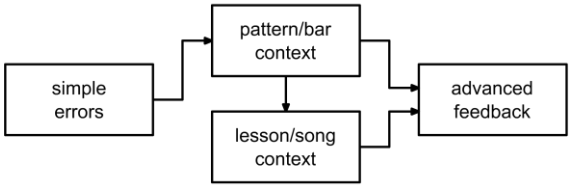


Figure 5. Advanced error detection process.

Some errors are based on the music patterns/bars and feedback can be given right after their detection (combination of notes that is hard for user to play, specific part of the pattern that is making problems, specific dynamics problem). Some errors must be searched in the context of the whole lesson (slowing down, speeding up, detection of pattern types that are problematic).

When significant statistical difference in the number of any error type is detected, advanced feedback can be given, along with generated lesson for that particular problem.

4. CONCLUSION AND FURTHER RESEARCH

In this paper an advanced error detection of users' playing in the context of adaptive learning systems that teach users how to play drum kit has been described. The principle that was described could be applied for other instruments too, by defining a list of common errors and applying detecting based not on single notes, but on patterns/bars and lessons/songs. More precise feedback could be given to the user and that can result in better knowledge/skill acquisition. To determine the impact of proposed advanced feedback detection in the learning process, an evaluation will be conducted. By comparing two groups, one that will use the system with simple error detection and one that will use advanced error detection, it will be possible to exactly determine if and how much this approach can help in the learning process.

This approach could be the basis for the development of a learning system that will define the problematic patterns of errors on its own and that will be able to find the patterns that are hard for the users. Then the specific lessons could be created for those patterns. Although proposed method does not include learning at the moment, it is a step forward from simple error detection.

5. REFERENCES

[1] Bass Method. *http://www.emediamusic.com/bass-lessons/bass-method.html* Accessed: 2015-08-12.

[2] Block, K. The 2011 NAMM Global Report. *http://www.nxtbook.com/nxtbooks/namm/2011globalreport/index.php#/1* Accessed: 2015-08-11.

[3] Dannenberg, R. B., Sanchez, M., Joseph, A., Capell, P., Joseph, R., and Saul, R. 1990. A computer-based multi-media tutor for beginning piano students. *Journal of New Music Research.* 19, 2-3 (1990) 155-173. DOI= 10.1080/09298219008570563.

[4] DT-1. *http://www.roland.com/products/dt-1/* Accessed: 2015-08-13.

[5] Finney, J., and Burnard, P. 2010. Music education with digital technology. Bloomsbury Publishing, New York.

[6] Fober, D., Letz, S., and Orlarey, Y. 2007. VEMUS-Feedback and groupware technologies for music instrument learning. *In Proceedings of the 4th Sound and Music Computing Conference SMC.* 7 (July 2007), 117-123.

[7] Guitar Hero. *https://en.wikipedia.org/wiki/Guitar_Hero* Accessed: 2015-08-16.

[8] Guitar Method. *http://www.emediamusic.com/guitar-lessons/beginning-guitar-method.html* Accessed: 2015-08-12.

[9] Klimmt, C. 2003. Dimensions and determinants of the enjoyment of playing digital games: A three-level model. *In Level up: Digital games research conference.* (Nov. 2003), 246-257.

[10] Konecki, M. 2014. Learning to play musical instruments through dynamically generated lessons in real-time based on adaptive learning system. In *Central European Conference on Information and Intelligent Systems* (Sep. 2014), 124-129.

[11] Konecki, M. 2015. Self-Paced Computer Aided Learning of Music Instruments. In *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*. (May 2015), 809-813. DOI= 10.1109/MIPRO.2015.7160382.

[12] Koster, R. 2013. *Theory of fun for game design*. O'Reilly Media, Inc., CA.

[13] My violin. *http://www.emediamusic.com/kids-violin-lessons/my-violin.html* Accessed: 2015-08-15.

[14] Ng, K., and Nesi, P. 2008. i-Maestro framework and interactive multimedia tools for technology-enhanced learning and teaching for music. In *Automated solutions for Cross Media Content and Multi-channel* Distribution, 2008. AXMEDIS'08. International Conference on (Nov. 2008), 266-269. DOI=10.1109/AXMEDIS.2008.41.

[15] Pachet, F. 2004. Beyond the Cybernetic Jam Fantasy: The Continuator. *IEEE Comput. Graph. Appl.* 24, 1 (January 2004), 31-35. DOI=10.1109/MCG.2004.1255806

[16] Percival, G., Wang, Y., and Tzanetakis, G. 2007. Effective use of multimedia for computer-assisted musical instrument tutoring. In *Proceedings of the international workshop on Educational multimedia and multimedia education* (Sep. 2007), 67-76. DOI=10.1145/1290144.1290156.

[17] Piano Marvel. *https://pianomarvel.com/* Accessed: 2015-08-14.

[18] Pitts, A., and Kwami, R. M. 2002. Raising students' performance in music composition through the use of information and communications technology (ICT): a survey of secondary schools in England. *British Journal of Music Education*. 19, 1 (2002), 61-71.

[19] Playground Sessions. *http://www.playgroundsessions.com/* Accessed: 2015-08-14.

[20] Rock Band. *http://www.rockband4.com/* Accessed: 2015-08-16.

[21] Rocksmith. *http://rocksmith.ubi.com/rocksmith/en-us/home/* Accessed: 2015-08-12.

[22] Schoonderwaldt, E., Hansen, K., and Askenfeld, A. 2004. IMUTUS–an interactive system for learning to play a musical instrument. In *Proceedings of the International Conference of Interactive Computer Aided Learning* (2004), 143-150.

[23] Smoliar, S. W., Waterworth, J. A., and Kellock, P. R. 1995. pianoFORTE: a system for piano education beyond notation literacy. In *Proceedings of the third ACM international conference on Multimedia* (Nov. 1995), 457-465.

[24] Yin, J., Wang, Y., and Hsu, D. 2005. Digital violin tutor: an integrated system for beginning violin learners. In *Proceedings of the 13th annual ACM international conference on Multimedia* (Nov. 2005), 976-985. DOI=10.1145/1101149.1101353.

Expanding the OntoDM ontology with network analysis tasks and algorithms

Jan Kralj[1,2]
jan.kralj@ijs.si

Panče Panov[1]
pance.panov@ijs.si

Sašo Džeroski[1,2]
saso.dzeroski@ijs.si

[1]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper presents the first steps towards integrating concepts from the field of network analysis into the OntoDM ontology of data mining concepts. We have performed an extensive analysis of different subfieds of network analysis to provide a broad overview of the variety of tasks and algorithms that are encountered in the field. The main part of this work was to categorize the tasks and algorithms into a hierarchy that is consistent with the structure of OntoDM and which can systematically cover as many aspects of network analysis as possible. This work is a first step in the direction of OntoDM becoming an ontology that systematically describes not only data mining, but also network analysis. We believe that this work will encourage other researchers working in the filed to provide additional insight and further improve the integration of this field into OntoDM.

1. INTRODUCTION

Network analysis, is a large and quickly growing scientific discipline connected to physics, mathematics, social sciences and data mining. The tasks, tackled by the experts in the field, range from detecting communities in a given network, through predicting links in incomplete or time evolving networks, to ranking or classifying vertices of a given network. Most such tasks are analyzed in the context of information networks in which all nodes are treated equally, but in recent years, the concept of *heterogeneous* information networks [43], a generalization of standard information networks (which are then referred to as *homogeneous*), is gaining popularity.

OntoDM [35, 36] is a reference modular ontology for the domain of data mining. It is directly motivated by the need for formalizing the data mining domain and is designed and implemented by following ontology best practices and design principles. It includes the terms neccessary to describe different types of data, data mining tasks and approaches to solving these tasks. Among the key OntoDM classes are the classes representing datasets, data mining tasks, generalizations and algorithms themselves. The latter three classes are interconnected, as each data mining algorithm solves some data mining task by producing an output which is some type of generalization. In our work, we have expanded these classes to include tasks and algorithms that are found in the study of network analysis.

2. ONTOLOGY EXTENSION

This section presents an overview of the ontology classes that we have added to the OntoDM ontology. Shown in Figure 1, they consist mainly of subclasses of the classes *data mining task* and *data mining algorithm*.

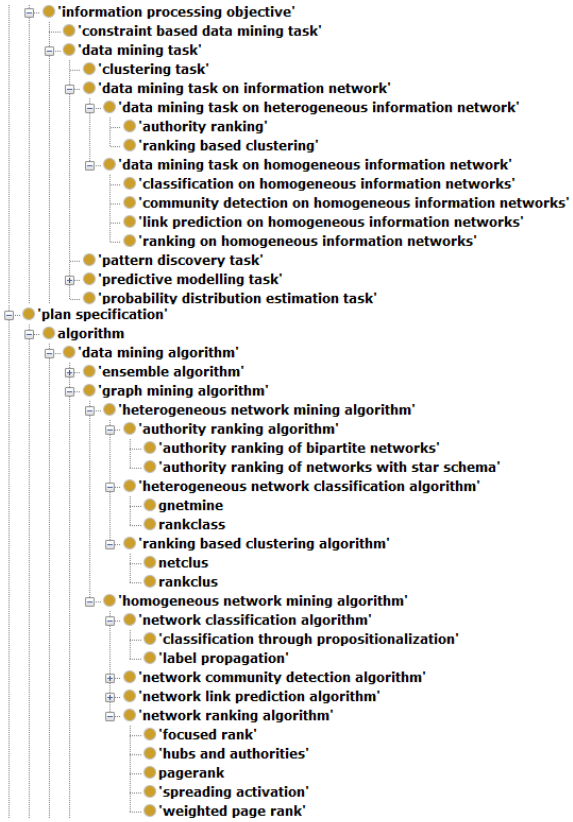


Figure 1: The hierarchical structure of the main network analysis additions to OntoDM

As a subclass of the data mining task class, we have added a new class of data mining tasks, *data mining task on information network*, which includes all tasks encountered in our overview of the field. The tasks are first split into tasks that can be defined on a general (homogeneous or heterogeneous) information network and those that can only be defined on a heterogeneous network. Tasks on general networks, which constitute the majority of the new entries, include link prediction, community detection, ranking, and classification.

The classes, added under the *data mining algorithm* class, are gathered into a separate parent class *network analysis algorithm*. Each algorithm constitutes a leaf node in the hierarchy rooted in this parent class. The hierarchical structure of network analysis algorithms follows the hierarchy of tasks, described in the previous paragraph. Each presented algorithm solves a particular task which lies in a analogous part of the ontology – heterogeneous network analysis [43], link prediction algorithms [29, 2], community detection methods [12, 37], network ranking algorithms [11] and network classification algorithms. Furthermore, for each network analysis algorithm, we added a short description presenting the key concepts of the algorithm, as well as references to the paper in which the algorithm was presented. The descriptions and references are added as annotations to the classes in the ontology. Note that some terminal nodes of the hierarchy are actually instances of the classes, while other are proper classes that still need to be populated.

The final concept we added to the ontology was the concept of generalization specifications. Generalization specifications describe the types of output given by network analysis algorithms. Because the outputs of network analysis tasks are fundamentally different from outputs of traditional data mining algorithms, we decided to construct a hierarchy of generalization specifications, following the hierarchy of network analysis tasks and algorithms.

3. DATA MINING TASKS

In this section, we present the classes, added to the *data mining task* class. We present a short description of each class of data mining tasks that was added to OntoDM.

3.1 Data mining tasks on general networks

Data mining tasks on general networks are data mining tasks that can be formulated on any (homogeneous or heterogeneous) network. Commonly, different algorithms are used to perform the same task on homogeneous and heterogeneous networks.

Classification. Classification of network data is a natural generalization of classification tasks encountered in a typical machine learning setting. The problem formulation is simple: given a network and class labels for some of the vertices in the network, predict the class labels for the rest of the vertices in the network. The output of a classification task on a network is a function that predicts the class label of each vertex in the network.

Link prediction. While classification tasks try to discover new knowledge about network entities, link prediction focuses on unknown connections between the entities. The assumption is that not all network edges are known. The task of link prediction is to predict new edges that are missing or likely to appear in the future. The output of an algorithm solving a link prediction task is a function which provides a proximity measure for each pair of vertices in a network. The pairs with the highest proximity measure are then assumed to be the most likely candidates for predicted links.

Community detection. While there is a general consensus on what a network community is, there is no strict definition of the term. The idea is well summarized in the definition by Yang et al. [49]: a community is a group of network nodes, with dense links within the group and sparse links between groups. The output of a community detection algorithm is similar to the output of a clustering algorithm: a function that assigns each vertex in the network to a community.

Ranking. The objective of ranking in information networks is to assess the relevance of a given object either with regard to the whole graph or relative to some subset of vertices in the graph. In either case, the output of a network ranking algorithm is a function that assigns a score to each vertex of the network. The vertices with the highest score are then ranked the highest.

3.2 Data mining tasks on heterogeneous networks

Data mining tasks on heterogeneous networks are tasks that can only be formulated on a heterogeneous network. Unlike the tasks on general networks, these tasks have only been addressed in recent years.

Authority ranking. Sun and Han [43] introduce *authority ranking* to rank the vertices of a heterogeneous network with either a bipartite structure or a star network schema in which one vertex type is central in that all network edges start or end on a vertex of this central type. The task of authority ranking is to rank vertices in each (not necessarily all) vertex type separately, and the output of the task is a collection of functions, each assigning a score to only vertices of a certain type.

Ranking based clustering. While both ranking and clustering can be performed on heterogeneous information networks, applying only one of the two may sometimes lead to results which are not truly informative. For example, simply ranking authors in a bibliographic network may lead to a comparison of scientists in completely different fields of work which may not be comparable. Sun and Han [43] propose joining the two seemingly orthogonal approaches to information network analysis (ranking and clustering) into one, in which vertices are simultaneously assigned to a cluster and given a score to rank them within the cluster.

4. ALGORITHMS

This section presents the algorithms that are classified in the modified ontology. The classification hierarchy of network analysis algorithms is similar to the hierarchy of data mining tasks, described in Section 3.

4.1 General network mining algorithms

This section describes the algorithms that can be used to solve data mining tasks on general networks.

Community detection algorithms. The classification of community detection algorithms on networks follows the classification of algorithms, described in the surveys by Fortunato [12] and Plantié and Crampes [37]. The algorithms can be split into several classes based on the underlying idea that guides the algorithms. It must be noted that a strict split of the different methods is impossible as different methods are not developed in isolation. For example, many methods that are not strictly classified as modularity based

algorithms still use the concept of modularity in one of their steps.

Divisive algorithms. Divisive algorithms are algorithms that find a community structure of a network by iteratively removing edges from the network. As edges are removed, the network decomposes into disconnected components. The decomposition pattern forms a hierarchical clustering over the set of all vertices in the network. The most widely used such algorithm is the Girvan Newman algorithm [16], which removes the edges in the network with the largest centrality measure, arguing edges which are more central to a graph are the edges that cross communities. An alternative algorithm is the Radicchi algorithm which calculates the edge clustering coefficient of edges to calculate which edges must be removed. Here, the intuition is that edges between communities belong to fewer cycles than edges within communities.

Modularity based algorithms. Modularity based algorithms form the majority of community detection algorithms. While the concept of modularity (first defined in Newman and Girvan [33]) is used in almost all algorithms at some point (especially to determine the best clustering from a hierarchical clustering of nodes), the algorithms in this class use modularity more centrally than other algorithms. The most prominent such methods are the Louvain algorithm [5] and the Newman greedy algorithm [33]. Other methods include variations of the greedy algorithm [46], using simulated annealing [18], spectral optimization of modularity via a modularity matrix [32, 31] or via the graph adjacency matrix [47], and deterministic optimization approaches [10].

Spectral algorithms. Spectral algorithms find communities in network by analyzing eigenvectors of matrices, derived from the network. The community structure is extracted either from the eigenvectors of the Laplacian matrix of the network [9] or from the stochastic matrix of the network [6]. In both cases, algorithms assume that eigenvectors, extracted from the network, will have similar values on indices that belong to network vertices in the same community. The computation of several eigenvectors belonging to the largest eigenvalues is first performed. The eigenvectors form a set of coordinates of points, each belonging to one network vertex. Clustering of the points then corresponds to community detection of network vertices.

Random walk based algorithms. Random walk based algorithms are algorithms that use the concept of a random walker on a network to perform community detection. The methods use a random walker model to determine the similarities of network vertices and then use either a divisive [51] or an agglomerative [52, 38] approach to construct a hierarchical clustering of the nodes.

Link prediction algorithms. Link prediction algorithms are presented in survey papers by Lü and Zhou [29] and Al Hasan and Zaki [2]. These surveys present a similar hierarchy of link prediction algorithms, which we also used in the construction of the ontology. All presented algorithms calculate a proximity measure between two vertices. They do so in 3 distinct ways, described below.

Similarity based algorithms. Similarity based algorithms calculate the proximity of two vertices in the network either from their neighborhoods (*local similarity based algorithms*) or from the way the two vertices fit into the overall network structure (*global similarity based algorithms*). Local similarity based algorithms are further divided into common neighbor based algorithms and vertex degree based algorithms. The first class of algorithms computes the similarity between two vertices purely from the number of neighbors of each node, and the number of common neighbors, while the second class also takes the degrees of both nodes into account. The most widely used algorithm in this class is the algorithm for calculating the Adamic-Adar proximity measure [1]. Other proximity measures listed are the common neighbors [30], the hub depressed and hub promoted indices [39], the Jaccard index, the Leicht-Holme-Newman index [28], the Salton index [40], the Sorensen index [42] and the preferential attachment index [3]. Unlike local similarity based algorithms, global similarity based algorithms use the entire network structure to calculate the proximity between two network vertices. The algorithms include the Katz index [25], the random walk with restart [41], the SimRank [22], the average commute time index [26] and the matrix forest index [7].

Probability based algorithms. Probabilistic algorithms for link prediction use various techniques to estimate the probability that a pair of vertices should be connected. These maximum likelihood methods, like the hierarchical structure model [8] and the stochastic block model [19], and probabilistic models, like the probabilistic relational model [15], probabilistic entity relationship model [20] and stochastic relational models [50].

Network ranking algorithms. The classification of network ranking algorithms adopted in this work was guided by the paper by Duhan et al. [11]. However, this paper is not as detailed as the survey papers for the link prediction and community detection tasks. The paper focuses on the classification of web pages and describes several algorithms for ranking vertices in a network. For this work, only the methods that deal with ranking nodes in a network were used. The methods include the famous PageRank algorithm [34] used by the Google search engine and a weighted version of the PageRank method called the Weighted PageRank [48], as well as the related Hubs and Authorities method [27]. Another method to rank nodes in the network is to use centrality measures. To construct a collection of network centrality measures, we followed the lecture given by dr. Cecilia Mascolo [1]. The network centrality measures listed in the ontology are Freeman's Network Centrality [14], betweenness centrality [13], closeness centrality [4] and the Katz centrality measure [25].

Network classification algorithms. The most widely used network classification algorithm is the label propagation algorithm [52]. Another algorithm based on nwtwork propositionalization [17] can also be used to classify nodes in a homogeneous network.

[1]https://www.cl.cam.ac.uk/teaching/1314/L109/stna-lecture3.pdf

4.2 Heterogeneous network mining algorithms

This section describes the algorithms used to solve the data mining tasks, described in Section 3.2.

Authority ranking. Authority ranking, as presented in [43], can be adressed by the algorithms for authority ranking in networks with a bipartite structure and in networks with a network schema.

Ranking based clustering. Ranking based clustering, as presented in [43], is adressed similarly to authority ranking. Sun et al. [44] present the algorithm RankClus, which performs ranking based clustering on bipartite networks, and Sun et al. [45] present the NetClus algorithm which tackles the same task on networks with a star network schema.

Classification in heterogeneous networks. The algorithms in this class can be used to classify nodes in a heterogeneous network. They are the algorithm by Grčar et al. [17] which uses network propositionalization to classify network vertices, the RankClass [23], the GNetMine [24] algorithm and the heterogeneous network propagation algorithm [21].

5. CONCLUSION AND FURTHER WORK

The field of network analysis is a rich and complex field. This work presents a starting point for integrating the descriptions of tasks and algorithms ito OntoDM. In the future, we wish to add several subfields of network analysis that were not analyzed in this paper, such as the analysis of data enriched networks, time evolving networks and a separate analysis of community detection algorithms for directed networks. Furthermore, the ontology can be expanded to include example datasets on which algorithms can be tested, as well as evaluation metrics to examine the performance of various algorithms.

Acknowledgement. We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

[1] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3):211–230.

[2] Al Hasan, M. and Zaki, M. J. (2011). A survey of link prediction in social networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 243–275. Springer.

[3] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[4] Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*.

[5] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[6] Capocci, A., Servedio, V. D., Caldarelli, G., and Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2):669–676.

[7] Chebotarev, P. Y. and Shamis, E. (1997). A matrix-forest theorem and measuring relations in small social group. *Avtomatika i Telemekhanika*, (9):125–137.

[8] Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

[9] Donetti, L. and Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10):P10012.

[10] Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104.

[11] Duhan, N., Sharma, A., and Bhatia, K. K. (2009). Page ranking algorithms: A survey. In *2009 IACC Advance Computing Conference*, pages 1530–1537. IEEE.

[12] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

[13] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[14] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

[15] Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *16th International Joint Conference on Artificial Intelligence*, volume 99, pages 1300–1309.

[16] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[17] Grčar, M., Trdin, N., and Lavrač, N. (2013). A methodology for mining document-enriched heterogeneous information networks. *The Computer Journal*, 56(3):321–335.

[18] Guimera, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

[19] Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078.

[20] Heckerman, D., Meek, C., and Koller, D. (2007). Probabilistic entity-relationship models, PRMs, and plate models. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*, pages 201–238. MIT Press.

[21] Hwang, T. and Kuang, R. (2010). A heterogeneous label propagation algorithm for disease gene discovery. In *10th SIAM International Conference on Data Mining*, pages 583–594.

[22] Jeh, G. and Widom, J. (2002). SimRank: a measure of structural-context similarity. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543. ACM.

[23] Ji, M., Han, J., and Danilevsky, M. (2011). Ranking-based classification of heterogeneous information networks. In *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1298–1306. ACM.

[24] Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 1, pages 570–586. Springer-Verlag.

[25] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

[26] Klein, D. J. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95.

[27] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

[28] Leicht, E., Holme, P., and Newman, M. E. (2006). Vertex similarity in networks. *Physical Review E*, 73(2):026120.

[29] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170.

[30] Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102.

[31] Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104.

[32] Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[33] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

[34] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

[35] Panov, P., Džeroski, S., and Soldatova, L. N. (2008). OntoDM: An ontology of data mining. In *2008 IEEE International Conference on Data Mining Workshops*, pages 752–760. IEEE Computer Society.

[36] Panov, P., Soldatova, L., and Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28(5-6):1222–1265.

[37] Plantié, M. and Crampes, M. (2013). Survey on social community detection. In Ramzan, N. e. a., editor, *Social Media Retrieval*, pages 65–85. Springer.

[38] Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *20th International Symposium on Computer and Information Sciences*, pages 284–293. Springer.

[39] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

[40] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. Mcgraw-Hill College.

[41] Shang, M.-S., Lü, L., Zeng, W., Zhang, Y.-C., and Zhou, T. (2009). Relevance is more significant than correlation: Information filtering on sparse data. *Europhysics Letters*, 88(6):68008.

[42] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.

[43] Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers.

[44] Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. (2009a). RankClus: integrating clustering with ranking for heterogeneous information network analysis. In *2009 International Conference on Extending Database Technology*, pages 565–576.

[45] Sun, Y., Yu, Y., and Han, J. (2009b). Ranking-based clustering of heterogeneous information networks with star network schema. In *15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

[46] Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. In *16th international Conference on the World Wide Web*, pages 1275–1276. ACM.

[47] White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In *5th SIAM International Conference on Data Mining*, volume 5, pages 76–84. SIAM.

[48] Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *2nd Annual Conference on Communication Networks and Services Research*, pages 305–314. IEEE.

[49] Yang, B., Liu, D., and Liu, J. (2010). Discovering communities from social networks: Methodologies and applications. In Furht, B., editor, *Handbook of Social Network Technologies and Applications*, pages 331–346. Springer.

[50] Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2006). Stochastic relational models for discriminative link prediction. In *2006 Conference on Advances in Neural Information Processing Systems*, pages 1553–1560.

[51] Zhou, H. (2003). Network landscape from a Brownian particle's perspective. *Physical Review E*, 67(4):041908.

[52] Zhou, H. and Lipowsky, R. (2004). Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *2004 International Conference on Computational Science*, pages 1062–1069. Springer.

Power Negotiations in Smart Cities

Matej Krebelj, B.Sc.
"Jožef Stefan" Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
+386 1 477 39 00
matej.krebelj@ijs.si

Matjaž Gams, Ph.D.
"Jožef Stefan" Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
+386 1 477 39 00
2nd E-mail

Aleš Tavčar, B.Sc.
"Jožef Stefan" Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
+386 1 477 39 00
3rd E-mail

ABSTRACT
In this paper, we describe detailed overview of Power negotiations in Smart homes located in Smart city. Smart homes are located in Smart city, which has Management system for different subsystems (water supply, traffic control etc.). One of domains of City management is also electric power supply and demand. In order to avoid power spikes we developed negotiating system which tries to reduce peak power consumption.

Categories and Subject Descriptors
I.2.11 [**Multiagent systems**]: Multiple actors acting as agents – *consumer agents, distributed system, negotiation agents.*

General Terms
Your general terms must be any of the following 16 designated terms: Algorithms, Management, Measurement, Performance, Design, Economics, Reliability, Experimentation and Standardization.

Keywords
Power negotiations, agent based system, smart city management.

1. INTRODUCTION
This work focuses on Power negotiation algorithm implemented into Smart City Management System (SCMS). SCMS is complex system, part of Adaptive Cooperative Control in Urban (sub)Systems (ACCUS) project [1], which consists of many independent subsystems. Its main focus is to improve citizens' quality of life. It's based on Smart Cities and Smart buildings, along with supporting Smart Infrastructure including Intelligent street lighting, advanced traffic control etc. SCMS tries to find most optimal price/comfort ratio during normal and abnormal city operation. For example it monitors traffic flow, pollution in different parts of the city, power demand from buildings and electricity production in thermal power plant. One of scenarios is as follows: Traffic congestion goes up, it might be a product of accident or simply peak hour. At the same time most of homes are using most of their appliances and thus electric demand is high. This two combined contribute to increased pollution in that part of the city where power plant is located, since there is no wind on that day. SCMS takes appropriate action when pollution sensors detect increased levels of exhaust fumes produced by traffic and power plant. It diverts oncoming traffic to different roads and at the same time starts negotiations for lowering electric power demand. The end result after some time is decreased pollution in critical area, which is result of decreased power plant output (demand for electricity is lowered and/or is redistributed from other sources), excess traffic is rerouted to other roads and previously localised pollution is dispersed to wider area.

2. PLATFORM SELECTION
At the beginning of development for negotiation system we used Agent based platform. Since most of our code was written in Java, we decided to use Java Agent DEvelopment framework [1]. When using Jade we had the opportunity to dynamically add and remove actors in our scenarios and we could see how well our system performed.

Jade ensures most of the basic components for Agent based system and follows FIPA [1] specification. JADE framework has the ability to dynamically add or remove agents, has its own messaging system, so agents can communicate with each other and also includes searching of agents based on their "yellow pages descriptions". That means that every agent has the ability to publicly list its functions for other agent to use. At the same time it has capability to run as a distributed network and thus enables to share workload and even use more agents of the same type to speed up critical tasks. The JADE architecture can be seen on picture below.

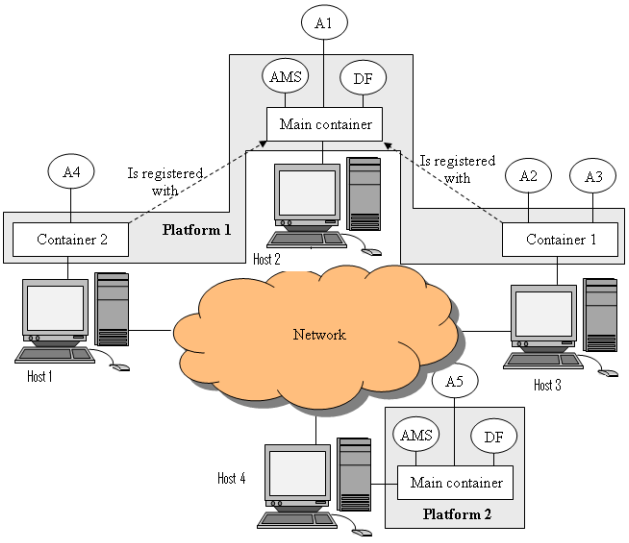


Figure 1: The JADE Architecture [1]

Since all of the agents ran on the same platform and on the same host, our load tests showed us that the algorithm alone was capable of reaching Nash equilibrium in decent time frame. Metrics showed average processing time for around 30.000 agents acting as clients in communication with single negotiator agent to

finish in sub 5 seconds. The tests were made in Windows 7 on laptop with Intel i5 processor and 8 GB of RAM. Similar results were reproduced on another system with Windows 7 and less powerful processor Intel Core2Duo with 4 GB of RAM.

When our work became mature for development inside ACCUS project, some of the changes had to take place. We could not afford to run our system on a separate platform but had to modify it to more practical form of service. That means that we needed more widely spread protocols than ACL messaging which is used in JADE. There was some existing solutions available to choose from and after consideration we decided to use JBoss middleware [1]. The main reason behind this decision was the fact that it is open source and free to use enterprise ready platform with quite big knowledge base and community support. WSO2 [1] was another way to go, but at the end we found common ground with partners within ACCUS project and choose JBoss.

JBoss is middleware that supports exposing services on the web. It also incorporates access to different databases, providing that you use appropriate driver, has its messaging service based upon JMS [1], asynchronous messaging for java and has the ability to dynamically add or remove applications on a server without need to restart it. That allowed us to develop a web based Negotiating Agent which could be accessed from virtually anywhere. Clients could communicate with that negotiator via HTTP [1] requests. Alongside negotiator we also included Power Consumption Monitoring Service (PCMS), which collected data from every client – represented as smart building. Another application that was implemented was Pollution Sensor Subsystem (PSS). It collected data from various subsystems including PCMS and Traffic Simulator Subsystem (TSS) developed by Computer Systems department [1] at "Jožef Stefan" Institute [1]. PSS used simple model to calculate Air Quality Index [1] from power consumption and traffic within the city. When AQI reached critical level in monitored area, power negotiations started and traffic was rerouted. End result was decreased pollution in critical area.

3. APPLICATIONS IN ACCUS

3.1 Power Negotiator

As name suggests, Power Negotiator is used to negotiate electric power demand, usually to decrease current electricity demand used in smart buildings. It leans on stick and carrot approach and tries to reach Nash equilibrium [1]. We decided to use three classes of importance for devices with each representing different priorities. Priority class one is reserved for devices that must run regardless of the cost (fridge, heat pump…), priority class two is considered to include devices that are a bit less essential as class one and can be turned down in order to save money and electric power (washing machine…). Class three includes devices that are not at all important and can be switched off in the time of power shortage to save money (electric car charger). Price factor is highly depended on each smart building. Some of them could need more power than others and have different settings of how much they are prepared to pay for each priority class. Users can also have different ideas which device is more or less important to them. For example every smart building has some devices in priority one class, which means that devices are going to run regardless of the electricity cost. The difference comes in class two and three, where there could be many variations between smart buildings. Some could have set the price factor for class two to 1.8 and factor for class three to 1.2, others for class two to 2.5 and class three to 1.5. Bottom line is that some of the buildings could be more cooperative than others. The negotiations are consisted of three phases:

First phase of the negotiations is introductory. Power negotiator informs all participants that the price for electricity will change along with new prices. Participants report back on how much they are prepared to lower their current power consumption in order not to be charged as high as initial price is set to (stick).

Phase two gathers responses from all participants and calculates reduced fees for electricity. Then it sends those figures to each participant and offers to decrease the price if each participant is willing to lower the consumption a bit more. Participants respond with final figures for their individual power consumption.

Phase three of the negotiations rewards more willing participants with further reduce of the price for electricity (carrot). That factor is calculated for each participant individually and reflects willingness to power reduction.

3.2 Power Consumption Monitoring Service

PCMS is used to monitor current state of Power demand in the city. In our simplified model that metric is also used to calculate AQI.

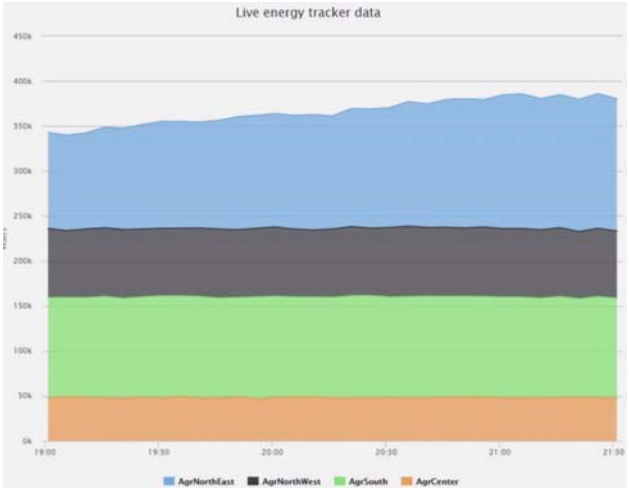


Figure 2: PCMS web interface

Above figure shows increasing power demand over time for Northeast part of the city as tested in one of possible scenarios. The data is collected for each part of city where smart buildings are located and displayed as aggregated sum of corresponding clients.

3.3 Pollution sensor subsystem

PSS is used for monitoring calculated AQI which reflects from current power demand in the city area and traffic congestion in the same part of the city. It also triggers an alarm when that action is needed in order to decrease pollution. Electric energy is included in this model since city uses Thermal power plant for electricity production. Additional energy demand results in higher TPP output electric power production with higher pollution as by-product. The interface was developed and is shown on figure 3.

Figure 3: PSS

Figure above consists of four rows, each representing different part of the city and four columns. First column corresponds to electricity demand, second represents how much congestion is reported on the main roads. Third column includes result of a simplified model for calculated AQI. When AQI reaches critical level, as shown above with value 7 or more, negotiations for electric take place alongside with rerouting of the traffic to other parts of the city.

3.4 Smart building client

Smart building clients consisted of simplified clients representing different households with different appliances connected to electric outlets.

Each house had its own values for different priority classes, explained in Power Negotiation section. For example Building One had price factor for priority class two set to 2.0 of normal price and class three to 1.4 of normal price.

Each electric device had its fluctuating power consumption, state (on or off) and priority class. For example Electric Stove had power consumption between 800W and 2500W and was placed to priority class one. Car charger had power consumption of 3500W and was placed in priority class three.

In the event of power negotiations there is high likelihood that car charger would be taken offline in order to save power and negotiate a better price for electricity. With ability to respond to power negotiation it could get a better price for electric power and contribute to common goal – to reduce the pollution in the area that this building is situated in.

4. RESULTS

The result of a demonstration could be seen in the figures below. After power negotiations smart buildings reduced their power consumption over time. At the same time, traffic was rerouted to other areas and exhaust fumes from traffic were dispersed to wider area.

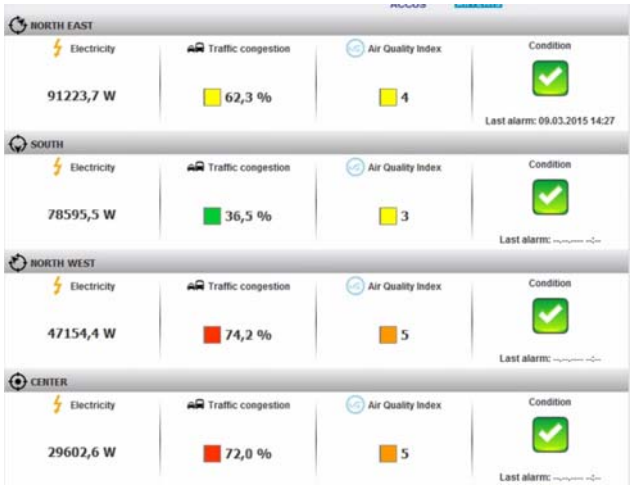


Figure 4: PSS after negotiations and traffic rerouting

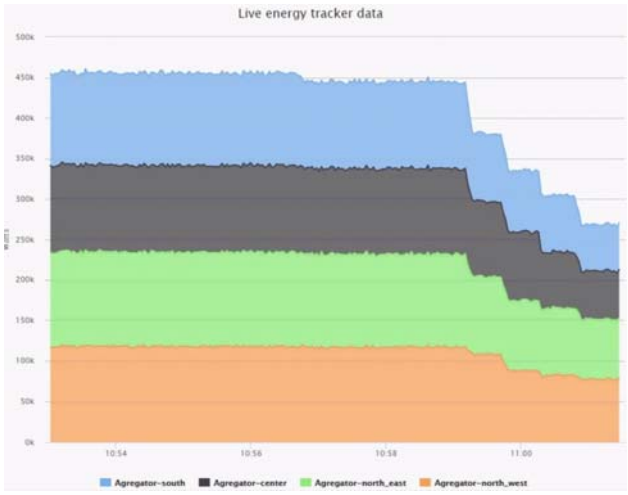


Figure 5: PCMS after power negotiations

The end result is increased quality of air which results in decreased discomfort levels. At the same time drivers benefit from reduced traffic congestion in critical area.

In our system we presume that most of smart buildings that are connected to the network respond to power negotiations. Results are similar even if some of the buildings offer no cooperation or treat every electric device as critical or essential. The key is that majority of smart buildings cooperate to achieve common goal – reducing pollution.

5. ACKNOWLEDGMENTS

Our thanks to ACCUS partners, who contributed to this work.

6. REFERENCES

[1] "Project Accus," 15 9 2015. [Online]. Available: http://projectaccus.eu.

[2] "Jade," [Online]. Available: http://jade.tilab.com. [Accessed 15 9 2015].

[3] fipa, "Fipa," [Online]. Available: http://www.fipa.org. [Accessed 15 9 2015].

[4] JADE, "Jade," [Online]. Available: http://jade.tilab.com/documentation/tutorials-guides/jade-administration-tutorial/architecture-overview/. [Accessed 15 9 2015].

[5] RedHat, "JBoss," Redhat, [Online]. Available: http://www.jboss.org/technology/. [Accessed 15 9 2015].

[6] wso2. [Online]. Available: http://wso2.com. [Accessed 15 9 2015].

[7] wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/JBoss_Messaging. [Accessed 15 9 2015].

[8] wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol#Request_message. [Accessed 15 9 2015].

[9] C. Systems. [Online]. Available: http://cs.ijs.si. [Accessed 15 9 2015].

[10] IJS. [Online]. Available: http://www.ijs.si/ijsw/JSI. [Accessed 15 9 2015].

[11] wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Air_quality_index. [Accessed 15 9 2015].

[12] wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Nash_equilibrium. [Accessed 15 9 2015].

Metis: zaznavanje učnih težav z uporabo strojnega učenja

Damjan Kužnar, Miha Mlakar,
Erik Dovgan, Jernej Zupančič,
Boštjan Kaluža in Matjaž Gams
Institut "Jožef Stefan"
Jamova 39
Ljubljana, Slovenija
+386 1 477 3807
damjan.kuznar@ijs.si

POVZETEK

V prispevku predstavljamo nov sistem Metis za zgodnje zaznavanje učnih težav. Sistem s pomočjo algoritmov umetne inteligence in strojnega učenja na podlagi indikatorjev učnega uspeha identificira učence s povečanim tveganjem za težave v izobraževalnem procesu. Evaluacija sistema je bila izvedena na 692 učencih srednjih šol v obdobju od 2011 do 2015 in kaže zadovoljive rezultate.

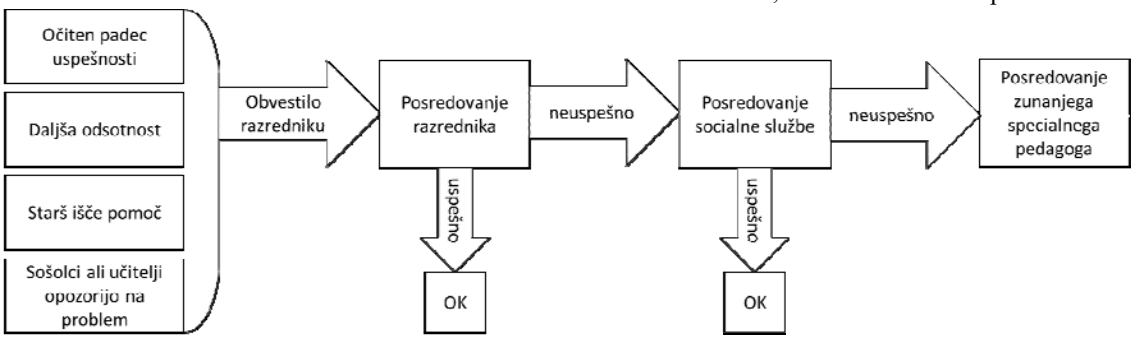
1. UVOD

V slovenskih šolah na srednji predvsem pa na višji in visoki stopnji se srečujemo z razmeroma visokim osipom dijakov in študentov med šolanjem. Razlogov je lahko več: napačno izbrana smer šolanja, učne in osebne težave, pričakovanja staršev, socialni problemi, vedenjske težave in nizko samospoštovanje. V kolikor se te težave odkrijejo dovolj zgodaj, se učencu lahko ponudi pomoč strokovnih služb, ki nato poskušajo pomagati učencu in mu povečati možnosti za uspeh.

Obstoječi proces detekcije učnih težav je prikazan na sliki 1. Učitelji običajno opazijo posledice teh težav šele, ko učenec začne pridobivati negativne ocene (očiten padec uspešnosti). Takrat govorimo o absolutni neuspešnosti. Od učenca se namreč pričakuje, da v primeru takih težav pravočasno sam poišče pomoč pri pedagoškem delavcu, kar pa se zgodi le v redkih primerih (to redko doživi 67 % strokovnih delavcev) ali celo nikoli (7 % strokovnih delavcev) [1]. Podobno redko je samoiniciativno iskanje pomoči staršev, ki opazijo težave pri svojih otrocih. V primeru absolutne neuspešnosti so učenci deležni pomoči s strani pedagoškega delavca, ki pa običajno zajema le pomoč pri razumevanju snovi. Takemu učencu učitelj običajno posveti več pozornosti med poučevanjem in ga usmeri na dopolnilni pouk, v nižjih razredih pa mu lahko pomaga tudi učitelj v podaljšanem bivanju. Če se situacija kljub prizadevanju učitelja ter učenca ne popravi, se učenca napoti k strokovnemu delavcu, ki je prisoten na vsaki šoli, šolskemu psihologu in po potrebi tudi k specialnemu pedagogu.

Izkušen pedagog, ki posveti dovolj časa svojim učencem za spremljanje njihovega razvoja ter uspešnosti preko celega učnega obdobja (skozi šolsko leto ali celo stopnjo izobraževanja) in se pogovarja z njimi in njihovimi starši, je sicer zmožen zgodaj prepoznati težave. Toda zaradi preobremenjenosti ter naraščajočih velikosti razredov in letnikov (največ 28 učencev v osnovnih ter 32 dijakov v srednjih šolah, v praksi celo več, ter tudi po 50 študentov na višjih in visokih šolah) pedagoški delavci ne morejo posvetiti dovolj časa vsem svojim učencem in pogosto ne namenijo dovolj časa prepoznavanju učencev, ki bi potrebovali pomoč, da bi se uspešno izognili ponavljanju letnika ali predčasnemu prenehanju šolanja.

Informatizacija vzgojno-izobraževalnega sistema je že nekaj časa aktualna tema. Prenos podatkov o šolanju s tradicionalnih na digitalne medije ima očitne pozitivne posledice: lažje hranjenje, pregledovanje in večjo dostopnost do podatkov. Številne šole tako že imajo svoj informacijski sistem ali pa sistem v oblaku, ki podpira številne funkcije, ki omogočajo lažje vodenje evidence prisotnosti in ocen pa tudi učne snovi ter administrativnih vidikov šolstva, kot so šolska prehrana ter šolsko



Slika 1. Obstoječi proces zaznavanja učnih težav.

računovodstvo. Kljub temu, da je dostop do teh podatkov mogoč, pa se redko oz. nikoli ne uporablja za avtomatizirano odkrivanje zakonitosti v teh podatkih. Najbolj enostavna metoda odkrivanja zakonitosti v podatkih je vizualizacija podatkov, kar že podpira večina informacijskih šolskih sistemov (eAsistent, Moodle). Slabost uporabe vizualizacije je, da pedagoškemu delavcu lahko vzame veliko časa in ne poda koristnih informacij, poleg tega pa je lahko zaradi velike množice različnih tipov podatkov prikaz nejasen in nerazumljiv.

2. SORODNO DELO

V preteklosti je že bilo nekaj poskusov uporabe metod strojnega učenja pri izobraževalnih procesih tudi v slovenskem raziskovalnem prostoru [2], [3], [4]. Raziskovalci so reševali problem napovedovanja učnega uspeha oz. neuspeha v srednjih šolah. Pri tem so želeli razviti orodje, ki bi učiteljem, razrednikom in svetovalnim službam omogočilo lažje svetovanje pri izbiri nadaljnjega šolanja učenca. Raziskovalci v vseh treh člankih ugotavljajo, da je s primernimi inteligentnimi metodami mogoče zgraditi kakovosten model napovedovanja uspešnosti učencev. Vsi so tudi izpostavili težave, s katerimi so se srečevali pri gradnji modela: pomanjkanje podatkov, ročno vnašanje podatkov v sistem, slaba prilagodljivost modela.

3. SISTEM METIS

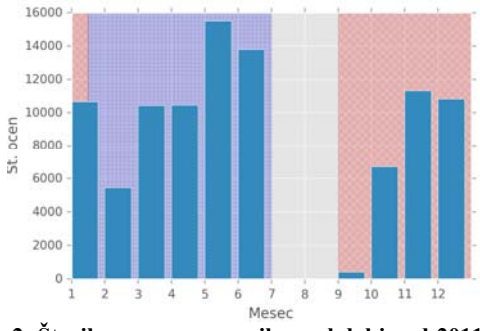
Sistem Metis ima v grobem dve nalogi: (1) zaznavanje težav in (2) pomoč pri odpravljanju težav. V sklopu prve naloge se z metodami statistične analize in algoritmi za strojno učenje [5] na zgodovinskih podatkih zgradi model za napovedovanje učnih težav. Z analizo indikatorjev učnega uspeha, ki so dostopni v informacijskih sistemih šole (ocene, izostanki, izogibanje preverjanjem znanja, ipd.), in uporabo zgrajenega modela nato identificira učence z učnimi težavami oz. učence, za katere obstaja določena stopnja tveganje za neuspeh. V primeru zaznane težave sistem obvesti učitelja razrednika, ki nato povabi učenca na individualne konzultacije, kjer učitelj skupaj z učencem definira cilje, ki jih bo učenec poskušal doseči in za katere se smatra, da so ključni pri odpravi oz. preprečevanju učnih težav.

V nadaljevanju prispevka podrobno poročamo o rezultatih prve naloge, t.j. zaznavanje oziroma napovedovanje učnih težav.

4. NAPOVEDOVANJE UČNIH TEŽAV

V sodelovanju s pedagoškimi delavci smo učne težave definirali na kot negativno oceno pri katerem koli predmetu ob zaključku tekočega ocenjevalnega obdobja.

Napovedovanje učnih težav se izvaja za različna časovna obdobja pred zaključkom ocenjevalnega obdobja z željo po čim bolj zgodnem zaznavanju morebitnih učnih težav, saj imata tako učenec in učitelj dovolj časa za popravljalne oz. preprečevalne ukrepe. V ta namen smo v sodelovanju s pedagoškimi delavci definirali naslednja časovna obdobja. Za prvo ocenjevalno obdobje je potrebno imeti oceno težav do konca novembra, ko zberejo že prve ocene in se lahko odkrije že prve težave. Za drugo ocenjevalno obdobje želijo pedagoški delavci prve napovedi že konec februarja, kar je mesec in pol po začetku drugega obdobja, kar se je v fazi testiranja izkazalo kot težka naloga zaradi premajhnega števila zbranih ocen. Dinamika zbiranja ocen je razvidna na Slika 2, kjer je prikazano število ocen po mesecih. Vidimo lahko, da je do konca februarja zbran le majhen delež ocen za drugo ocenjevalno obdobje, zato smo dodali še dodatne napovedi ob koncu marca, aprila in maja.



Slika 2. Število ocen po mesecih za obdobje od 2011 do 2015. Rdeče ozadje (karo vzorec) označuje prvo ocenjevalno obdobje, modro ozadje (kvadratast vzorec) označuje drugo ocenjevalno obdobje.

Za časovna obdobja smo definirali tudi različne mejne vrednosti zahtevane natančnosti (angl. precision) modela [6], t.j. delež učencev, ki bodo resnično imeli učne težave, od vseh učencev, za katere smo napovedali, da bodo imeli težave. Obdobja in zahtevane minimalne natančnosti so povzete v Razpredelnica 1.

Razpredelnica 1. Obdobja napovedovanja in minimalne zahtevane natančnosti.

Obdobje	*1. obdobje*	*2. obdobje*			
Datum	30.11.	28.2.	31.3.	30.4.	31.5.
Natančnost	30%	40%	50%	60%	60%

Zaznavanje učnih težav se izvaja s pomočjo metod strojnega učenja, kjer se v prvi fazi izvede učenje napovednega modela na zgodovinskih podatkih ter napovedovanje prihodnjih učnih težav v drugi fazi na podlagi prej naučenega napovednega modela. Večina metod strojnega učenja za učenje zahteva podatke v atributni obliki, kjer se vsak primer opiše z množico atributov (lastnosti), zato je potrebna predhodna obdelava

podatkov, ki se jih pridobi iz informacijskih sistemov šol, v primerno obliko.

4.1 Podatki

Učno množico za učenje smo zgradili iz podatkov o 692 učencih srednjih šol iz obdobja od 2011 do 2015. V pričujočem delu smo nadalje osredotočili na štiri z ocenami najbolj zastopane predmete (matematika, angleščina, slovenščina in kemija) zaradi zagotavljanja dovolj velikega vzorca. Surovi podatki so bili pretvorjeni v atributno obliko z naslednjimi atributi:

- *ID dijaka*: anonimiziran identifikator učenca (meta atribut, ki ni udeležen pri samem učenju)
- *Predmet*: predmet, na katerega se nanaša učni primer
- Za trenutno in prejšnje ocenjevalno obdobje smo izračunali naslednje atribute:
    - *Povprečje*: povprečna ocena za trenutno/prejšnje obdobje
    - *Število*: število ocen v obdobju
    - *STD*: standardna deviacija ocen
    - *Min*: najnižja ocena
    - *Max*: najvišja ocena
    - *Naklon*: izraža ali učencu ocene naraščajo ali padajo
    - *NPS*: število izostankov od preverjanja znanja
    - *Vsi izostanki*: vsi izostanki v obdobju
    - *Opravičeni*: opravičeni izostanki
    - *Neopravičeni*: neopravičeni izostanki
- *Razlika v oceni*: razlika v povprečni oceni med prejšnjim in trenutnim obdobjem
- *Razlika v št. ocen*: razlika v številu ocen med prejšnjim in trenutnim obdobjem
- *Negativno*: razredna spremenljivka oz. vrednost, ki jo želimo napovedovati in označuje ali je imel učenec pri predmetu negativno zaključeno oceno

Velikost učne množice se razlikuje glede na obdobja, ki so definirana v Razpredelnica 1, ker nekateri učenci prve ocene dobijo šele kasneje v ocenjevalnem obdobju. Prav tako se razlikuje porazdelitev učnih primerov v razred, t.j. ali je bil predmet *negativno* ali *pozitivno* zaključen ob koncu obdobja. Povzetek velikosti učnih množic in zastopanosti razredov je v Razpredelnica 2.

Razpredelnica 2. Velikosti učnih množic in zastopanosti razredov za različna časovna obdobja napovedovanja.

Obdobje	*Velikost učne množice*	*Pozitivno zaključeno*	*Negativno zaključeno*
30.11.	3225	3104 (96.25 %)	121 (3.75 %)
28.2.	2367	2317 (97.89 %)	50 (2.11 %)
31.3.	4949	4857 (98.14 %)	92 (1.86 %)
30.4.	5649	5549 (98.23 %)	100 (1.77 %)
31.5.	5734	5627 (98.13 %)	107 (1.87 %)

4.2 Metode strojnega učenja

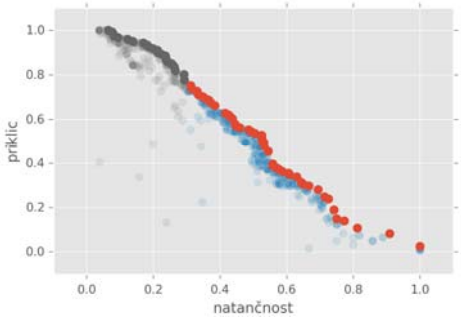
Za učenje smo uporabili Scikit-learn (SKLearn) knjižnico za Python, ki vsebuje kopico metod strojnega učenja. SKLearn smo povezali s knjižnico Hyperopt, ki služi za optimizacijo hiper parametrov, t.j. optimizacijo parametrov klasifikatorja. Ker Hyperopt podpira le enokriterijsko optimizacijo, smo kot kriterij uporabili mero F1, ki je geometrijsko povprečje mer natančnosti in priklica.

Optimizacija s Hyperopt je zajemala iskanje najboljšega metode strojenega učenja in najboljših parametrov zanjo. Množica metod je zajemala naključne gozdove (angl. random forest), odločitvena drevesa (angl. decision tree), naivni Bayes (angl. naive Bayes) in logistično regresijo (angl. logistic regression). Prvotno smo vključili tudi druge metode, kot so metode podpornih vektorjev (SVM), vendar so imele težave z velikostjo učne množice oz. je učenje s temi metodami bilo časovno predolgo.

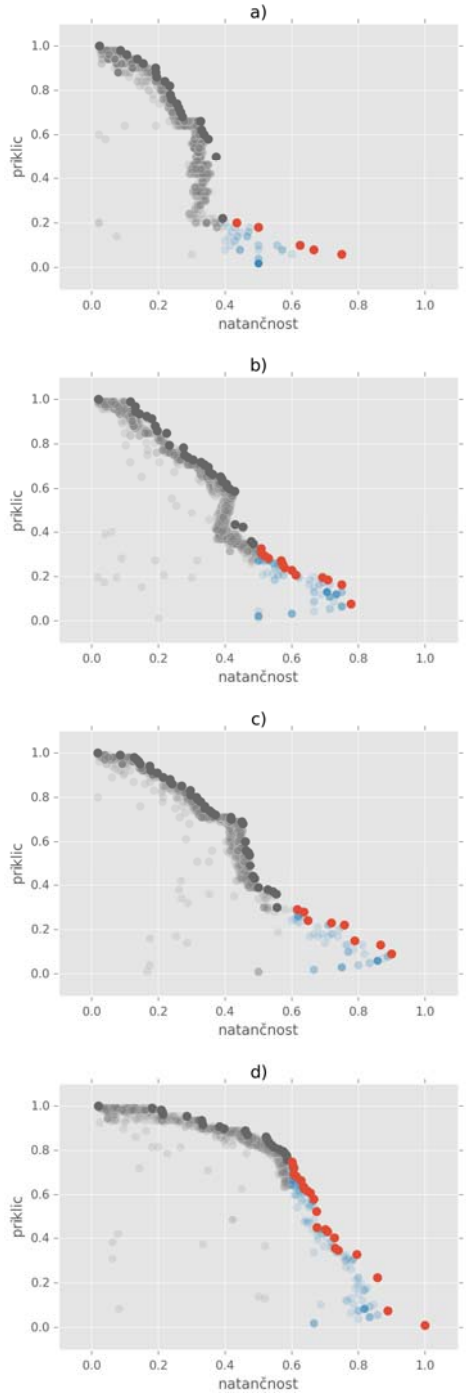
5. EVALUACIJA

Pri iskanju optimalnih klasifikatorjev smo pri evaluaciji posameznega klasifikatorja uporabili 5 kratno prečno preverjanje ter izračunali natančnost, priklic in mero F1. V nadaljevanju poročamo o klasifikatorjih, ki so bili v procesu optimizacije s Hyperopt najdeni.

Slika 3 in Slika 4 prikazujeta uspešnost klasifikatorjev najdenih za napovedovanje za različna obdobja napovedovanja. Siva barva prikazuje klasifikatorje, ki ne ustrezajo kriteriju minimalne zahtevane natančnosti, modra barva prikazuje ustrezne dominirane klasifikatorje in rdeča barva prikazuje ustrezne nedominirane klasifikatorje.



Slika 3: Rezultati za napovedovanje ob koncu novembra v 1. ocenjevalnem obdobju.



Slika 4: Rezultati za napovedovanje v 2. ocenjevalnem obdobju ob koncu: a) februarja, b) marca, c) aprila in d) maja.

Vidimo lahko, da optimizacija sicer vedno najde klasifikator, ki ustreza minimalni zahtevani natančnosti, vendar pri se med obdobji napovedovanja v 2. ocenjevalnem obdobju močno razlikuje priklic. Zelo dober priklic dobimo, če napovedujemo konec maja, ko je priklic blizu 0.8, kar pomeni, da zajamemo blizu 80 % od vseh učencev, ki bodo imeli težave.

6. ZAKLJUČEK

Predstavili smo sistem, ki na podlagi različnih učnih indikatorjev napove ali bo imel določeni dijak učne težave v obliki negativno zaključene ocene pri katerem koli predmetu. Glede na različna obdobja napovedovanja smo poročali o različni natančnosti in priklicu, saj je sistem razvit tako, da omogoča izbiro najbolj ustreznega klasifikatorja z dajanjem različne teže priklicu in natančnosti.

V prihodnje želimo uspešnost napovedovanja še povečati z uporabo bolj kompleksnih atributov, kot so npr. n-grami, in vključitvijo večjega nabora metod strojnega učenja v proces optimizacije z metodo Hyperopt.

7. ZAHVALA

Zahvaljujemo se Zlatki Bukovec-Gačnik, Mojci Lukšič, Darji Progar in Maji Grošičar za prispevke k definiranju funkcionalnosti sistema Metis.

Projekt Metis je financiran s strani Ministrstva za izobraževanje, znanost in šport RS in Evropskega sklada za regionalni razvoj.

8. REFERENCE

[1] L. Magajna, S. Pečjak, C. Peklaj, G. Č. Vogrinčič, K. B. Golobič, M. Kavkler in S. Tancig, Učne težave v osnovni šoli : problemi, perspektive, priporočila, Ljubljana: Zavod RS za šolstvo, 2008, pp. 245-250.

[2] D. Rudolf, „Napovedovanje učnega neuspeha,“ v *Informacijska družba*, Ljubljana, 2006.

[3] T. Viher, „Večparametrski model za predvidevanje uspešnosti zaključka šolanja po končanem prvem letniku srednje šole,“ v *Zborniku Informacijska družba IS2004*, Ljubljana, 2004.

[4] S. Gasar, M. Bohanec in V. Rajkovic, „Napovedovanje uspešnosti zakljucka šolanja,“ *Organizacija,* Izv. 8, št. 35, pp. 508-513, 2002.

[5] I. H. Witten, E. Frank in M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.

[6] D. M. Powers, „Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation,“ *Journal of Machine Learning Technologies,* Izv. 2, št. 1, pp. 37-63, 2011.

[7] S. Arlot, „A survey of cross-validation procedures,“ *Statistics Surveys,* Izv. 4, pp. 40-79, 2010.

Evaluation of algorithms for speaker diarization in sound recordings

Vanja Mileski, Eng.
Faculty of Computer and Information Science,
Ljubljana, Slovenia
vanja0mileski@gmail.com

Matija Marolt, PhD
Faculty of Computer and Information Science,
Ljubljana, Slovenia
matija.marolt@fri.uni-lj.si

ABSTRACT
Speaker diarization is a process that separates the audio clip in sections regarding the identity of the speakers. Speaker diarization in sound recordings answers the question of *who spoke when?*

This paper is dedicated to the automatic segmentation of speakers in a variety of sound recordings. A test data of audio recordings in Slovenian was prepared, which was obtained from field recordings. The recordings contained two or more speakers and very often they contained other sounds, silence, overlap between the speakers and other features the algorithms may struggle with. We manually annotated a set of recordings by placing segment boundaries and assigning speakers to each segment. The set represented our ground-truth for algorithm evaluation. We ran all the algorithms for diarization that we evaluate on this test data. We wrote a program which takes the results from the algorithms as an input which have different formats and different types of representation, and the results were converted to a common format. We evaluated the accuracy of the algorithms and analysed how well they work in different situations.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing: indexing methods

General Terms
Algorithms, Measurement, Performance, Theory

Keywords
speaker segmentation, diarization, algorithm evaluation

1. INTRODUCTION
In our everyday life, we come across multiple technologies that use sound as a main source of information exchange, from films, television shows, news and radio to voice mail, meeting recordings etc. Detection, identification and diarization of speech have become a very interesting subject of research. In this paper, we will look into diarization of speakers only. Diarization is the process of automatic division of a sound recording into speaker segments and determining which segments are from the same speaker. It can be used to improve the readability of automatic transcription of speech [1, 2].

A diagram illustrating the process of speaker diarizaition is shown in Figure 1. The first step of diarization is feature extraction, which provides features describing the audio recording. Using the features, a disambiguation between speech and non-speech segments is made. The speech segments are processed with segmentation and clustering of speakers. The purpose of the former is the search for points in the sound recording where speaker change occurs. It divides the recording in acoustic homogeneous segments, with each segment consisting of (in the best case) a single speaker. The purpose of the latter is unsupervised classification of these speaker segments and their grouping based on the speaker's features. This means that it recognises all

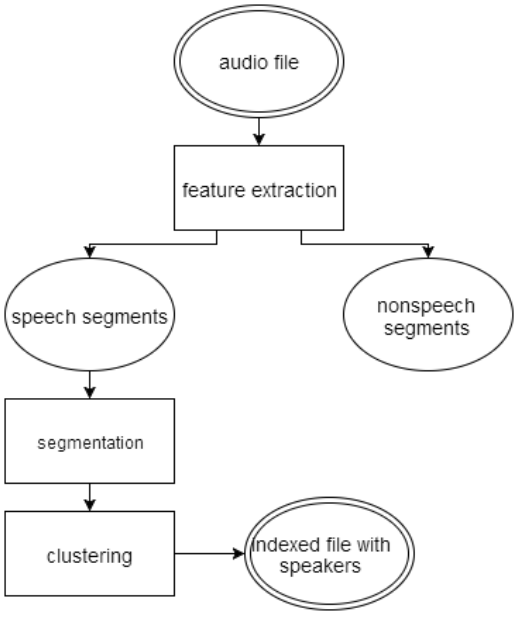


Figure 1: Typical flow of the diarization steps

speaker segments spoken from a certain speaker, and it clusters and annotates them together. This can be performed in a bottom-up or top-down manner [1]. Even though these two processes (segmentation and clustering) can be jointly optimized or done in the other direction, usually segmentation is the first step followed by the clustering of speakers.

Diarization, if we look at the speed technology changes nowadays, has been around for some time. There already exist many approaches and algorithms performing speech diarization. The National Institute of Standards and Technology (NIST, USA) has a project for *Rich Transcription Evaluation* [1, 2, 3], but even though that the technology is developing for a while now, the results are still not as good as we want it to be. The biggest problem facing almost all of the algorithms is when the sound recordings are of a poor quality, contain noise, speaker overlap and difference between vocal powers of the speakers. In this paper, we use such recordings, and furthermore, the recordings are in Slovenian language.

2. BACKGROUND

In this section, we describe the three basic parts of speaker diarization: feature extraction, segmentation and clustering.

Feature extraction. The first step in acquiring features from an audio recording is the transformation of the analogue signal into a digital one. The step where we take only certain information from the audio, making it less redundant, is called feature extraction. Most commonly used functions for feature extraction are: *Linear Prediction Coefficients (LPC), LPC Derived Cepstral Coefficients, Line Spectral Pairs (LSP), Mel-Frequency Cepstral Coefficient (MFCC), short-time energy, zero-crossing rate*, and *pitch* [1, 2].

Segmentation. The algorithms for speaker segmentation are divided into three categories: model-based, metric-based and hybrid algorithms.

In *model-based* algorithms, a set of models is derived and trained for different speech categories from the speech corpus and then the input speech is sorted using these models. This means that prior knowledge is needed for initialising these models of speakers. Most used models at this stage of the diarization are: *Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Viterbi algorithm* and other [1, 2].

Metric-based algorithms for segmentation estimate the similarity between two neighbouring analytical windows in the sound recording with the use of a distance function. Local maximums of that distance function, which surpass a certain threshold are considered speaker change points. These methods do not need prior knowledge about the number of speakers in the sound recording, their identities or similar. Most used distance functions are the *Kullback-Leibler Divergence, Gaussian Divergence, Generalized Likelihood Ratio* and *Entropy Loss* [1]. Most commonly used criterion is the *Bayes Information Criterion (BIC)* [1, 2, 4].

Hybrid algorithms combine model-based and metric-based techniques. Usually metric-based segmentation is first used to segment the input audio signal. These segments are then used to create model speakers and then model-based resegmentation refines the previous segmentation.

Clustering. The approaches for speaker clustering are separated into two main categories: deterministic and probabilistic [1]. Deterministic approaches cluster similar sound segments based on measurements; popular methods are *Self-Organising Maps* (SOM methods) and hierarchical methods. Probabilistic approaches on the other hand, use GMM or HMM to model the clusters.

3. DIARIZATION TOOLS

We list and discuss the most widely used tools for diarization of speech recordings that were used in this study.

LIUM speaker diarization [5] is a program written in Java. It consists of a wide range of tools for a complete diarization of speakers. It is optimized for radio and television shows but can be optimised for other uses too. It uses the BIC algorithm, the cluster is modelled with a single state HMM represented with 8-component GMM and even has gender detection.

PyAudioAnalysis[1] is an open Python library which offers feature extraction, classification, segmentation and visualisation of sound recordings. The segmentation process can be supervised or unsupervised. For supervised segmentation a supervised model is used for classification and segmentation (with a predictive model or HMM). For unsupervised segmentation, the model is not present and we cluster the detected clusters with the k-means algorithm.

The Diarize-jruby library provides a tool for speaker segmentation and identification from other sounds. It is wrapped around the LIUM library. Additional features are: normalisation of speaker models with the use of M-Norm, symmetric Kullback-Leibler divergent approximation and support for speaker supervectors, which can be used for fast identification of speakers.

VoiceID is a recognition and identification system based on LIUM. VoiceID can handle video or sound files, determine at which intervals a speech is present and then analyse those segments and determine who is speaking. To do that it uses a data collection for speech models.

ALIZE is an open source platform for biometric authentication [6, 7]. It is written in C++ and the architecture of the tool is based on the distribution of functions between several servers. The main servers are the feature server which deals with the acoustic data, the mixed server that deals with the models (storing, component binding, reading, writing) and the statistical server that conducts all of the statistical calculations.

Matlab Audio Analysis Library is a Matlab library that covers a wide spectrum of sound analytical tasks like general audio handling (input, output, conversion, recording...), sound processing, feature extraction, classification, segmentation and information search in music [8]. Particular algorithms used in the library are: short-time audio processing, short-time FFT, chroma vector, fundamental frequency,

[1]Available at https://github.com/tyiannak/pyAudioAnalysis

HMM, dynamic programming, short-term energy, zero crossing rate, entropy of energy and many more.

4. EVALUATION

We developed and implemented a framework (in Java) for the evaluation of the algorithms/tools mentioned above. The recordings used for evaluation are ethnomusicological field recordings taken from the EthnoMuse archive of the Institute of Ethnomusicology of the Research Centre of the Slovenian Academy of Sciences and Arts. The recordings differ between each other in the environment they were recorded, number of speakers, speech ratio between the speakers, quality of the recordings and the amount of overlap between speakers. The total length of the audio recordings on which the algorithms were tested was 30 minutes. First, for each of the recordings a manual diarization was made: the number of speakers was written and the time intervals for each speakers in which he speaks. Because of human limitations this was rounded to the nearest $25ms$. This will be our ground truth. Because all tools have different output formats, a common format was created to which all the outputs were converted. Data such as gender, type of environment and similar additional data that some programs provide were not taken into consideration.

Programs for diarization of speakers have to make segmentation and clustering of speakers. Most commonly used metric for measuring the results of the programs is Diarization Error Rate (DER) [1]. DER is the ratio between cumulative time of incorrect annotation and the total time of all speech segments in the sound recording. It is represented in percentiles, with 0% being the perfect score, higher percentiles meaning less satisfactory results. DER is defined as $DER = MS + FA + SE$, where MS is *Missed Speech*, part of the recording where there is a speaker, but the algorithm does not detect it, FA is for *False Alarm*, when there is no speech but the algorithm annotated as if there were, and SE for *Speaker Error*, where the algorithm correctly detected the speech but annotated it as the wrong speaker. In this paper, we will use inverse DER (1-DER), hence the value 100% to be the best result possible – the algorithm detected all of the speech segments in the recording and correctly identified the speaker.

The evaluation framework separated the length of the sound recording into intervals of $25ms$ same as from both the tools and the ground truth. The results where then compared and DER was calculated. If the ground truth annotation states that multiple people talk at the same time for a certain interval and it is not possible to determine a single speaker, such intervals were skipped.

5. RESULTS

Figure 2 shows the results of the evaluation for each audio recording separately for every program, whereas Figure 3 shows the average result for all recordings for each program. Exact description for each sound recording cannot be precisely given in a written or visual form. However, since all programs are primarily made for different purposes, they all work differently on the different sound recordings, as can be seen from the results. A brief analysis for each algorithm is given below.

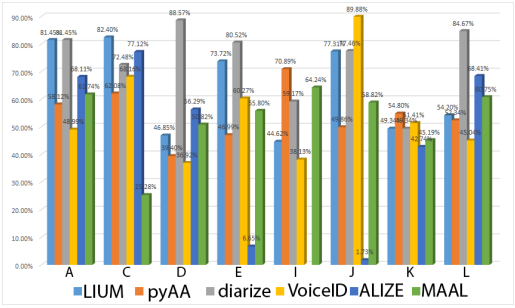


Figure 2: Comparison of the tools for individual recordings, higher is better

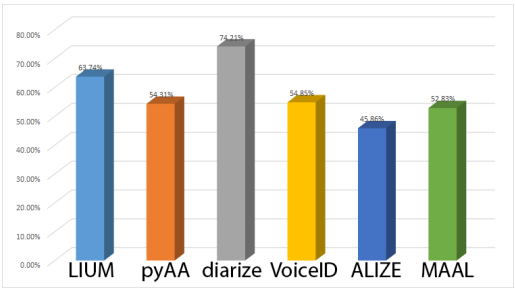


Figure 3: Comparison of the tools based on the average score, higher is better

LIUM speaker diarization provided good clustering of the segments and on half of the audio recordings has given a correct value for the number of speakers. Bigger problems have appeared at the recordings I, K and L, or recordings containing overlaps between speakers. The segmentation part was very good for those recordings where LIUM estimated the correct number of speakers. The only exception is D which has very frequent overlaps. As expected, it gave a very bad result for the recording I where he estimated that there are 10 speakers instead of 3.

PyAudioAnalysis had an excellent estimation of the number of speakers in all recordings and gave an exact number for almost all of them. It also had a very good estimation on the recording I, where most algorithms struggle and because of this it has the best score out of all programs. But, it achieved lower scores on recordings like D, E and J where there is a dominant speaker who speaks most of the time.

Diarize-jruby has given a correct answer about the number of speakers on only few of the recordings. However, this doesn't mean that it made big errors, on the contrary, it gave very good results. The only problems it had was with I and K, recordings that have overlaps and fast changing of speakers.

VoiceID achieved similar results as Diarize-jruby: it guessed the number of speakers on only few of the recordings, but gave relatively good results overall. On some of the recordings it behaved really good, but on some like I where he estimated wrong about the number of speakers by a big margin

and D where there are big overlaps between the speakers, he achieved lower results.

ALIZE was the only surprise out of all the algorithms. It's estimation about the number of speakers was extremely low, with 7 out of 8 recordings estimated at only one speaker in the entire recording. However, for recordings such as A, C, D and L it gave relatively satisfactory results. But for recordings like E, I and J it gave extremely low results and couldn't even detect that there is speech in the recording.

Matlab Audio Analysis Library estimated the number of speakers in a recording relatively low even on recordings that are clear and without overlaps. However, it gave relatively good results overall. Almost on all of the recordings it has achieved a relatively good result compared to the other algorithms, but it got a very low result on the recording C because of the frequency of speaker changes and the weak voice of one of the speakers, which probably was the biggest problem.

Table 1 presents the cumulative advantages of the evaluated algorithms. From these, we can note that each of the algorithms has its advantages when applied to a specific type of recordings (e.g., VoiceID achieves very good results on movie/TV recordings).

Table 1: Summarized advantages of the algorithms.

LIUM	good when the size of the clusters is not proportionate and a speaker's voice in the recording is weak.
Py Audio Analysis	best when first one speaker talks a longer period, then the other speaker for a longer period and so on.
Diarize-jruby	good when the size of the cluster is not proportionate, a speaker's voice is weak, certain speakers have noticeably more powerful voices and there is a domination from one speaker.
VoiceID	good when the size of the clusters is not proportionate, a speaker's voice is weak and it is a movie or a television show.
ALIZE	satisfactory when a speaker's voice in the recording is weak and it is used for speech detection and verification.
Matlab Audio Analysis Library	good when the recording isn't dynamic and there are no frequent changes between the speakers.

6. CONCLUSIONS

Diarization methods are an attractive field in the last couple of years. Because of the fast expansion of sound recordings available, speaker segmentation is a very active research field [3]. In this paper, some of the most popular programs for diarization were evaluated.

State-of-the-art algorithms for diarization that contain all of the diarization steps (feature extraction, speaker segmentation, speaker clustering) are working satisfactory in "clean" audio speech recordings, but there is still a big room for improvement, especially when the recordings are taken in noisy environments, meeting recordings, when the ratio of spoken time between different speakers is big, or when certain speakers are quieter that the others or some have more dominant voices.

Even though many implement very similar algorithms, they all had different approaches and goals when were designed. All of the algorithms have strengths and weaknesses, and it is not possible to find a algorithm that will work good on all the different types of recordings. The goal of the paper is first to evaluate the different algorithms on various recordings and look for the algorithm that is overall best, i.e., the algorithm that could be used as a default when dealing with unknown recordings. Based on the results, Diarize-jruby achieved overall the best results.

7. REFERENCES

[1] Margarita Kotti, Vassiliki Moschou, and Constantine Kotropoulos. Speaker segmentation and clustering. *Signal processing*, 88(5):1091–1124, 2008.

[2] Marijn Anthonius Henricus Huijbregts. Segmentation, diarization and speech transcription: surprise data unraveled. 2008.

[3] Jonathan G Fiscus, Jerome Ajot, and John S Garofolo. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans*, pages 373–389. Springer, 2008.

[4] Xuan Zhu, Claude Barras, Sylvain Meignier, and Jean-Luc Gauvain. Combining speaker identification and bic for speaker diarization. In *INTERSPEECH*, volume 5, pages 2441–2444, 2005.

[5] Sylvain Meignier and Teva Merlin. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010, 2010.

[6] Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier. Alize, a free toolkit for speaker recognition. In *ICASSP (1)*, pages 737–740, 2005.

[7] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas WD Evans, Benoit GB Fauve, and John SD Mason. Alize/spkdet: a state-of-the-art open source software for speaker recognition. In *Odyssey*, page 20, 2008.

[8] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.

[9] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. Technical report, Idiap, 2013.

[10] Sue E Tranter, Douglas Reynolds, et al. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

[11] Douglas A Reynolds, Patrick Kenny, and Fabio Castaldo. A study of new approaches to speaker diarization. In *Interspeech*, pages 1047–1050, 2009.

[12] Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems for signal, image and video technology*, 20(1-2):61–79, 1998.

Analyzing and Predicting Peak Performance Age of Professional Tennis Players

Miha Mlakar
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
miha.mlakar@ijs.si

Tea Tušar
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
tea.tusar@ijs.si

ABSTRACT

Tennis is a sport where physical ability, match experience and mental toughness all have to be at the peak to achieve best results. To determine when this typically happens, we analyzed aging trends and calculated the average aging curve for professional players from 1974 to 2014. We showed that 25 is the age when it is most likely for a player to reach his career peak. In addition to analyzing performance, we also developed an algorithm for predicting how a player will perform in the next years. The algorithm finds players that had similar career paths and uses these players to predict the performance for the next years.

1. INTRODUCTION

Tennis is a very popular sport played all over the world by thousands of players of different ages. In recent years we are witnessing older players like Federer (33), Ferrer (33), Lopez (33) and Karlovič (36) still winning ATP tournaments and playing important roles on the ATP World Tour. This contradicts what we think and poses an interesting question: what is the age at which a tennis player is most likely to be at the peak of his career?

Of course the answer is not the same for all players. If we draw a line presenting the performance of a player depending on his age, we get his aging curve [5]. Players have different aging curves, due to various reasons such as different constitutions, styles of play, mental preparedness and other factors. Also some players tend to develop faster physically and some players get injured more frequently when they get older. All these facts influence the player careers and consecutively their aging curves.

Measuring players' peak performances and drawing their aging curves is an integral part of player analysis in many sports. These analyses first began in baseball, where the database of all statistics is very detailed and covers players from major and minor leagues [4]. By trying to determine the player aging curves the clubs and scouts are trying to find out which players are worth buying. We draw our inspiration for this paper from one of the most well-known systems for determining aging curves and also predicting career peaks in baseball called PECOTA [4].

In tennis, there were some analyses measuring how age influences performance in tennis, but to our knowledge none of them did it thoroughly on a great amount of data. In [3] authors identified the age of peak performance in a broad range of sports including tennis and associated the performance peak with how much explosiveness and how much stamina is needed in this sport. In [1], the authors showed how tennis players in Grand Slam tournaments are now older then they used to be in nineties. In addition, they analyzed peak performances for players of Wimbledon in 2014.

In this paper we analyzed aging trends and presented average aging curves for professional tennis players. In addition, we are interested in predicting how well a specific player will perform in the next years. The predictions are based on determining the similarities between the players that already had similar career paths and have similar characteristics.

Predicting a player's performance or his aging curve can be very useful when trying to estimate, for example, if and when a young prospect player will break the top 10 or how long can a player over 30 keep his current level.

The structure of this paper is as follows. In Chapter 2 we describe the preprocessing of data, including normalization and determining which players should be included in the analysis and which shouldn't. Chapter 3 presents the average aging curve for professional tennis players. We also detected and presented other facts about player performances in dependance of the age or time playing. Chapter 4 describes the algorithm for predicting the performance and presents the obtained results of predictions. The final chapter concludes the findings and names the possibilities for future work.

2. DATA PREPARATION

The data used in this paper includes rankings and player characteristics for men ranked on the ATP rankings from the beginning of 1974 to the end of year 2014. We included all professional players, not just the best, because we wanted to obtain a general tennis aging curve and also we wanted to be able to predict the aging curve for players of different levels.

2.1 Data filtering

Since we were analyzing players aging trends, we had to be careful not to analyze players with missing data. So, we removed all players that (i) were in the middle of their careers in 1974, because maybe the career peak has already passed and (ii) players that still had active careers at the end of 2014, because their peak may still be coming. Including these players with inaccurate career performances would impair the results.

The further inspection of data showed that some players have rankings just for a few years. In order to find the appropriate aging curve, we had to determine, if we should limit the career lengths, or use all players.

Let say we have a player that is ranked only for one year. The analysis would show that this year is his performance peak and the best year for playing tennis. Because his aging curve would have been different if he would play more years, his example would influence the shape of the aging curve in an incorrect way.

To see how limiting the career lengths would influence the distribution of players, we draw two graphs representing player rankings depending on their ages. On the first one (Figure 1) we draw lines for player with at least two years on the rankings. The second one (Figure 2) shows players with at least ten years long careers.

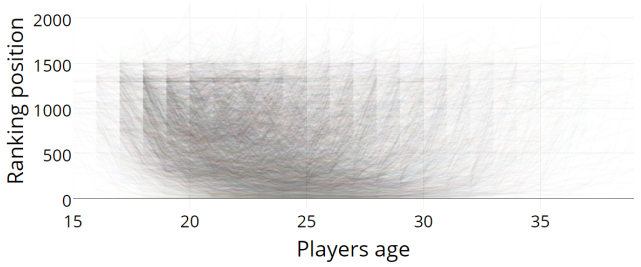


Figure 1: Player ranking for players who were on rankings at least 2 years

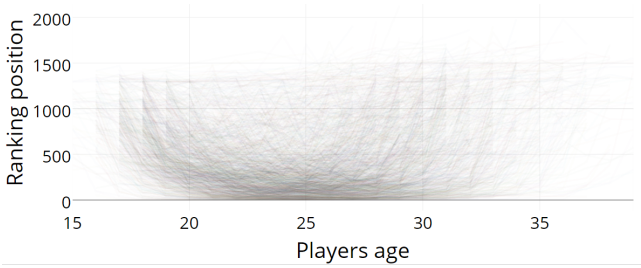


Figure 2: Player ranking for players who were on rankings at least 10 years

We can see that the main difference between graphs is with the players aged between 18 to 25 that are not ranked very high. This is normal because in tennis, if you are ranked lower than 350 you have more expenses than incomes [2]. This results in the fact that promising players might get discouraged and drop their rackets even before they reach their best years. Due to the fact their career endings are not a result of physical nature, keeping these players in the analysis would wrongly alter its results. So, we decided to keep only players with their careers lasting at least ten years. With this rule, we also ensured that phase of getting experience and reaching the top and the phase of physical declination because of age gets included.

2.2 Data normalization

In order to be able to compare and calculated aging curves from players of different quality we had to normalize the performance of each player. We could normalize the player positions or the number of points obtained, but both measures are not very appropriate for normalization. Problem with the position is that it is not equally hard to get 10 positions if you are ranked in top 20 or if you are ranked in top 500. So to include that the normalization would have to be very complex. The problem with gained points is that in 2009 the ATP decided to change the number of point gained on tournaments. So after that date the points are incomparable, so not appropriate to normalize.

To overcome this issue we created surrogate points. For every position on the ranking we calculated average number of points that were needed for that position. The surrogate points obtained in this way are not influenced by the rule changes or by the missing data for number of points, thus we used them for normalization.

3. TENNIS AGING CURVE

As already mentioned in the introduction, there is a lot of talk nowadays about how in modern tennis experienced players are more dominant and better than they used to be and also that the younger players don't have enough quality and experience to win big tournaments.

To determine if this is true, we took the last 30 years of rankings data and for every year divided players into four groups; players younger than 20 years, players between 20 and 24 years, players between 24 and 30 years old and players older than 30 years. In this way we measured young prospect players, players reaching their career peak, experienced players and players that are considered old(er). For every group we summed all the point that the group obtained throughout the year. The results are presented in Figure 3.

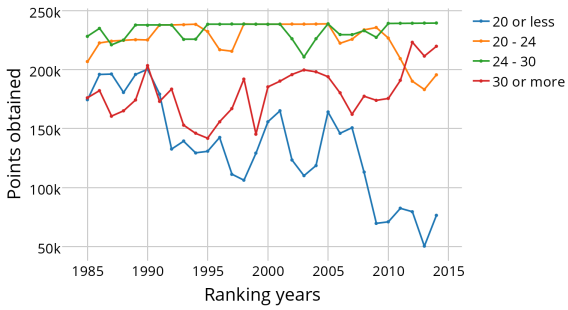


Figure 3: Performance of different age groups.

We can see that performance of younger players really got

worse in the last 10 years. On the contrary the players over 30 years are performing better in the last 10 years. Since this differences in performances could be the reason for shifting the aging curve, we decided to test this and calculate two aging curves. The first one is for players born before 1975 and the second one is for players born after that year. The curves are presented in Figure 4.

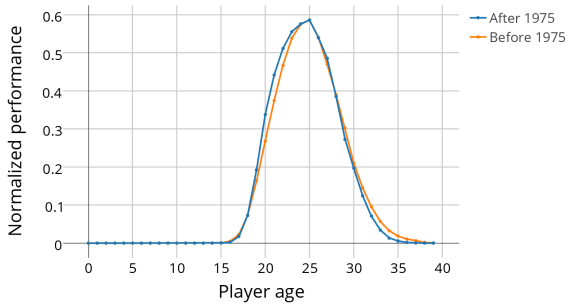


Figure 4: Aging curves for players born before and after year 1975.

We can see a small difference between the curves, but both curves clearly show that a player is most likely to reach his career peak at the age of 25. We can also see that the curve is less steep before the peak and more steep after it. This indicates that players are more gradually approaching their peak performance and that after 25 years the chances of reaching career peak performance are decreasing quickly.

In addition to combining all normalized performances to see what is the average performance over the years, we also wanted to see when players reach their best ranking. The distribution where for every year we counted how many players reached their best ranking is presented in Figure 5.

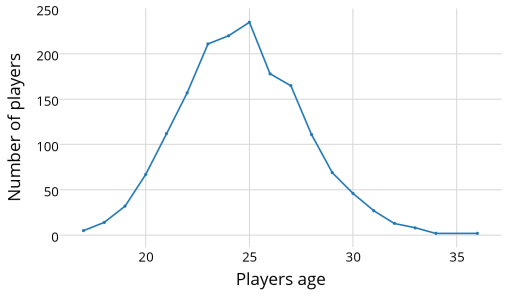


Figure 5: When players reach their best position.

We can see that most players reach their best position when they are 25 years old which coincides with the aging curve.

Since we know that the man's body is reaching its maximum physical capabilities before 25 years, we wanted to know, what is the reason that the peak performance age is only at 25. The reason is experience. Tennis is as much a mental game as it is a physical one. And to be able to handle the pressure and learn to adapt your game to the various opponent styles, you need to play on professional tour for some time. Figure 6 presents how many years after getting a first ATP ranking, players reach their best career position.

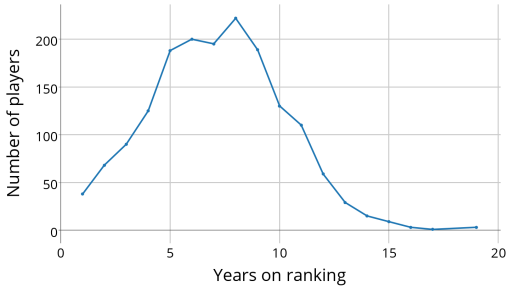


Figure 6: Years needed to reach the best position.

We can see that most player need 9 years on the ATP tour to reach their best position. The distribution presenting players' first appearance on the rankings (Figure 7) shows that most prayers get their first points around the time, when they are 17 or 18 years old. By combining age of first ranking and the years needed to get all the experience, the players are typically 25 years old at their career peak.

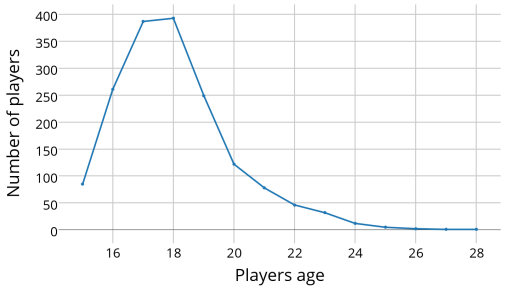


Figure 7: Age of first ranking.

4. PREDICTING PERFORMANCE

Predicting a player's performance can be very interesting, if we want to estimate how high on the rankings a player can get in the next few years. To make such predictions, we designed an algorithm based on similarity measures that finds other players that already had similar career paths. The prediction is then obtained by combining the performances these similar players had in predicted years.

4.1 Prediction procedure

When comparing two players we took the beginning (age of first ranking) of their careers and for every (next) year compare their highest number of obtained points for this year. We summed the absolute differences for every year to get an overall similarity difference between two players for a specific number of years.

After comparing a specific player to all other players, we rank players by their overall similarity difference. The players with smallest differences are picked to contribute to

predictions. The prediction is the average number of points obtained by these player for a specific year.

4.2 Testing predictions

In order to determine the accuracy of the predictions, we decided to use leave-one-out methodology to predict the future number of points for all players in our database.

There are three main parameters that could be tuned: (i) the number of years taken for learning the similarities, (ii) the number of similar players taken for making the predictions and (iii) for how many years in the future we make predictions.

For determining the similarities between players it is true that increasing the number of years used, will improve the similarities and thus also the predictions. We opted for the 5 years long period, since this should be long enough to determine the career trend and also short enough not to miss the expected career peak.

For the number of similar players we did some preliminary tests with 5 and 10 players and the obtained results were similar (Figure 8). For the predictions we choose to use 5 similar players, since we wanted the predictions to include only players with very similar career paths.

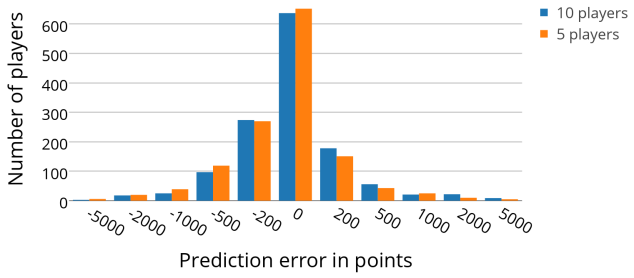


Figure 8: Prediction error for using 5 or 10 similar players for predicting the number of points

To test the prediction accuracy of the algorithm, we predicted players' points for 2, 3 and 4 years in advance. The histogram of prediction errors is shown in Figure 9.

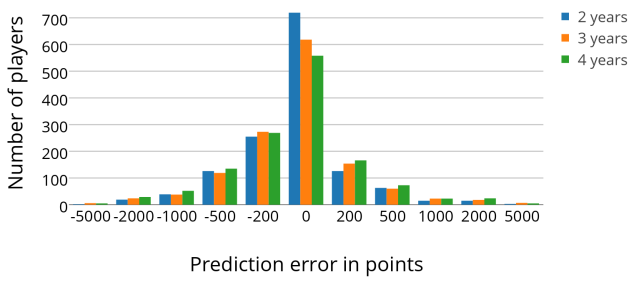


Figure 9: Prediction error for predicting players' points for 2, 3 and 4 years in advance.

The x axis shows the classes of errors. For example, the one around zero represents all predictions that were between -100 and 100 surrogate points away from the actual number of points. We can see that most of the predictions are within the 100 points radius. The medians were 93 for 2 year predictions, 117 for 3 year predictions and 129 for 4 year predictions. There are some players with larger prediction errors, but of course some players get injured or suddenly improve dramatically, so their predictions will be off for a big margin. This events are almost impossible to predict so no matter the algorithm, there will always be some players with wrong predictions.

5. CONCLUSIONS AND FUTURE WORK

In this paper we analyzed the data from professional tennis players ranked on the ATP World Tour rankings between 1974 and 2014. We defined our own measure for comparing players' performances throughout the years disregarding the changes in ranking system.

We presented aging trends and calculated average aging curve for professional tennis players. We showed that a player is most likely to be in his career peak when he is 25 years old.

In addition, we designed algorithm for predicting how many points will a specific player get in the next years. The algorithm is based on finding players with similar careers and for most players prediction error is less then 100 points.

In order to make predictions more accurate, additional player characteristics would have to be used. We could use player rankings data from junior (under 18 years) rankings, detailed statistics obtained for all matches and include additional custom-defined parameters like style of play or some other player specifics.

6. ACKNOWLEDGMENTS

Thanks to Jeff Sackmann for making the data publicly available and to Jernej Zupančič for long and productive discussions.

7. REFERENCES

[1] C. Bialik. Get off my tennis lawn! http://fivethirtyeight.com/features/get-off-my-tennis-lawn. Accessed:2014-07-04.

[2] C. Bialik. Tennis has an income inequality problem. http://fivethirtyeight.com/features/tennis-has-an-income-inequality-problem Accessed:2014-07-04.

[3] R. Schulz and C. Curnow. Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology*, 43(5):113–120, 1988.

[4] N. Silver. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[5] K. J. Stewart. Physical activity and aging. *Annals of the New York Academy of Sciences*, 1055(1):193–206, 2005.

Selection of Classification Algorithm Using a Meta Learning Approach Based on Data Sets Characteristics

Dijana Oreški
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin,
Croatia
+385 42 390850
dijana.oreski@foi.hr

Mario Konecki
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin,
Croatia
+385 42 390834
mario.konecki@foi.hr

ABSTRACT
Many classification algorithms have been proposed, but not all of them are appropriate for a given classification problem. At the same time, there is no good way to choose appropriate classification algorithm for the problem at hand. In this paper, a meta learning data set classification algorithm recommendation, based on characteristics, is presented. An experimental study is performed using 128 real-world data sets and all research made has pointed to the same result: data sets characteristics significantly affect classification accuracy. Guidance in selection of classification algorithm based on data set characteristics is provided.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *Data mining*; F.1.1 [**Computation by Abstract Devices**]: Models of Computation – *Self-modifying machines*.

General Terms
Algorithms, Measurement, Performance, Standardization, Theory.

Keywords
Data set characteristics, meta learning, classification accuracy, neural networks, decision trees, discriminant analysis.

1. INTRODUCTION
In the recent years the area of data mining has experienced a considerable demand for technologies used to extract knowledge from large and complex data sources. There is a substantial commercial interest as well as research activities in the area that aims to develop new and improved approaches for extracting information, relationships, and patterns from datasets.

Classification is the basic task of data mining and a wide variety of approaches have been taken towards this task. All approaches can be categorized into three groups: statistical, machine learning and neural networks [11]. Numerous studies have compared different algorithms, only to conclude that none of the algorithms dominates the rest across several data sets. Thus, none of the algorithms has superiority over competing algorithms without taking into consideration data set characteristics. The optimal classifier for a classification task is determined by the characteristics of the data set employed; understanding the relationship between data characteristics and the performance of classifiers is therefore crucial for the process of classifier selection [15]. Van der Walt has identified data set characteristics important for classification task. He has divided data characteristics into five groups: standard measures, data sparseness measures, statistical, information theoretic and noise measures [15]. Empirical and theoretical approaches have been employed in the literature to define this relationship. For instance, Ajay Kumar Tanwani, et al. [14] have provided some guidelines for selecting a machine learning technique by applying six classifiers: Naïve Bayes, Neural Network, Support Vector Machine, Instance Based Learner, Decision Tree and Inductive Rule Learner to 31 biomedical datasets. Authors have done promising research but they have limited their data to only specifically medical data, thus restricting their analysis to a specific domain.

In the study presented in this paper, real-world data sets from various domains have been considered and worked on. In [1] author has provided conclusion by generating artificial data and by studying the relationship between data characteristics with classifiers' performance. Selected classifiers are: NB, Gaussian Network, DT, SVM, MLP, and kNN. Three different experiments were conducted on artificial data. This study has made several points related to classifiers' performance clearer but data set characteristics like number of features, number of instances, presence of noise and colinearity were not considered. Another weak point is the fact that author has taken in account only the artificial data, again limiting the area of observation.

It remains an intriguing challenge to fully describe relationships between data characteristics and classifier behavior, and to develop approaches that automatically select classifiers that are appropriate for a given set of data. Since none of the previous approaches have been successful in accurately predicting or explaining classifier performance on real-world data, in this paper three different classification methods which belong to different approaches have been compared: statistical approach (discriminant analysis), machine learning (decision trees) and neural approach (neural networks). Methods have been compared on 128 data sets of different characteristics. Based on the results of the analysis, meta learning approach was used to explain relationship between classification accuracy and data set characteristics.

2. RESEARCH FRAMEWORK
In this section, the experiment design of proposed research is presented. The process of experiment includes the following (as shown in Figure 1):

- identifying characteristics of data sets,
- performing classification on each data set,
- evaluation of results,

- meta learning.

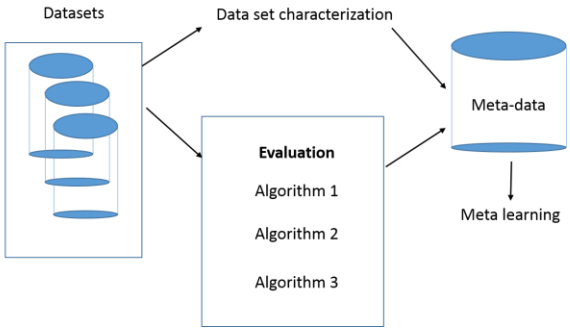


Figure 1. Research framework.

In the following four subsections each step of the experiment is explained.

2.1 Data set characterization

Data sets used in the analysis have been downloaded from four publicly available repositories. Each of the data sets is characterized by the following approach of Van der Walt, as shown in Table 1 [15].

Table 1. Data set characteristics

Data set characteristic	Acronym
Standard measures	
Number of features (Dimensionality)	d
Number of instances	N
Number of classes	C
Data sparsity meaures	
Data sparsity ratio	DSR
Data sparsity	DS
Statistical measures	
Correlation of features	p
Multivariate normality	MVN
Homogenity of class covariances	SDR
Information theoretic measures	
Intrinsic dimensionality	ID
Intrinsic dimensionality ratio	IDR
Noise measures	
Feature noise	ID2

Van der Walt lists three generic measures that are used to normalize many of other measures. These include the following standard measures: the number of features, the number of instances and the number of classes [15]. Data sparsity measure defines relationship between the dimensionality of the data and the number of samples required to model the data accurately. This measure indicates how sparse data is by taking the dimensionality, number of classes and number of instances in a data set into account. Statistical measures measure the correlation between features, the multivariate normality of class-conditional probability density functions and the homogeneity of class covariance matrices. Pearson correlation has been used to measure correlation coefficient, Kolmogorov-Smirnov test for multivariate normality and Box's M test for examining homogeneity. The mutual information between classes and features is used to determine the intrinsic dimensionality of a data set. Feature noise indicates the proportion of features that do not contribute to classification.

2.2 Classification methods

For evaluating the classification performance three different methods have been used. These methods are explained briefly in the next subsections.

2.2.1 Discriminant analysis

Discriminant analysis accepts a random sample of observations defined by a set of variables and generates a discriminant function that classifies observations into two groups by minimizing the expected misclassification cost [8]. This method assumes that all variables are normally distributed. Furthermore, it requires identical covariance matrices. Thus, discriminant analysis is performed only on data which satisfied these assumptions. In this research, Fisher's [3] discriminant analysis procedure, a widely used DA function, has been implemented.

2.2.2 Neural networks

Artificial neural networks (ANN) are non-linear mapping systems with a structure loosely based on principles observed in biological nervous systems. ANN offer many advantages over conventional statistical methods [13]. The ANN use the data to develop an internal representation of the relationship between the variables, and they do not make assumptions about the nature of the distribution of the data. Another advantage is that while traditional discriminant analysis is not adaptive, ANN read just their weights, as new input data becomes available [9; 10; 12]. The model that is used in this paper is based on a special case of a feed-forward neural network known as a multi-layer perception (MLP). An MLP neural network is a non-linear, non-parametric regression model commonly referred to as a function approximator. The ANN used in this study are fully connected, feed-forward MLP neural networks with three layers: an input layer, a hidden layer and an output layer. The topology of the neural networks used in this paper is the following: all input nodes ("neurons") are connected to every hidden node and every hidden node is connected to the output nodes. The literature states that a single hidden layer in the neural network is sufficient for the network to be a universal function approximator [2; 4; 6; 7].

2.2.3 Decision tree

A decision tree is created in two phases: (i) Tree Building Phase: repeatedly partition the training data until all examples in each partition belong to one class or the partition is sufficiently small, (ii) Tree Pruning Phase: remove dependency on statistical noise or variation that may be particular only to the training set. One of the many advantages decision trees have to offer is that decision trees make no prior assumptions about the nature of the data.

2.3 Evaluation

In this paper decision trees are applied in order to identify important data set characteristics for each classification method.

Decision trees parameter called contribution indicates effect of each characteristic on classification method accuracy. Cross validation technique is applied in tree induction. Cross validation consists of partitioning the data set into a number of subsets and then running a given algorithm a number of times, each time using a different training set.

2.4 Meta learning approach

Meta learning is a framework developed in the field of supervised machine learning with the aim of automatically predicting algorithm performance, thus assisting users in the process of algorithm selection [5; 16]. In Meta learning, knowledge is acquired by the meta-examples that store: (a) The features that describe the dataset (problem), (b) Performance information obtained by executing candidate algorithms on training datasets. After generation of meta-examples, Meta learner (learning algorithm) is applied to acquire knowledge that relates performance of candidate algorithms to the features of the datasets (problems).

3. RESEARCH RESULTS

The discussion of the study findings is organized around an effect of data set characteristics on classifier performance. To determine if data set characteristics affect elapsed time and classification accuracy, decision tree was used. The results are presented in Figures 2, 3 and 4. Classification accuracy of neural network tends to be determined by the number of features and number of instances. Other than that, correlation, data set sparsity and multivariate normality seem to affect accuracy. Results of the data sets characteristic effects on neural networks accuracy are presented in Figure 2.

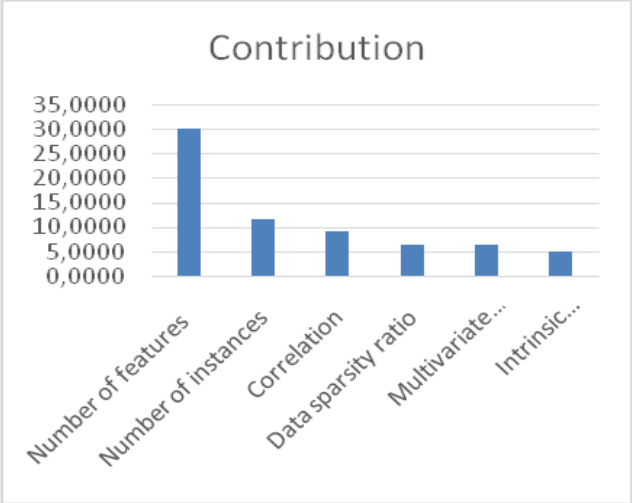


Figure 2. Data sets characteristics impact on the neural networks performance.

When it comes to the discriminant analysis, it is interesting that accuracy mostly depends on the intrinsic dimensionality ratio. It also seems that the number of instances and correlation coefficient in data set is related to classification accuracy (see Figure 3).

performance. Increase in a sample size improves performance of decision tree. It can also be stated that the discriminant analysis classification is mostly affected by high intrinsic dimensionality ratio.

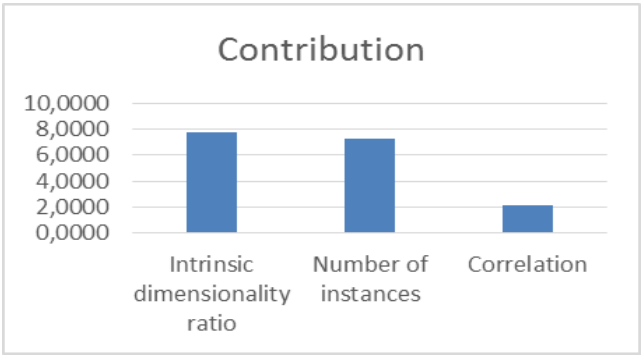


Figure 3. Data sets characteristics impact on the discriminant analysis performance.

In the case of the decision tree classification, the number of instances and correlation tends to be important. Furthermore, accuracy increases more when the data sparsity ratio is low and when the number of features is low. It can be concluded that decision tree is inadequate for a large data set volume, because it is highly influenced by both factors: the number of features and the number of instances (see Figure 4).

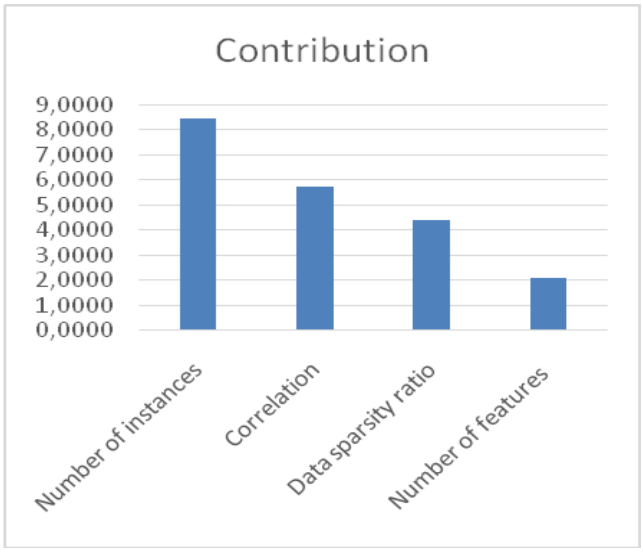


Figure 4. Data sets characteristics impact on the decision tree performance.

4. CONCLUSION

This paper presents the results of a meta learning research based on data set characteristics. Extensive empirical comparison of 128 data sets of different characteristics was conducted. Experiment included 384 classification analysis and almost 900 analysis of data sets characteristics (for each data set, necessary tests were performed in order to identify characteristics). Therefore, there is a total of 1280 analysis included into comparison. Results support hypothesis that classification algorithms are sensitive to changes in data characteristics. High correlation and a high number of instances and features affect neural networks performance. The method remains superior to other methods in relative

5. REFERENCES

[1] Assareh, A., Moradi, M. H., and Volkert, L. G. 2008. A hybrid random subspace classifier fusion approach for

protein mass spectra classification. *EvoBIO 2008. LNCS*. 4973, 1–11. DOI= 10.1007/978-3-540-78757-0_1.

[2] Cybenko, G. 1989. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals, and Systems*. 2, 4, 303-314.

[3] Fisher, R. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. 7, 2, 179-188.

[4] Funahashi, K. I. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks*. 2, 3, 183-192.

[5] Giraud-Carrier, C., Vilalta, R. and Brazdil, P. 2004. Introduction to the special issue on meta-learning. *Machine Learning*. 54, 3, 187- 193. DOI= 10.1023/B:MACH.0000015878.60765.42.

[6] Hornik, K., Stinchcombe, M. and White, H. 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks*. 2, 5, 359-366.

[7] Hornik, K., Stinchcombe, M., and White, H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks*. 3, 5, 551-560.

[8] Kiang, M. Y. 2003. A comparative assessment of classification methods. *Decision Support Systems*. 35, 4, 441-454. DOI= http://dx.doi.org/10.1016/S0167-9236(02)00110-0.

[9] Krishnaswamy, C. R., Gilbert, E. W., and Pashley, M. M. 2000. Neural Network Applications in Finance: A Practical Introduction. Financial Practice and Education. 10, 75-84.

[10] Kuo, C. and Reitsch, A. 1995. Neural Networks vs. Conventional Methods of Forecasting. *The Journal of Business Forecasting*. 14, 4, 17-22.

[11] Michie, D., Spiegelhalter, D. J., and Taylor, C. C. 1994. *Machine Learning, Neural and Statistical Classification*.

[12] Pao, Y. 1989. *Adaptive Pattern Recognition and Neural Networks*.

[13] Shachmurove, Y. 2005. Business Applications of Emulative Neural Networks. *International Journal of Business*. 10, 4.

[14] Tanwani, A. K., Afridi, J., Shafiq, M. Z., and Farooq, M. 2009. Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. In *Proceedings of European Conference, EvoBIO 2009*, (April 2009), 128-139. DOI= 10.1007/978-3-642-01184-9_12.

[15] Van der Walt, C. M. 2008. Data measures that characterize classification problems, Master's Dissertation. *http://repository.up.ac.za/handle/2263/27624*. Accessed: 2015-09-12.

[16] Vilalta, R. and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Journal of Artificial Intelligence Review*. 18, 2, 77- 95. DOI= 10.1023/A:1019956318069.

Using Shadowed Clustering and Breeder GA in the Imbalanced Data Classification Problems

[Extended Abstract]

Ante Panjkota
Department of Economics and Department of Maritime Studies, University of Zadar
Splitska 1
Zadar, Croatia
apanjkot@unizd.hr

Ivo Stančić, Josip Musić
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split
Rudjera Boškovića 32
Split, Croatia
istancic@fesb.hr, jmusic@fesb.hr

Miha Drole
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
Ljubljana, Slovenia
miha.drole@fri.uni-lj.si

Petar Vračar
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
Ljubljana, Slovenia
petar.vracar@fri.uni-lj.si

Igor Kononenko
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
Ljubljana, Slovenia
igor.kononenko@fri.uni-lj.si

Matjaž Kukar
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
Ljubljana, Slovenia
matjaz.kukar@fri.uni-lj.si

ABSTRACT

We discuss possibilities and advantages of using a novel oversampling approach based on the synthetic minority instances generation by using shadowed clustering and Breeder genetic algorithm for dealing with imbalanced classification problems. Viability of the proposed oversampling method is confirmed through comparison with state-of-the-algorithms in experiments on four imbalanced datasets obtained from a Blender-based simulator. In all conducted experiments the proposed oversampling approach performs at least as well as state-of-the-art algorithms. Our results indicate a great potential for future ensemble algorithm formed from orthogonal projections and oversampled as proposed in this.

Categories and Subject Descriptors

G.4.1 [**Computing methodologies**]: Cluster Analysis; G.4.1 [**Computing methodologies**]: Feature selection

General Terms

Imbalanced data classification, minority class oversampling, ensembles for classification

1. INTRODUCTION

With huge advancement in technology, modern society is focusing on raising the quality of the life for vulnerable individuals and groups in everyday situations. People with low vision/blindness are in focus of our work. Activities of daily living are often very difficult and in some cases impossible for such groups. One such activity is independent travel between two points (indoor or outdoor). Self-adapting machine learning techniques can be used in such situations [1] to identify potentially dangerous situations. Outdoor navigation is generally the more challenging task since surroundings are highly dynamic with little or no regular structure. Related approaches dealing with similar problems include utilizing Bayesian statistic to aid in navigation through indoor environment, crosswalk and stairs detection with RGBD cameras using Hough transform and line fitting with geometric constraints [2], and multi-sensory learning [1]. However, often they are plagued by a great number of false alarms due to oversimplified input space.

There are in general few cases of the potentially dangerous trajectories that can cause injury to visually impaired person. From machine learning perspective these tasks are usually characterized with: (1) significant imbalance between class sizes, and (2) large penalty for wrong classification which can lead to serious injury. Among most successful methods for dealing with class imbalance are sampling methods, more precisely minority class oversampling methods [3, 4]. One of the limiting factors of such systems is environmental complexity resulting in large amounts of data much of which can not be used for intended application. In our paper we present a novel approach for dealing with binary imbalanced data classification problems based on minority class oversampling by combining shadowed clustering [5] and Breeder GA algorithm [6]. Our research is directed toward development of reliable and accurate algorithm that can be used in assistive systems for visually impaired in tasks of obstacles/objects avoidance.We use a Blender-based simulator to obtain depth-sensor data as proposed in [7]. The use of 3D surface scanners in simulation environments has already been successfully achieved which makes this sensing

technology a good candidate for both our current testing and future system. Data obtained from simulator are used to create four scenarios that emulate real-life situations. Viability and perspectives of the proposed minority oversampling method based on a shadowed clustering and Breeder GA is demonstrated through comparison with several state-of-the-art machine learning algorithms for handling imbalanced data classification. Finally, we analyze and discuss obtained results and highlight the road-map toward future ensemble algorithms which will implement the oversampling method proposed in this paper.

2. SHADOWED CLUSTERING AND BREEDER GA OVERSAMPLING APPROACH

Sections obtained from shadowed clustering in the minority class are oversampled by synthetically generated minority instances using Breeder GA algorithm. Every region in minority class is oversampled proportional to its size. This means that in the case of two clusters from the shadowed clustering phase (with joint ratio of 1:3) we automatically have defined the number of synthetically produced instances in every cluster so in finale all oversampled clusters are producing desired relation between majority and minority class size (commonly 1:1). In this way unchanged ratio in relative size of the all produced clusters is secured. At the end of the section pseudo code for the minority class oversampling technique based on shadowed clustering and Breeder GA (Shadowed Clustering based Minority Oversampling Technique – SCMOT) are presented.

2.1 Shadowed clustering

Shadowed clustering is based on shadowed set theory. This theory developed out of the simplified view on fuzzy sets. Shadowed set can be obtained from fuzzy set via $BZMV^{dM}$ algebra which uniquely defines transformation operator from shadowed to fuzzy set while reverse does not hold. Shadowed sets are introduced as necessary tool for observing and analyzing problem sets for which classical set theory does not provide satisfactory interpretation.

Shadowed clustering is derived from fuzzy clustering with introduction of threshold $\lambda \in [0, 1]$ on the membership function [5]. In this manner cluster structure can be determined more reasonably. Shadowed set A is defined as mapping $A : X \mapsto \{0, (0, 1), 1\}$ (where X is observed space). Codomain interpretation has the following meaning: elements for which $A(x) = 1$ represent shadowed set (cluster) kernel fully reflecting set A concept, elements for which $A(x) = 0$ do not belong to set A, while elements which have membership function value in the interval $(0, 1)$ do not have membership defined and are members of a shadowed set.

2.2 Basic principles of synthetic data sample generation using Breeder GA algorithm

Widely used group of algorithms for synthetic oversampling of minority class are from family of SMOTE algorithms [8, 9]. Main drawback of such algorithms is reduced flexibility in synthetic data generation. More flexibility potentially could be achieved with algorithms from evolutionary computing (i.e. genetic algorithms). However, basic Breeder GA algorithm [6] can be modified in a way as to generate new, synthetic data from some initial data set. In the process only two operations will be used: recombination and mutation.

Selection phase in a way also exist here, but since all individuals have same fitness function, all pairs from input data will be taken into account in the next recombination phase. Recombination process generates new instances by combining data from two or more parents (in our proposal we confine ourselves with only two parents). Depending on a way in which parent variable values are represented, different recombination can be achieved. We propose use of discrete and line recombination. So, two possible synthetic samples generation approaches are defined: (1) parents in Breeder GA algorithm are all possible pairs within the same core **CORE Breeding** (CB), or (2) parents in the Breeder GA algorithm are all possible shadow-core combinations for the same cluster **CORE and SHADOW Breeding** (CSB)

In case that insufficient number of cluster instances are obtained for balancing out classes, required number could be obtained with mutations on offsprings. It is worth noting that this situation should not happen since with randomness in recombination process arbitrarily large number of offsprings can be obtained. Thus, mutation procedure only enables more flexibility in the synthetic data generation.

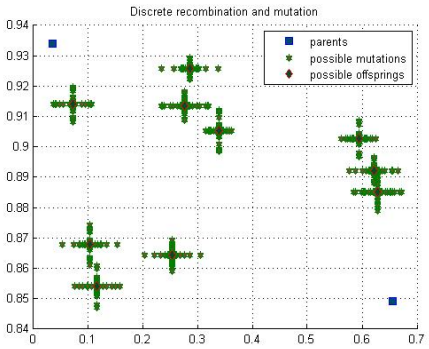


Figure 1: Mutation results after discrete recombination. Note that after line recombination, all possible offsprings would lie roughly on a line connecting the parents.

As in case of discrete recombination this procedure outputs synthetic instances within the hyper-cube along the coordinate axis. Results of mutation after discrete and line recombination are depicted in Figure 1.

2.3 SCMOT algorithm

After oversampling, shadowed clustering method is performed on the all minority instances with certain number of clusters. For relatively few dimensions (≤ 6) number of clusters can be estimated with known methods, e.g. Davies-Bouldin index and Silhouette method.

The first oversampling mode (CB) has all pairs of the particular core as a recombination parents in the Breeder GA algorithm, while in the second mode (CSB) parents in the Breeder GA algorithm are all possible pairs from core and shadow of the defined clusters. In the existence of the outliers only CORE Breeding mode is at our disposal. Upon

selection of the oversampling mode each cluster has been populated in a relative size to defined clusters. At the end of the oversampling procedure the desired ratio between majority and minority class is achieved. The described procedure is depicted in Algorithm 1.

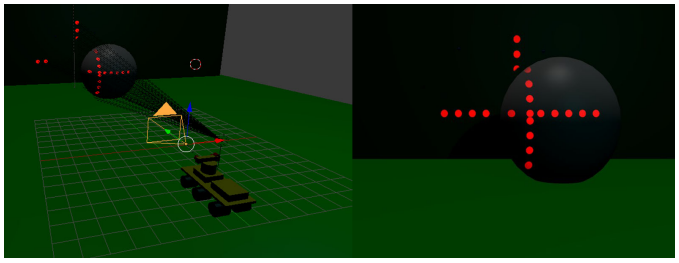
Algorithm 1 SCMOT

```
procedure SCMOT(data)
Ensure: Extraction of the minority class from data
Ensure: Choose number of clusters (numCLS) OR evaluate number of clusters in minority class
    SCM ← v, λ, u        ▷ Performing SCM on minority class with numCLS
    if NumOutliers = 0 then
        for do(k = 1 : numCls)
            synthetic ← (CB) OR (CSB)        ▷ Recombination and Mutation
        end for
    else
        for do(k = 1 : numCls)
            synthetic ← (CB) ▷ Recombination and Mutation
        end for
    end if
    ovsamMin ← merge all minority and all synthetic
    return ovsamMin
end procedure
```

3. SIMULATOR

The simulator used in this paper is based on adaptive robot 3D scanner [7]. It operates as a simple stereo vision system, consisting of a laser pattern projector as the active component, and a camera as the passive component. The laser projector projects a star-like structural light pattern (consisting of 21 points) on the surface of the target area (Figure 2 right). The three-dimensional shape and location of the sensed object is reconstructed by comparing the known locations of each projected point (projector P_{pr}) with the corresponding point on camera image (P_{cam}), as seen in left of Figure 2. Scanner's output data is a point cloud, which represents 3D locations of all projected points (surface points).

Figure 2: Blender model of robot with 3D scanner (left), projected patterns as seen from scanners camera (right)



Up to two 3D sphere-like objects that represent potentially dangerous obstacles for the robot are simulated. In the simulations, in total four scenarios were simulated: miss (object is passing far from the robot) - DS1, near hit (object is passing near the robot) - DS2, hit (object is on a collision course with the robot) - DS3 and hit with 2 objects (one object is missing, while second object is colliding with the robot) - DS4. Each scenario was repeated 6 times, with random object placements, movement speeds and directions. Output of Blender simulation is video stream. Simulation inherits some of scanner's defects like projection errors, projected point distortion, change in point intensity level etc. Based on video stream data 3D locations of projected points were calculated.

4. EXPERIMENTAL PROCEDURE

Basic information of the data sets used in four Experiments 1-4 are given in the Table 1.

Table 1: Basic data set informations used in experiments

data set	Nmin	Nmaj	imbalance ratio
DS1	125	1000	8
DS2	100	500	5
DS3	95	665	7
DS4	80	320	4

In all 2D combinations and one 3D case (of the primary 3D data) class balance have been produced with ratio 1:1 by using SCMOT approach described with Algorithm 1. Class balance is only performed on the training folds in classical 10-folds CV procedure, while test folds are kept in the originally imbalanced ratio [10]. SCMOT oversampling approach was combined with known algorithms Naive Bayes (NB), Support Vector Machines (SVM), Multi Layer Perceptron (MLP), Nearest Neighbor Search (NNS), Ripper, C45 and Decision Stump (DS). Quality of the proposed oversampling approach (SCMOT) was confirmed through comparison with the state-of-the-art algorithms for handling imbalanced data classification. For that purpose IIVOTES, OverBagging, SMOTEBoost, MSMOTEBoost and standard SMOTE procedure were used. Combinations of all 2D orthogonal projections and one 3D projection were oversampled by using SCMOT approach to produce classification ensemble. Final decision of a such ensemble is defined with majority voting process. These ensembles were also compared with the mentioned state-of-the-art algorithms.

5. RESULTS AND DISCUSSION

Results presentation and corresponding discussions is performed for all four experiments on datasets in Table 1. Experiment 1 corresponds to dataset 1 (DS1), Experiment 2 to dataset 2 (DS2) and so on.

Table 2: Classification results of the data set DS1 to DS4 - best ensemble from all orthogonal projections with SCMOT and best classifier among state-of-the-art algorithms

	ACC	F-measure	AUC
Experiment 1 - DS1			
Ripper + SCMOT	**0,9680**	**0,8571**	**0,9680**
SMOTEBoost	*0,9493*	*0,7782*	*0,8843*
Experiment 2 - DS2			
Ripper + SCMOT	**0,9667**	**0,8980**	**0,9618**
SMOTEBoost	**0,9667**	*0,8913*	*0,9100*
Experiment 3 - DS3			
Ripper + SCMOT	**0,9737**	**0,8969**	**0,9872**
SMOTEBoost	*0,9711*	*0,8791*	*0,9158*
Experiment 4 - DS4			
Ripper + SCMOT	*0,9525*	*0,8774*	**0,9771**
SMOTEBoost	**0,9625**	**0,9007**	*0,9203*

As can be seen from Table 2 we only reported results of the best performing ensemble with SCMOT procedure and best classifier from state-of-the-art group. For Experiments 1 to 4 obtained results on DS1-DS4 indicate Ripper as the

best base learner for ensemble based on using SCMOT on orthogonal projections while best state-of-the-art algorithm was SMOTEBoost. Results in the Table 2 are showing comparable values in all performed Experiments (1-4) between SMOTEBoost and ensemble with Ripper + SCMOT procedure.

Overall results on all four datasets, without calculation of statistical significance, and with highlighted number of highest values for ACC, F-measure and AUC parameter are presented in Table 3. Data in Table 3 verified our main intention to demonstrate feasibility of application of the SCMOT procedure on orthogonal projections and ensemble construction from such areas for the purpose of imbalanced set classification.

Table 3: Aggregate number of wins/ties/losses of each algorithm against others over datasets in Table 1 based only on maximal values for particular evaluation measure

	ACC	F-measure	AUC
IIVOTES	0/0/4	0/0/4	0/0/4
MSMOTEBoost	0/0/4	0/0/4	0/0/4
OverBagging	0/0/4	0/0/4	0/0/4
SMOTEBoost	1/1/2	1/0/3	0/0/4
SMOTE	0/0/4	0/0/4	0/0/4
Ensemble with SCMOT	**2/1/1**	**3/0/1**	**4/0/0**

6. CONCLUSIONS

In this paper we discuss possibilities and features of a novel approach for imbalance data classification based on minority class oversampling by using shadowed clustering in combination with Breeder GA. Viability of the proposed Shadowed Clustering based Minority Oversampling Technique (SCMOT) has been confirmed by comparison with several state-of-the-art algorithms in classification of four imbalanced datasets obtained from Blender based simulator. In all experiments SCMOT approach had comparable quality with chosen state-of-the-art algorithms evaluated through ACC, F-measure and AUC. Besides comparing favorably in quantitative terms, SCMOT also has the following advantages like high flexibility in synthetic instances generation because of the shadowed clustering combination with Breeder GA and possibility for producing quality ensemble multiple diverse through data level (synthetic data generation flexibility) and on the features multiplicity by combining SCMOT in different orthogonal projections. By combining all orthogonal projections oversampled by SCMOT in one ensemble, with majority voting policy, a good solution for classification of the imbalanced datasets was obtained. Ensembles built in this way achieved the highest number of top performance results. This is a good guarantee for its perspectives in classification of the imbalanced datasets.

The benefits of balancing the data sets are not limited to propositional learning approaches, but also apply to relational learning (e.g. inductive logic programming). Since relational learning approaches often guide their search by rewarding a hypothesis' (correct) coverage and penalizing its complexity, they are likely to find sub-optimal solutions when presented with imbalanced datasets, since the minority class severely limits the maximal allowed complexity of the produced hypothesis. A basic assumption for the further development of ensembles based on the SCMOT method in orthogonal projections is multidimensionality of the input data. For this purpose some sub-problems need to be addressed, such as finding the upper projection dimensionality where shadowed clustering can be used, discovering favorable orthogonal projections (good attributes) and fast estimation of the number of clusters. These sub-problems are among main directions of our further research.

7. REFERENCES

[1] S. Krishnas and K. Sujitha, "Machine Learning approach for Assisting Visually Impaired," *International Journal for Trends in Engineering and Technology*, vol. 3, pp. 119–122, 2015.

[2] Y. Tian, "RGB-D sensor-based computer vision assistive technology for visually impaired persons," in *Computer Vision and Machine Learning with RGB-D Sensors*, pp. 173–194, Springer, 2014.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[4] Z. Xie, L. Jiang, T. Ye, and X. Li, "A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning," in *Database Systems for Advanced Applications*, pp. 3–18, Springer, 2015.

[5] S. Mitra, W. Pedrycz, and B. Barman, "Shadowed c-means: Integrating fuzzy and rough clustering," *Pattern Recognition*, vol. 43, pp. 1282–1291, Apr. 2010.

[6] D. Schlierkamp-Voosen and H. Mühlenbein, "Predictive models for the Breeder genetic algorithm," *Evolutionary Computation*, vol. 1, no. 1, pp. 25–49, 1993.

[7] I. Stančić, J. Musić, and M. Cecić, "A novel low-cost adaptive scanner concept for mobile robots," *Ingeniería e Investigación*, vol. 34, no. 3, pp. 37–43, 2014.

[8] K. Li, W. Zhang, Q. Lu, and X. Fang, "An improved SMOTE imbalanced data classification method based on support degree," in *Identification, Information and Knowledge in the Internet of Things (IIKI), 2014 International Conference on*, pp. 34–38, IEEE, 2014.

[9] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences*, vol. 291, pp. 184–203, 2015.

[10] D. Košir, Z. Bosnic, and I. Kononenko, "The use of prediction reliability estimates on imbalanced datasets: A case study of wall shear stress in the human carotid artery bifurcation," in *Medical Applications of Intelligent Data Analysis: Research Advancements*, 2012.

Izboljšano ocenjevanje pomembnosti zveznih značilk

Matej Petković[1,2,3]
matej.petkovic@ijs.si

Panče Panov[3]
pance.panov@ijs.si

Sašo Džeroski[2,3]
saso.dzeroski@ijs.si

Fakulteta za matematiko in fiziko Univerze v Ljubljani, Jadranska 19, 1000 Ljubljana, Slovenija
Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, 1000 Ljubljana, Slovenija
Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija

POVZETEK

V tem delu predstavimo novo različico algoritma ReliefF, v kateri želimo odpraviti podcenjevaje zveznih značilk, ki velja za eno izmed glavnih težav tega algoritma. S tem namenom uvedemo nov način merjenja razdalje, ki se uporablja v algoritmu. To vzbudi potrebo po ocenjevanju verjetnostnih porazdelitev značilk, za kar uporabimo EM-algoritem. Novo različico algoritma primerjamo s staro na nekaj podatkovnih množicah, in sicer glede na kakovost ocenjevanja in časovno zahtevnost. Ugotovimo, da je naša metoda boljša, vendar počasnejša.

1. UVOD

V zadnjem času se na področju napovednega modeliranja srečujemo z vse kompleksnejšimi napovednimi problemi, kjer je razsežnost vhodnega prostora zelo visoka. S takšnimi vhodnimi podatki se spopadamo tudi tako, da ocenimo pomembnost vhodnih spremenljivk za napovedovanje ciljne spremenljivke. Vhodne spremenljivke nato uredimo po pomembnosti in pri napovedovanju upoštevamo le nekaj najkoristnejših. Tako lahko izboljšamo tako natančnost napovedovanja kot tudi učinkovitost algoritmov za napovedno modeliranje.

Obstaja veliko metod za rangiranje značilk. V primeru, ko je ciljna spremenljivka diskretna, je ena izmed najbolj znanih ReliefF. Dokazano je bilo [4], da podcenjuje zvezne značilke v primerjavi z diskretnimi, kar poskušamo v tem delu popraviti. Ker jedro algoritma ReliefF predstavlja merjenje razdalje, tj. definicija bližine, skušamo podcenjevanje vsaj omiliti s tem, da predlagamo novo metriko. Ta temelji na ocenjevanju verjetnostnih gostot značilk. Ker teh porazdelitev večinoma ne poznamo, posežemo po EM-algoritmu [3], s pomočjo katerega najdemo približek dane gostote v prostoru Gaussovih jedrnih funkcij.

Obe različici algoritma preizkusimo na nekaj podatkovnih množicah z repozitorija UCI. Njuno kakovost napovedovanja primerjamo z metodologijo, ki jo je razvil Slavkov s sodelavci [5]. Posebej izmerimo še časovno zahtevnost obeh različic.

2. OZADJE

Vseskozi predpostavljamo, da so podatkovne množice S, na katerih bomo uporabljali naša algoritma, podane kot tabela, ki jo sestavlja $|S|$ primerkov $(\mathbf{x}, y)$, kjer je $\mathbf{x}$ vektor n vrednosti vhodnih spremenljivk x_i, $1 \leq i \leq n$, in je y ena od končno mnogo možnih vrednosti ciljne spremenljivke.

Razdalja med vrednostma i-te komponente primerkov $\mathbf{r}$ in $\mathbf{s}$ je definirana s predpisom

$$d_i(\mathbf{r}, \mathbf{s}) = \begin{cases} 0; & \mathbf{r}_i = \mathbf{s}_i \\ 1; & \text{sicer} \end{cases}, \tag{1}$$

če je x_i diskretna značilka, in

$$d_i(\mathbf{r}, \mathbf{s}) = \frac{|\mathbf{r}_i - \mathbf{s}_i|}{\max\limits_{\mathbf{t} \in S} \mathbf{t}_i - \min\limits_{\mathbf{t} \in S} \mathbf{t}_i} \tag{2}$$

v zveznem primeru. Razdalja med primerkoma $\mathbf{r}$ in $\mathbf{s}$ je podana s predpisom $d(\mathbf{r}, \mathbf{s}) = \sum_i d_i(\mathbf{r}, \mathbf{s})$.

Pojasnimo motiv, na katerem temelji algoritem ReliefF. Izberemo primerek $\mathbf{r}$ in najdemo njegovega bližnjega soseda $\mathbf{s}$. Če sta istega razreda y, je za značilko x_i škodljivo, če je razdalja $d_i(\mathbf{r}, \mathbf{s})$ velika, saj sklepamo, da x_i ne more močno vplivati na določanje razreda. Če sta primerka različnih razredov, je za značilko x_i koristno, če je $d_i(\mathbf{r}, \mathbf{s})$ dovolj izražena, saj to pomeni, da je (so)odgovorna za spremembo razreda.

Merjenje razdalj in lokalnost sta ključnega pomena, saj tako zagotovimo, da ReliefF upošteva tudi interakcije med značilkami. Njegovo osnovno različico predstavimo v algoritmu 1. Vidimo, da na vsakem koraku iteracije najdemo primerku $\mathbf{r}$ k najbližjih sosedov iz istega razreda (*hits*) in k najbližjih sosedov iz vsakega izmed preostalih razredov (*misses*). Na koncu posodobimo uteži w_i v skladu z zgornjo motivacijo, pri čemer prispevke posameznih razredov utežimo glede na njihovo zastopanost.

Algoritem 1 ReliefF(S, m, k)

w ← ničelni seznam dolžine n
for $j = 1, 2, \ldots, m$ **do**
  r ← naključno izberi primerek iz S
  $\mathbf{h}_1, \ldots, \mathbf{h}_k \leftarrow k$ primerku **r** najbližjih zadetkov
  for all $c \in Y \setminus \{\mathbf{r}_y\}$ **do**
    $\mathbf{m}_{c,1}, \ldots, \mathbf{m}_{c,k} \leftarrow k$ primerku **r** najbližjih zgreškov razreda c
  end for
  for $i = 1, 2, \ldots, m$ **do**
    dobro $\leftarrow \sum_{c \in Y \setminus \{\mathbf{r}_y\}} \frac{P(c)}{1 - P(\mathbf{r}_y)} \sum_{\ell=1}^{k} d_i\left(\mathbf{m}_{c,\ell}, \mathbf{r}\right) / mk$
    slabo $\leftarrow \sum_{\ell=1}^{k} d_i\left(\mathbf{h}_\ell, \mathbf{r}\right) / mk$
    $w_i \leftarrow w_i +$ dobro $-$ slabo
  end for
end for
vrni w

Kot smo že omenili, algoritem 1 sistematično podcenjuje zvezne značilke. Ker diskretne različice razdalje (1) ne moremo popraviti, poskusimo to storiti z zvezno.

3. PREDLOG IZBOLJŠAVE

Opazimo, da v primeru (2) velja

$$d_i(\mathbf{r},\mathbf{s}) = \left| \int_{\mathbf{r}_i}^{\mathbf{s}_i} f_{x_i}(t)dt \right|, \tag{3}$$

če je x_i porazdeljena enakomerno zvezno $EZ(a,b)$ in parametra (a,b) ocenimo po metodi največjega verjetja. To nas napelje na misel, da bi predpis (3) vzeli za definicijo razdalje tudi za preostale verjetnostne gostote f_{x_i}. Zgornje lahko še enostavneje zapišemo kot

$$d_i(\mathbf{r},\mathbf{s}) = P(\min\{\mathbf{r}_i,\mathbf{s}_i\} < x_i < \max\{\mathbf{r}_i,\mathbf{s}_i\}).$$

Razdalja d_i je očitno še vedno normirana, prav tako je prejšnji predpis (2) le poseben primer novega. Domnevamo, da novi predpis bolje opisuje razdalje med vrednostmi značilk. Naj bo $t = |\mathbf{r}_i - \mathbf{s}_i|$. Če sta vrednosti $\mathbf{r}_i$ in $\mathbf{s}_i$ pri repih porazdelitve, je med njima malo verjetnostnega prostora, zato se mednju uvrsti malo vrednosti x_i, torej sta si blizu. Nasprotno, če je $x_i \sim N(0,1)$ in sta omenjeni vrednosti blizu nič, potem je pri istem t med njima precej več verjetnostnega prostora, zato sta si daleč.

Opozoriti moramo na to, da je d_i sedaj precej odvisna od gostote f_{x_i}, zato bi lahko novi predpis načeloma privilegiral nekatere porazdelitve. Na srečo velja naslednji izrek.

Izrek 1. *Naj bosta X in Y neodvisni enako porazdeljeni zvezni slučajni spremenljivki z gostoto f. Potem je porazdelitev spremenljivke $Z := d_f(X,Y) = |\int_X^Y f(t)dt|$ neodvisna od gostote f.*

Skrb o pristranskosti novega predpisa je torej odveč, a moramo rešiti še dve težavi. Ni namreč očitno, kako oceniti gostoto f, če je ne poznamo, in kako učinkovito izračunati integral (3), če se kumulativna porazdelitvena funkcija ne izraža z elementarnimi, kot je to npr. v primeru normalnih gostot.

Ker ocenjujemo koristnost značilk algoritmično, se moramo omejiti na parametrično družino gostot. Dober kompromis med ugodnimi računskimi lastnostmi normalne porazdelitve in številom prostih parametrov, s pomočjo katerih se lahko približamo gostoti, katere približek iščemo, so *Gaussove mešanice*, ki so oblike

$$f = \sum_{i=1}^{k} \alpha_i N(\mu_i, \sigma_i),$$

kjer so *jedrne funkcije* $N(\mu_i,\sigma_i)$ normalne gostote, *apriorne verjetnosti* $\alpha_i > 0$ pa se seštejejo v ena. Parameter k imenujemo število jeder. Generiranje slučajnih spremenljivk v skladu z f izvajamo v dveh korakih. Najprej sorazmerno s števili α_i naključno izberemo jedro i_0, ki mu bo realizacija pripadala, nato pa generiramo število v skladu s porazdelitvijo $N(\mu_{i_0},\sigma_{i_0})$.

Četudi predpostavimo, da je k poznan, preostalih parametrov gostote f še vedno ne moremo izračunati, denimo, po metodi največjega verjetja, zato postopamo iterativno. Popularen pristop je EM-algoritem (iz ang. *expectation maximization*), ki ga v povsem splošni obliki opiše Dempster s sod. [3], tu pa podajmo le psevdokodo za naš konkretni primer. Predstavimo ga v algoritmu 2.

Algoritem 2 EM-algoritem za seznam $\mathbf{x}$ vrednosti značilke x_i

1: $N \leftarrow$ dolžina seznama $\mathbf{x}$
2: $\mathbf{a},\mathbf{m},\mathbf{s} \leftarrow$ ničelni vektorji dolžine k
3: $\sigma \leftarrow$ naključno pozitivno število
4: **for** $i = 1,2,\ldots,k$ **do**
5: $\quad \mathbf{a}_i \leftarrow 1/k$
6: $\quad \mathbf{m}_i \leftarrow$ naključno število
7: $\quad \mathbf{s}_i \leftarrow \sigma$
8: **end for**
9: **while** zaustavitveni kriterij ni izpolnjen **do**
10: $\quad$ **for** $i_0 = 1,2,\ldots,k$ **do**
11: $\quad\quad$ **for** $j = 1,2,\ldots N$ **do**
12: $\quad\quad\quad p_{i_0,j} \leftarrow \frac{P(\mathbf{x}_j|y_{i_0})\mathbf{a}_{i_0}}{\sum_i P(\mathbf{x}_j|y_i)P(y_i)}$
13: $\quad\quad$ **end for**
14: $\quad\quad \mathbf{m}_{i_0} \leftarrow \frac{\sum_j p_{i_0,j}\mathbf{x}_j}{\sum_j p_{i_0,j}}$
15: $\quad\quad \mathbf{s}_{i_0} \leftarrow \frac{\sum_j p_{i_0,j}(\mathbf{x}_j - \mathbf{m}_{i_0})^2}{\sum_j p_{i_0,j}}$
16: $\quad\quad \mathbf{a}_{i_0} \leftarrow \frac{1}{N}\sum_j p_{i_0,j}$
17: $\quad$ **end for**
18: **end while**
19: **vrni** $(\mathbf{a},\mathbf{m},\mathbf{s})$

Na kratko opišimo, kako poteka. Vektorji $\mathbf{a}$, $\mathbf{m}$ in $\mathbf{s}$ so po vrsti ocene za α_i, μ_i in σ_i. Po inicializaciji v zanki **while** ponavljamo naslednji postopek. V vrstici 12 po Bayesovem pravilu izračunamo verjetnosti, da j-ta vrednost vektorja $\mathbf{x}$ pripada jedru i_0. Nato izračunamo nove ocene za povprečja in standardne odklone podobno kot pri metodi največjega verjetja, le da tu pri izračunu μ_0 in σ_0 v vrsticah 14 in 15 upoštevamo zgolj tiste deleže vrednosti $\mathbf{x}_j$, ki pripadajo jedru i_0.

Wu [6] opiše nekatere konvergenčne lastnosti. Če s θ_j označimo vektor neznanih parametrov na j-ti ponovitvi iteracije in z $\ell(\theta_j)$ logaritem verjetja, se izkaže, da je zaporedje $\{\ell(\theta_j)\}_j$ naraščajoče. Omejenost (in s tem konvergenco) si zagotovimo, če iz dopustnega območja parametrov

$$\Theta = ((0,1) \times \mathbb{R} \times (0,\infty))^k$$

izvzamemo majhne ε-okolice točk, kjer je kakšno izmed povprečij μ_i enako eni izmed vrednosti, ki jo značilka, katere porazdelitev ocenjujemo, zavzame na podatkovni množici S, hkrati pa velja $\sigma_i \searrow 0$.

Vemo že, kako najti približek za gostoto f s pomočjo Gaussovih mešanic, opisati pa moramo še, kako učinkovito izračunati njihove integrale, ki se pojavijo v (3). Če najdemo dober približek za *funkcijo napake*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt,$$

Tabela 1: Tabela koeficientov za približek funkcije erf

k	1	2	3	4	5
a_k	0,25482	−0,28449	1,42141	−1,45315	1,06140

Tabela 2: Opis podatkovnih množic

ime	diskretne	zvezne	primerki
contraceptive	4	5	1473
diabetes	0	8	768
hepatitis	13	6	155
hypo	18	7	3163

potem bomo znali natančno in učinkovito izračunati tudi kumulativno porazdelitveno funkcijo

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$$

in zato tudi $d_i(\mathbf{r}, \mathbf{s})$. Za $x \geq 0$ lahko vzamemo približek oblike

$$\text{erf}(x) = 1 - e^{-x^2} \sum_{k=1}^{5} \frac{a_k}{(1+px)^k} + \varepsilon(x),$$

pri čemer je $p = 0{,}32759$, koeficienti a_k pa so podani tabeli 1 [2]. Na negativne x ga razširimo prek lihosti funkcije napake. Za odstopanje približka velja $\varepsilon \leq 1{,}8 \cdot 10^{-7}$. Približek je torej hitro izračunljiv in natančen.

Določeni integral Gaussove mešanice, ki je konveksna kombinacija normalnih gostot f_i, dobimo po formuli

$$\int_a^b \sum_{i=1}^{k} \alpha_i f_i(t) dt \approx \sum_{i=1}^{k} \left(\text{erf}\left(\frac{b-\mu_i}{\sqrt{2}\sigma_i}\right) - \text{erf}\left(\frac{a-\mu_i}{\sqrt{2}\sigma_i}\right)\right).$$

Nova različica algoritma ReliefF je precej podobna prejšnji, razliki sta le dve. Preden se lotimo ocenjevanja značilk, moramo v novi različici oceniti vse gostote zveznih značilk. Pri dani značilki x_i poženemo EM-algoritem najprej pri številu jeder $k = 1$, nato pa ga povečujemo toliko časa, dokler si nista zaporedna približka gostot, ki ju dobimo, dovolj podobna ali pa ne pridemo do maksimalnega števila jeder K, ki si ga izberemo sami. Za merjenje podobnosti porazdelitev smo med možnimi kriteriji izbrali razmerje verjetij in se ustavili, ko je bilo l_k/l_{k+1} dovolj blizu ena.

Druga razlika glede na dosedajšnjo različico je ta, da v nadaljevanju pri zveznih značilkah uporabljamo nov predpis (3) za razdaljo.

4. VREDNOTENJE

Različici algoritma ReliefF primerjamo na štirih množicah z repozitorija UCI [1]. Njihov opis najdemo v tabeli 2. Ogledali si bomo, ali nova metoda vrača boljša rangiranja kot dosedajšnja.

Za vrednotenje vsake od obeh različic algoritma ReliefF izberemo metodologijo, ki jo je razvil Slavkov [5]. Na kratko jo opišimo.

Z nekim algoritmom dobimo rangiranje značilk

$$\mathbf{r} = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}),$$

kjer so značilke urejene padajoče po koristnosti. Definiramo množice X_j, $1 \leq j \leq n$, v katerih je j najkoristnejših značilk. Za vsak j zgradimo napovedni model za našo podatkovno množico S, pri čemer pa kot vhodne spremenljivke vzamemo zgolj tiste iz množice X_j. Natančnost napovedovanja našega modela označimo z α_j. Sedaj lahko narišemo *krivuljo dodajanja*, tj. graf natančnosti napovedovanja v odvisnosti od števila najkoristnejših značilk.

Podobno postopamo pri *krivuljah odvzemanja*, le da tu pri napovedovanju najprej upoštevamo vse značilke, na naslednjih korakih pa tiste iz komplementa množice X_j. Tako dobimo števila $\bar{\alpha}_j$.

Novo in staro različico algoritma ReliefF po vrsti označimo z M_1 in M_2. Ko ju preizkušamo, dobimo para krivulj odvzemanja in dodajanja. Iz dobljenih natančnosti bi radi izluščili številsko vrednost, ki pove, katera metoda vrača boljša rangiranja. Označimo razlike $\delta_j = \alpha_j^1 - \alpha_j^2$ ter $\bar{\delta}_j = \bar{\alpha}_j^1 - \bar{\alpha}_j^2$. Definiramo

$$\delta(M_1, M_2) = \frac{1}{2}\left(\sum_j \frac{w_j}{w}\delta_j - \sum_j \frac{w_{n-k+1}}{w}\bar{\delta}_j\right), \qquad (4)$$

kjer je $w_j = 1/j$ in $w = \sum_j w_j$. Označimo vsoti v (4) po vrsti z Δ in $\bar{\Delta}$. Z utežmi damo večji pomen razlikam, ki se pojavijo, ko je značilk, s katerimi zgradimo napovedni model, malo. Težimo k temu, da se krivulja dodajanja na začetku čim bolj strmo vzpne (tj. na začetku rangiranja so najbolj koristne značilke) in da natančnost pri krivuljah odvzemanja čim hitreje pade (tj. pri repu rangiranja so najmanj koristne značilke).

Pri preizkušanju smo v algoritmu ReliefF izbrali vrednosti $k = 10$ in $m = |S|$ ter pri novi različici dodatno še največje dovoljeno število jeder $K = 7$.

5. REZULTATI

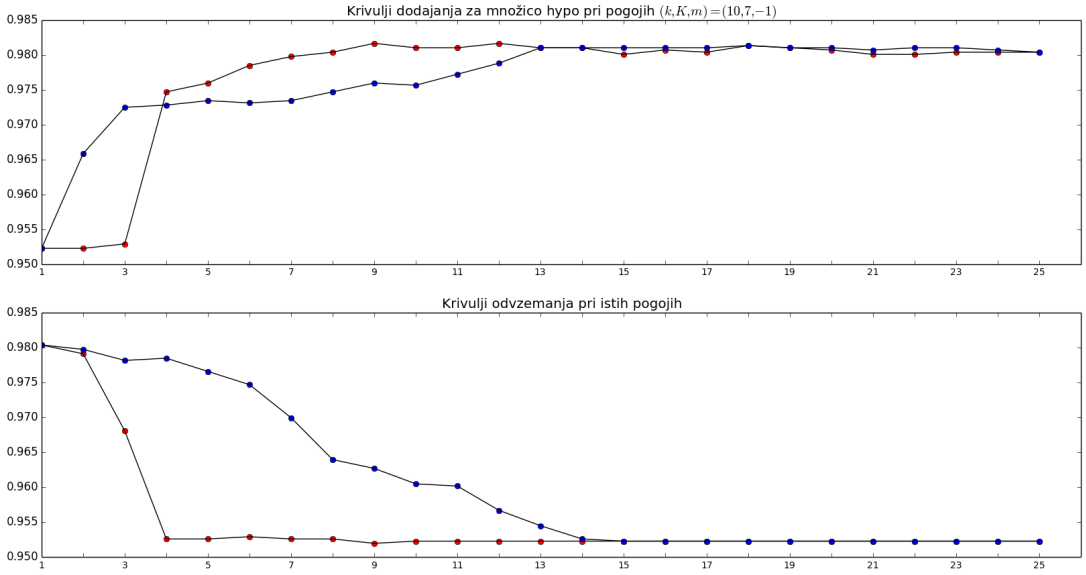
V tabeli 3 so zbrani povprečni časi trajanj obeh algoritmov. Vsako različico smo na računalniku s 3,3-gigaherčnim Intelovim procesorjem `i3-2120` pognali dvajsetkrat. Pri starejši različici merimo zgolj skupni čas, pri novejši pa smo posebej merili še čas, ki ga potrebujemo za ocenjevanje porazdelitev.

Razmerja časov za ocenjevanje značilk so med 10 in 15. To je pričakovano, saj je povprečno število jeder med 2 in 3, izračun razdalje (3) pa je pri enem jedru približno petkrat dražji kot (2). Razmerje skupnih časov gre seveda še nekoliko bolj v korist stare različice.

V tabeli 4 prikažemo še uspešnosti, ko izračunamo δ, Δ in $\bar{\Delta}$ po formuli (4). Sledi, da so za našo metodo dobre pozitivne vrednosti in večina je takih. Morda se zdi, da sta si metodi na množici `hypo` zelo podobni, saj so utežene razlike, ki jih dobimo, precej majne, vendar je z obeh grafov na sliki 1 jasno razvidno, da naša metoda vrne boljše urejanje, le da najkoristnejših dveh značilk ne uvrsti na sam vrh seznama, kar je razvidno iz krivulj dodajanja. Do podobnih sklepov bi lahko prišli pri preostalih množicah.

Tabela 3: Povprečni časi ene izvedbe algoritma v sekundah

ime	t_1^{skupni}	t_2^{EM}	t_2^{brezEM}	t_2^{skupni}	$t_2^{\text{brezEM}}/t_1^{\text{skupni}}$	$t_2^{\text{skupni}}/t_1^{\text{skupni}}$
contraceptive	0,0129	0,3347	0,1177	0,4525	9,0	34,9
diabetes	0,0078	0,6039	0,1231	0,727	15,7	93,2
hepatitis	0,003	0,0767	0,0243	0,1011	8,1	33,6
hypo	0,0547	1,1784	0,4169	1,5953	7,6	29,1



Slika 1: Uspešnost napovedovanja za dosedajšnjo (modra) in predlagano različico (rdeča) algoritma ReliefF

Tabela 4: Razlike uspešnosti napovedovanja v odstotkih

ime	Δ	$-\Delta$	δ
contraceptive	0.43	0.46	0.45
diabetes	0.34	0.01	0.18
hepatitis	0.59	−0.59	0.00
hypo	−0.22	0.20	−0.01

6. ZAKLJUČKI IN NADALJNJE DELO

Spoznali smo, kako lahko s pomočjo ocenjevanja porazdelitev z jedrnimi funkcijami pridemo do spremenjene različice algoritma ReleifF, ki vrača boljša urejanja kot dosedanja.

V prihodnosti bi jo lahko še dodelali, tako da bi se ocenjevanja porazdelitev lotili na kak hitrejši način. Ena možnost za to je gotovo empirična porazdelitvena funkcija, morda celo v kombinaciji s pragovno funkcijo, ki sta jo vpeljala Kononenko in Robnik-Šikonja v [4].

Prav tako je mogoča tudi povsem drugačna posplošitev trenutnega ReliefF-a, ki bi delovala na strukturiranih podatkih, ki niso v tabelarični obliki, kot smo privzeli na začetku. To bi bilo zanimivo, ker je takih podatkovij vedno več.

7. ZAHVALA

Zahvaljujemo se Ivici Slavkovu, ki je dal na razpolago svojo implementacijo postopkov, ki jih je razvil za primerjavo metod za rangiranje značilk.

8. VIRI

[1] UCI machine learning repository. `https://archive.ics.uci.edu/ml/datasets.html`. Ogledano: 2015-09-15.

[2] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dower Publications, Inc., 1972.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[4] I. Kononenko and M. Robnik-Šikonja. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning Journal*, 53:23–69, 2003.

[5] I. Slavkov. *An Evaluation Method for Feature Rankings*. PhD thesis, Mednarodna podiplomska šola Jožefa Stefana, Ljubljana, 2012.

[6] C. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11:95–103, 1983.

Analyzing the formation of high tropospheric Ozone during the 2015 heatwaves in Ljubljana with data mining

Johanna A. Robinson[1,2] johanna.robinson@ijs.si
Sašo Džeroski[3,1] saso.dzeroski@ijs.si
David Kocman[2] david.kocman@ijs.si
Milena Horvat[2] milena.horvat@ijs.si

ABSTRACT

This paper presents a data mining approach to analyzing how environmental parameters, measured by electrochemical gas sensors, influence the formation of ozone. Ozone is a hazardous gaseous air pollutant whose formation is accelerated at high temperatures. The study period includes the four heatwaves experienced in the summer of 2015 in Ljubljana, Slovenia. A temperature threshold, which could result in above-limit ozone levels, was determined with the help of decision trees. It can be used to warn the public about exceeding the allowed ozone levels. In different types of decision trees, we found the same temperature threshold of 34,7°C, which resulted in the highest ozone levels (that did not exceed the limit value).

1. INTRODUCTION

While naturally occurring stratospheric ozone (O_3) protects us from the harmful ultra violet (UV) radiation from the Sun, the tropospheric, i.e., ground level, ozone is detrimental for human health. Ozone is formed from other pollutants through a complex set of photochemical reactions that need UV light. Common precursor gasses are, e.g., oxides of nitrogen (NO_x), carbon monoxide (CO) and volatile organic compounds (VOCs) [9].

Ozone is a major public concern because of its adverse impacts on human health. It causes a variety of health problems, varying from triggering asthma attacks through decreased lung function to even death. It is especially harmful for children, elderly and people with allergies and lung diseases [14,4]. Ozone formation increases during warm sunny weather [6]. Mortality is higher during heatwaves, with some incidents caused by high ozone levels [13].

Slovenia experienced four heatwaves during the summer of 2015. The Slovenian environment agency ARSO [10,11,12] reported the periods of the heatwaves as follows (1) 2-14.6.2015, (2) 1-8.7.2015, (3) 11-26.7.2015, (4) 4-15.8.2015. During such episodes exercising is recommended to be kept to the minimum.

In order to protect the public, The World Health Organization [14] has created guidelines to pollutants, which are also adapted internationally, e.g., by the European Union [1,3]. The newest guideline for ozone is set to 100 µg/m^3 for daily maximum 8-hour mean, the old 120 µg/m^3 being still valid in the European Union for 8 hour mean. There is never a safe level of any air pollutant, and possible health effects might still occur in some individuals, even at pollutant concentrations below limit values.

The guidelines are set to provide adequate protection of public health, taking into consideration also other living beings and the possible detrimental effect on historic heritage and bearing in mind the economic and technical feasibility of following the guidelines [1].

2. AIMS AND HYPOTHESIS

The four heatwaves experienced in Slovenia during the summer of 2015 provided us with a basis for studying the photochemical processes of ozone formation. Namely, ozone formation is accelerated at higher temperatures. We studied the effect and relation of environmental parameters on/to the formation of ozone by using data mining tools.

The limit value for ozone is set in the Directive 2002/3/EC [2] to 120 µg/m^3 for the 8 hourly maximum, which is allowed to be exceeded at most 25 times a year. The hourly average threshold level, for which the public need to be informed immediately after exceedance, is set to 180 µg/m^3, whereas the alert threshold is at 240 µg/m^3. Since higher ozone levels occur during higher temperatures, we wanted to find a temperature threshold for such high ozone values, which would exceed the hourly limit values. Such a threshold could help to give warnings to the general public before the occurrence of a high ozone episode even sets in. Usually the warnings are given together with a heatwave warning, whereas incidents could happen also on individual days or in places which are not covered by the national monitoring programme. That's why finding an indicative threshold for a parameter, such as temperature, which every household can easily read at home, would provide the general public with a good starting point to know when to avoid spending excess time outside and to restrict the amount of exercise performed during hours with high ozone levels. This could be especially important for those individuals who belong to the susceptible groups.

3. DATA AND METHODS

The data for analyzing ozone formation was obtained from a network of low cost electrochemical gas sensor units deployed in Ljubljana since the winter of 2013/2014. The sensors belong to the CITI-SENSE project (www.citi-sense.eu), currently testing the performance and applicability of low cost electrochemical gas sensors.

The period chosen for the data analysis covers the period from the first heatwave until the last, fourth heatwave, i.e., from 1.6.2015 to 19.8.2015, as illustrated in Figure 1. The numerical data was downloaded from an online server as a csv-file, after which it was

[1] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana

[2] Jožef Stefan Institute, Department of Environmental Sciences Jamova 39, 1000 Ljubljana

[3] Jožef Stefan Institute, Department of Knowledge Technologies Jamova 39, 1000 Ljubljana

pre-processed to meet the requirements of the chosen open source software data mining package, WEKA [5].

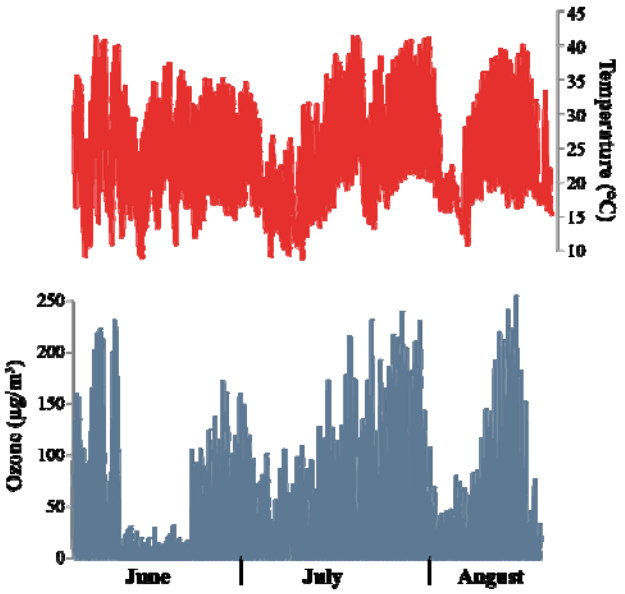


Figure 1. Measured temperature and ozone in Ljubljana during the period between 1.6.2015-19.8.2015.

3.1 Pre-processing of the Data

During the pre-processing, the data underwent screening of their suitability for analysis. Hourly averages were calculated as a new attribute, in order to be comparable with the European legislations limit values based on averaged hourly data. Also, due to technical limitations of electrochemical gas sensors, all data which were below the limit of detection (LOD) (9,99 µg/m^3) were excluded. The final number of data points, each corresponding to an one-hour of measurements, was 756. The data used for the analysis and its specifications are summarized in Table 1. All used values were numeric.

Table 1. The data: Basic statistics.

Parameter/Attribute	Missing	Min	Max	Mean
Ozone, O3 (µg/m^3)	0%	10.1	236	60.9
Temperature (°C)	0%	9.30	41.2	26.6
Humidity (%)	0%	26.8	91.1	55.3
Air pressure (mb)	0%	971	990	980
Nitrogen monoxide, NO (µg/m^3)	83%	0.12	27.0	6.32
Nitrogen dioxide, NO_2 (µg/m^3)	57%	0.20	166	38.7
Nitrogen oxides, NO_x (µg/m^3)	86%	0.01	165	27.0
Carbon monoxide, CO (µg/m^3)	9.0%	0.24	376	80.8

3.2 Choosing a Data Mining Method

An initial test to determine whether or not the ozone data is linearly dependent on temperature was run using LeastMedSq. However, according to the results, we believe that the relation is not linear. To proceed further to see which other parameters/ attributes play a role, we ranked the attributes with the RreliefFAttributeEval attribute evaluator within WEKA [5] This implements the instance-based RReliefF method [8] for estimating the relevance of attributes/ features and feature ranking. We used 10-fold cross validation to calculate the relevance scores of features and their variance. The parameters were ranked as shown in Table 2. NO (forming most of NO_x) and CO are the two common precursor gases forming ozone in the troposphere and this explains their high ranking. NO_2 on the other hand commonly counteracts this, thus reducing the amount of ozone: This probably explains why it ranks low, together with the fact that many data values for these gasses are missing (57%).

Table 2. Ranking the attributes with RReliefF in terms of their relevance for predicting the ozone concentration.

Rank	Parameter
1.	NO
2.	NO_x
3.	CO
4.	Humidity
5.	Temperature
6.	Air pressure
7.	NO_2

Two types of decision trees were considered: (1) model trees and (2) regression trees. For the purpose of building such trees we used the M5P algorithm, a Java reimplementation of the algorithm M5 [7] within the WEKA [5] package of data mining software. The availability and distribution of the environmental attributes (other than ozone) vary. The distribution of the values of all attributes is illustrated in Figure 2.

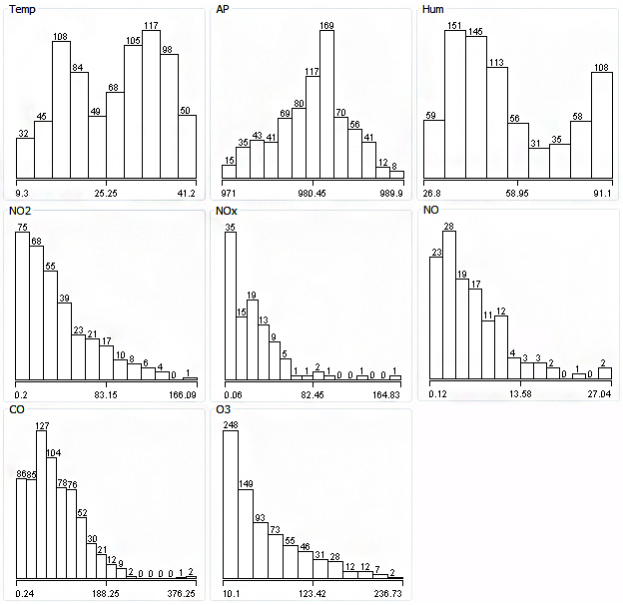


Figure 2. The distribution of the attribute values in the data.

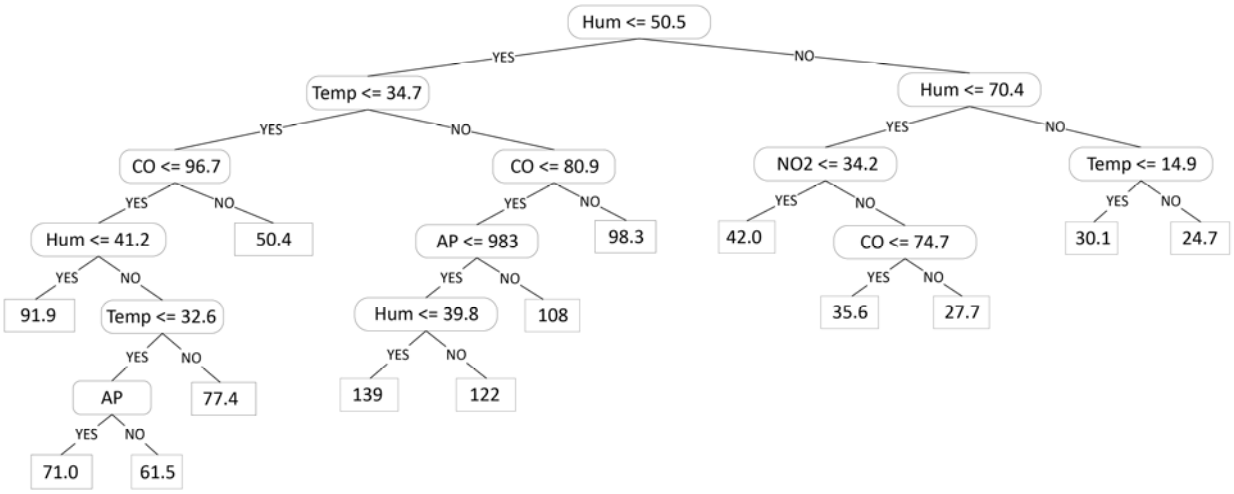


Figure 3. Pruned regression tree with all parameters.

4. RESULTS AND DISCUSSION

The initial runs with WEKA with all the available parameters are given below. The default settings of M5P were used unless otherwise stated. Both regression tree and model trees were created. The initial regression tree structure was more complex having 20 rules compared to the model tree with only 5 nodes.

M5P selected humidity as the most important parameter for prediction in both decision trees (top). This position of humidity was expected, since the electrochemical gas sensors highly depend on the humidity, given the way the sensors themselves function by absorbing or evaporating humidity in the electrolyte inside the sensors. The highest ozone levels predicted were 139 µg/m^3, 122 µg/m^3 and 108 µg/m^3, and did not exceed the limit values set by EU regulations (180 µg/m^3). According to the regression tree predictions, we get to these high ozone levels if humidity is less than or equal to 50,5%, the temperature is more than 34,7°C and carbon monoxide is less or equal to 80,9 µg/m^3. We get to the two highest predicted ozone levels if the air pressure is less than or equal to 983 mb: if the humidity is less than or equal to 39,8% the predicted ozone is 139 µg/m3, and if higher than 39,8% the ozone is 122 µg/m^3. The third highest ozone level (108 µg/m^3) is reached if the air pressure is higher than 983 mb. The model tree follows the same rules, but has fewer nodes.

As the initial regression tree was rather large, we pruned it further by setting the minimum number of instances to 80 (initially 4). This results in a tree with 14 nodes, which follows the same rules to reach the same three highest predicted ozone levels, now in nodes 6, 7 and 8 (Figure 3). The correlation coefficient is still high at 0.801 having pruned the tree to include more instances in the leaves.

Since a large number of data points were missing for all oxides of nitrogen (Table 2), we excluded them from the next runs. They were also not playing a role in the decision trees in reaching the high ozone values. Considering that carbon monoxide is one of the pre-cursor gases, and it appears in the decision trees we decided to keep it in the next data runs even if it was missing 9% of its values.

The rules leading to the prediction of high ozone values still stay the same. E.g., in the case of the highest predicted ozone level in the second pruned regression tree: IF hum <=50.5 AND IF Temp >34.7 AND IF CO <= 80.9 AND IF AP <= 983 AND IF Hum <= 39.75 THEN O3= 139.

When also carbon monoxide, which was missing some of the data points, was excluded from the run, we get a correlation coefficient slightly lower than 0.8 both for the regression tree and the model tree.

As the original research question was to find a threshold value in temperature, which would determine high ozone values, by observing the visualized regression trees and model trees we found a threshold temperature at 34,7°C which leads to higher ozone values. The threshold stayed the same in all the decision trees, i.e., model trees as well as regression trees. In addition, the pruned regression tree produced an extra threshold at 37,7°C with highest predicted ozone level at node 5 (133 µg/m^3). Table 3 summarizes all the 3 runs, showing the used parameters as well as the size and accuracy of the obtained decision trees.

Table 3. Summary of the three runs with M5P, all using temperature (Temp), humidity (Hum) and air pressure (AP). MT, RT, and PT = model, regression and pruned regr. tree.

	Temp, Hum, AP, CO, NO, NO_2, NO_x			Temp, Hum, AP, CO			Temp, Hum, and AP only		
	MT	RT	PT	MT	RT	PT	MT	RT	PT
Corr. Coeff.	0.846	0.805	0.801	0.844	0.810	0.806	0.797	0.775	0.776
Num. of rules	5	20	14	8	21	13	7	11	9

5. CONCLUSIONS

Data from an experimental project using low cost electrochemical gas sensors was used to determine the influence of environmental parameters on ozone formation. As ozone is formed under photochemical reactions, we predicted that temperature would play the largest role in determining the ozone levels. The summer of 2015 provided a good database for studying this relation, since Slovenia experienced four heatwaves during the period between 1.6.2015 and 19.8.2015.

Although the early runs of linear regression did not produce a simple rule, which would have been easy to implement in warning the population of high ozone levels, directed us to look into the other environmental parameters e.g., the precursor gases, to find how ozone is formed. Since the limit values of ozone are given in an hourly manner, the original raw data (15 minute averages) was aggregated at the hourly level to be comparable with the national and international information and alert thresholds.

In order to see which parameters play the largest role, an instance-based feature selection algorithm was run. It produced a ranking where the highest ranks were reserved for the precursor gases, followed by humidity, and only then temperature. The M5P program was used to analyze the data by creating regression trees as well as model trees, helping us to find threshold values which determine high ozone concentrations. Initial runs were made with all the available environmental parameters (humidity, temperature, air pressure, NO, NO_2, NO_x and CO). The high importance of humidity is partially due to the technical features of how electrochemical gas sensors work. Many elements and processes in the atmosphere determine the formation of gases, e.g., the presence of pre-cursor gases as well as the intensity of UV-light. In order to keep a representative set of environmental parameters, we eliminated the ones which, even though they play a role in the atmosphere, were missing a substantial number of values (up to 86%) due to bad sensor performance in the chosen time period.

The trees performed with a correlation coefficient of 0.8 or higher in predicting ozone levels. Regardless which decision tree was used, we found a temperature threshold of 34,7°C. When the temperature is above 34,7°C, higher ozone levels can occur in the troposphere. However, no temperature threshold was found by the model, which would directly result in ozone levels above the limit value. We also found that the predicted ozone values were relatively low, considering that the background level for ozone is 70 $\mu g/m^3$ [14] and the mean ozone value in our dataset was 61 µg/m3. Information which can be extracted from parameters available for all households e.g., temperature, could be a good start for creating a smartphone application to warn people when ozone is expected to reach a certain level. The application could take data from a home thermometer placed outdoors whereas the threshold value for notification could be modified by the user, enabling it to trigger even at lower levels of ozone depending on the sensitivity of the user to ozone.

6. ACKNOWLEDGMENTS

The CITI-SENSE project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 308524. This work has also received funding of the Slovenian Research Agency through a programme P1-0143.

7. REFERENCES

[1] EC. *Council Directive 96/62/EC of 27 September 1996 on ambient air quality assessment and management*. Retrieved from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31996L0062.

[2] EC. *Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to Ozone in ambient air*. OJL 67, 3 March 2002. Retrieved from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32002L0003

[3] EC. *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*. Retrieved from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008L0050.

[4] European environment agency (EEA). 2012. *Air quality in Europe — 2012 report.* EEA Report No 4/2012, Copenhagen.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 1: 10–8.

[6] Lee, J.D., Lewis, A.C., Monks, P.S., et al. 2006. Ozone photochemistry and elevated isoprene during the UK heatwave of august 2003. *Atmos. Environ.* 40, (Dec. 2006), 7598–7613. DOI= doi:10.1016/j.atmosenv.2006.06.057.

[7] Quinlan, R.J. 1992. Learning with Continuous Classes. In: *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapore, 343-348.

[8] Robnik-Sikonja, M., and Kononenko, I. 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.* 53, (Oct. 2003), 23-69.

[9] Salvato, J., A., Nemerow, N., and Agardy, F., J. 2003. *Environmental engineering*. Fifth edition. John Wiley & Sons: New Jersey.

[10] Slovenian Environment Agency – ARSOa. 2015. *Vreme in podnebje - zanimivosti*. Available in www.arso.gov.si/vreme/zanimivosti/

[11] Slovenian Environment Agency – ARSOb. 2015. *Agrometeorologija - aktualni prispevki*. Available in www.arso.gov.si/vreme/agrometeorologija/aktualno.html

[12] Slovenian Environment Agency - ARSOc. 2015. *Arhiv novic za zadnjih 6 mesecev*. Available in www.arso.gov.si/o%20agenciji/novice/arhiv.html

[13] Williams, S., Nitschke, M., Weinstein, P., Pisaniello, D.L., Parton, K.A., and Bi, P. 2012. The impact of summer temperatures and heatwaves on mortality and morbidity in Perth, Australia 1994 –2008. *Environ. Int.* 40 (Apr 2012), 33–38. DOI= doi:10.1016/j.envint.2011.11.011.

[14] World Health Organization (WHO). 2006. *Air quality guidelines for particulate matter, Ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Summary of risk assessment*. WHO, Geneve.

Recommender System as a Service based on the Alternating Least Squares algorithm

Gašper Slapničar
Jožef Stefan Institute,
Department of Intelligent Systems
Jamova cesta 39
1000 Ljubljana
+386 51 721 041
slapnicar.gasper@gmail.com

Boštjan Kaluža, Mitja Luštrek
Jožef Stefan Institute, Department of Intelligent Systems
Jamova cesta 39
1000 Ljubljana
+386 1 477 3944
{bostjan.kaluza, mitja.lustrek}@ijs.si

Zoran Bosnić
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113
1000 Ljubljana
+386 1 479 8237
zoran.bosnic@fri.uni-lj.si

ABSTRACT

In this paper, we describe a production-ready recommender system as a service for recommending eco-friendly tourist accommodations. It offers two main features: (1) it returns personalized recommendations for a user by creating a latent factor model through matrix factorization (Alternating Least Squares algorithm, ALS) and (2) it returns accommodations that are similar to a given accommodation by calculating content-based similarity using the Jaccard coefficient and the Euclidian distance. The system is evaluated on the collected data by using cross-validation and Precision@k as a performance measure. It achieves 19% Precision@k for personalized recommendations to a user based on his past interactions with accommodations. This score far surpasses a random recommender implementation that achieves 1% Precision@k.

Keywords

Recommender system, parallel computing, machine learning, big data, matrix factorization

1. INTRODUCTION

Due to significant growth and evolution of e-Commerce and digitalization in many fields (medicine, economics, etc.) in the past years, the size and complexity of collected data is growing rapidly. Data are being generated in real time and are mostly unstructured. Such data collections are thus being referred to as the "big data".

Alongside the development of big data technologies, which can successfully store and process large amounts of unstructured data in real time, companies want to use these technologies in machine learning and offer machine learning algorithms to users in a simple way, as a service. This led to recent development of many Machine Learning as a Service (MLaaS) providers.

MLaaS platforms can be used to develop recommender systems, which are an essential part of e-Commerce and can bring a significant competitive advantage. These systems focus on creating personalized recommendations which should include items that will most likely interest a potential customer. Since e-Commerce and data science are evolving rapidly, experts from different fields are required for development of a production grade recommender system. This can be addressed by developing the system in a cloud and thus ensuring separation of concerns.

We address the problem of developing such a recommender system by dividing in into two parts. The first part is data collection. We record the user-item interactions in the event-based style, where each interaction corresponds to an action of the user on the web portal and is an element of a predefined list. User is represented by a unique tracking id while items correspond to a list of unique accommodation ids.

The second part and core of the problem are the recommendation algorithms. For customized recommendations to a user based on past actions, we create a latent factor model using matrix factorization. The model learns by using Alternating Least Squares algorithm. We propose an algorithm that can be efficiently executed in parallel and is a good candidate to use with distributed big data technologies. For recommending similar accommodations based on their properties we use the Jaccard coefficient and normalized Euclidian distance merged together into a common similarity measure.

2. BACKGROUND

A large number of MLaaS platforms emerged recently. These support a wide variety of machine learning algorithms out of the box and can be used to develop a recommender system. In the following subsections we overview considered platforms, and describe the outline of our proposed approach.

2.1 MLaaS PLATFORM

First, we compared several MLaaS platforms, such as BigML, Google Prediction API, Azure ML, Amazon ML and Prediction.IO. Due to white-box design and open source access we chose Prediction.IO machine learning server. It is implemented as a distributed scalable stack based on latest big data technologies, such as Apache HBase, Apache Hadoop, Apache Spark and Elasticsearch [3].

Since it implements Apache Spark's MLlib, it comes with native support for many algorithms that are suitable for development of a recommender system. It also offers *templates* which are implementations of some machine learning algorithms on actual problem domains and are available at templates.prediction.io/ [3].

Prediction.IO machine learning server consists of three main parts [3]:

1. **Prediction.IO platform** – open source machine learning stack for creating ***engines*** with machine learning algorithms,
2. **Event Server** – API for collecting events from different sources and unifying the format,
3. **Template Gallery** – templates with implementations of machine learning algorithms on real problem domains.

Figure 1 shows how clients interact with the machine learning server and it also shows that the server can have several engines, each corresponding to one machine learning application.

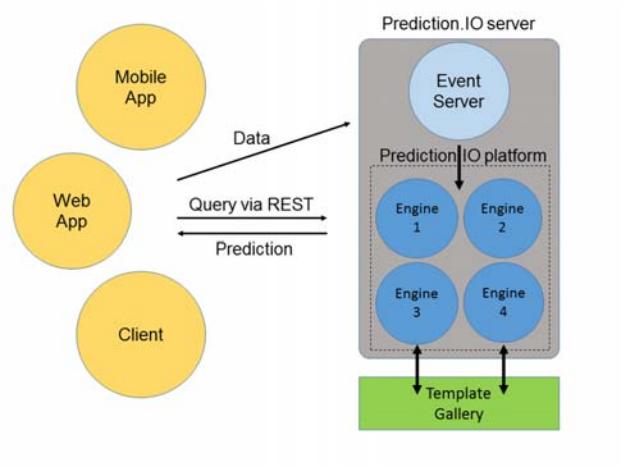


Figure 1: Conceptual architecture and interaction between Prediction.IO server and client.

MVC (model – view – controller) architectural pattern is well established in web development for years, since it helped to speed up the development process and lowered the learning curve.

It would make sense to implement such a pattern in data science, especially in machine learning. Prediction.IO represents one of the first attempts of this with their DASE engine architecture. Each engine follows this architectural pattern and must contain all the components with the exception of Evaluation, which is optional:

- **D**ata Source/**D**ata Preparator – reads data from an input source (database) and transforms it into a desired format,
- **A**lgorithm – the machine learning algorithm used to create the predictive model,
- **S**erving – takes client queries and returns prediction results,
- **E**valuation – quantifies prediction success with a numerical score.

2.2 ALGORITHMS

The "Netflix prize" competition has demonstrated that within collaborative filtering, latent factor models are the most successful method for recommending products. This method is gaining popularity, due to good scalability and better precision in comparison with neighborhood methods [1, 2].

Latent factor models are most successfully implemented by using matrix factorization. Two of the best known algorithms for solving the matrix factorization problem are Stochastic Gradient Descent (SGD) and Alternating Least Squares (ALS). A lot of effort has been put towards parallelization of the SGD algorithm. This has proven to be a difficult problem. ALS is an algorithm that is closely related to SGD and offers high level of potential parallelization [2].

3. DATA AND RECOMMENDATION TASK

The data are collected from a website, which offers eco-friendly accommodations. Any accommodation which meets 5 out of 10 required criteria is considered eco-friendly. Examples of these criteria are re-usage of water, usage of solar energy, waste recycling, etc.

Training data are being collected in real time in event-based style. This means that each action that a user does on the web portal, corresponds to a single user-item interaction.

User – action – item format was chosen to describe interactions between users and accommodations as it comprises all the required information. Each event on the web portal can be represented with this format. *User* corresponds to a unique user performing actions on the web portal and is traced by using a long lasting cookie id. *Action* is whatever this user does on the web portal and it corresponds to the user-item interaction. It is an element of a predefined list containing all relevant possible actions on the web portal. *Item* corresponds to unique identifier of the accessed object and is an element of a predefined set of all existing accommodation ids.

Example of this format is given as: user *U1 views* accommodation *I1* (*U1 – view – I1*).

A custom API was developed in order to implement basic authentication and authorization together with data sanity check. This API connects to the Event Server of Prediction.IO machine learning server and saves the collected data to HBase data store in real time. Each event corresponds to an HTTP POST request which contains parameters corresponding to *user – action – item* format. It also contains the timestamp of the event.

Different feedback mechanisms can be used to record user-item interactions. Recommender systems work best with data collected through explicit feedback mechanism such as ratings. Due to the implementation of the web portal, only implicit feedback mechanism is available (e.g. *views*, *inquiries*). A mapping was implemented to map the collected implicit data to explicit numerical ratings on the scale of 2 to 5. This is explained further in the following section.

Due to the design of the web portal, a user can either do a complete search from the landing page or he can access the page with accommodation details directly from a web search engine. In first case the web portal design allows us to implement recommendations for this specific user that can be displayed with the search results. In second case we are limited by design to recommend similar accommodations to the currently viewed accommodation. This also makes more sense since this type of access typically occurs for users without any past interactions with the web portal. Thus we develop two distinct recommending functionalities.

4. RECOMMENDATION OF ACCOMMODATIONS FOR A USER

First broad recommendation approach is known as *collaborative filtering* and it relies only on past user-item interactions. Based on these past actions it then finds similar users to other users. This approach usually produces better results but suffers from the *cold start* problem. This means that for a new user without past interactions (no history), the system will not be able to produce any meaningful recommendations. An important advantage is that the system can obtain the required data by observing and recording user history [2, 6].

To recommend accommodations to a given user, we used latent factor models approach, which is a method within collaborative filtering group.

Latent factor models try to explain the past ratings by characterizing users and items on hidden variables called factors, which are inferred from these past ratings patterns. The factors measure dimensions that are not obvious or easily explained. For users, each factor measures how much the user likes the item which scored high or low on the corresponding factor [2].

This method requires explicit preference values which we defined through the following mappings:

- view an accommodation ⇒ value 2.0,
- open inquiry window ⇒ value 4.0,
- send an inquiry ⇒ value 5.0,
- close inquiry window without sending ⇒ value 4.0.

The last action is important as an anomaly can happen where a user can access the inquiry window directly from a link and can therefore bypass other preceding actions.

In the case where a user does more than one action on the same accommodation, the highest preference value is kept.

4.1 Matrix factorization model

Latent factor models are most commonly created using matrix factorization. First a ratings matrix *R* of dimensions *m x n* is created from collected data. This means that the rows of matrix R represent *m* users and the columns represent *n* items. A specific value R_{ij} represents the rating given by user *i* to item *j*. Matrix *R* is then factorized into smaller matrices *U* and *V*. Matrix U represents *users* and is of dimensions *m x rank*, while matrix *V* represents *items* and is of dimensions *rank x n*. *Rank* is the parameter that tells us how many latent factors we want to use to describe a *user* or an *item* and is always much smaller than the dimensions of the original matrix which is very large (*rank* << *m, n*) [2].

Some elements of matrix *R* are not defined, meaning they cannot be interpreted as 0. This means that decomposition methods such as SVD (singular value decomposition) cannot be used.

4.2 Alternating Least Squares algorithm

Alternating Least Squares algorithm is used to modify latent factors and learn the model. It first initializes matrix *V* with small random values. When *V* is fixed, it iterates through matrix *U* and modifies the latent factors to best correspond to known ratings by minimizing the error. When this is done, it fixes *U* and does the same for *V*. It alternates doing this, solving the least squares problem defined by Equation (1) [2].

$$\min_{q,p} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\left|\left|q_i\right|\right|^2 + ||p_u||^2) \qquad \textbf{(1)}$$

In Equation 1, *K* is the set of *(u,i)* pairs for which rating r_{ui} is known. $(r_{ui} - q_i^T p_u)$ represents the error between known and predicted rating where q_i^T is the vector of latent factors for an item and p_u is the vector of latent factors for a user. $\lambda(\left|\left|q_i\right|\right|^2 + ||p_u||^2)$ represents the normalization term where λ is the normalization parameter which ensures that there is no data overfitting.

This algorithm is very suitable to be executed in parallel since many vectors of latent factors of *U* or *V* can be computed at the same time, considering one of the matrices is fixed. Potentially this allows the computation of all vectors of a single matrix in parallel.

The result of the algorithm is a model of latent factors which predicts preference values of any known user for any item.

5. SIMILAR ITEM RECOMMENDATION

Content-based filtering is the second broad recommendation approach and it focuses on creating a profile of each user or item to describe its nature. This profile usually contains the properties of an item. The system then finds similar items based on these profiles. This approach relies on obtaining data about items from an outside source, meaning it cannot obtain this data on its own [2, 6].

To recommend similar accommodations, an approach based on content-based filtering and the properties of accommodations was used. Based on our data we chose two similarity measures: Jaccard coefficient and Euclidian similarity.

5.1 Jaccard coefficient

Each accommodation has its corresponding attributes (e.g. pool, internet, bathroom, solar cells etc.). Each of these attributes is represented by an integer value and is either present or not. This was presented using a binary vector, where the index corresponds to the integer value of the attribute. Value 1 denotes presence of this attribute, while value 0 denotes its absence.

Since this data is binary, Jaccard coefficient is most suitable to measure attribute similarity between accommodations. It is defined by Equation (2), where $A \cap B$ represents the common attributes and $A \cup B$ represents the union of attributes of both accommodations:

$$Jaccard(A,B) = \frac{A \cap B}{A \cup B} \qquad \textbf{(2)}$$

5.2 Euclidian similarity

Each accommodation is also defined by its geographic position, consisting of longitude and latitude. Euclidian distance was used to first measure the distance between accommodations as shown by Equation (3).

$$Euclidian\_dist(a,b) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad \textbf{(3)}$$

A transformation of the distance into a similarity measure was then done by using Equation (4).

$$Euclidian\_sim = \frac{1}{1 + Euclidian\_dist(a,b)} \qquad \textbf{(4)}$$

Finally both similarity scores were merged into a unified similarity score. This was done by using Equation (5), where w_j and w_e correspond to weight parameters given to each similarity score:

$$Sim = \frac{w_j \cdot Jaccard(A,B) + w_e \cdot Euclidian\_sim}{w_j + w_e} \qquad \textbf{(5)}$$

6. EVALUATION

The system was evaluated using *Precision@k,* which is a standard measurement in information retrieval systems. Precision@k is defined by equation 6 [4].

$$Precision@k = \frac{Nr.of\ relevant\ among\ k\ recommended}{k}$$

In systems such as search engines and recommenders, the first *k* results are most important for their performance. It is highly important to show the *relevant* items among the first *k*. *Relevant* in the case of evaluation means the accommodations with which a user actually had an interaction in the past.

Recommender system was evaluated using *k-fold* cross validation, where the data is randomly split into *k* sets. The learning set always contains 80% of the data while the testing set contains 20%.

After the data was split, we chose evaluation parameters based on our problem. The following evaluation parameters were used:

- kFold = 5 – Chosen as a standard value.
- threshold = 2.0 – The threshold tells us which items are considered *relevant*. Threshold 2.0 means that any viewed or more strongly preferred is considered relevant.
- k = 3, 10 – Chosen as the user sees 3 accommodations on the site without scrolling and sees 10 accommodations on the first page of results.

The evaluation was executed for different parameters of algorithm:

- λ = 0.01 – Standard value which ensures that latent factor values do not overfit the learning data.
- numIterations = 5, 10 – Number of iterations during which latent factors are being modified.
- Rank = 5, 10, 20 – Number of latent factors which are used to describe user-item interactions.

Results are shown in Table 1.

rank	numIterations	Precision@3	Precision@10
5	5	1,74%	1,25%
5	10	5,38%	2,68%
10	5	8,90%	4,05%
10	10	17,66%	6,29%
20	5	**18,79%**	6,69%
20	10	16,69%	6,16%
Random recommender		0,97%	0,92%

Table 1: The evaluation results for different parameters of algorithm.

The results compare Precision@k of the developed recommender system with random recommender. The best result 19% was very superior to the 1% of random recommendation.

Better results are shown with low *k*. This is due to the fact that the average user only has three actions. This means that when the system recommends these items, it usually does so early (among the first three). Subsequently, by increasing the number of recommended items, the precision decreases, since there are not any relevant items left to recommend.

This measure proves to be pessimistic as Precision@k = 100% cannot be achieved when *k* is greater than the average number of actions per user. In case we had exactly one relevant item per user in the testing set, this measure could be normalized by dividing the scores with the maximum possible score that could be obtained. Example is shown by Equation 6 [5].

$$Precision@3_{norm} = \frac{18.79\%}{\frac{1}{3}} = 56.37\%$$
$$Precision@10_{norm} = \frac{6.69\%}{\frac{1}{10}} = 66.9\% \quad \textbf{(6)}$$

Comparison of our system with random recommender is shown on Figure 2.

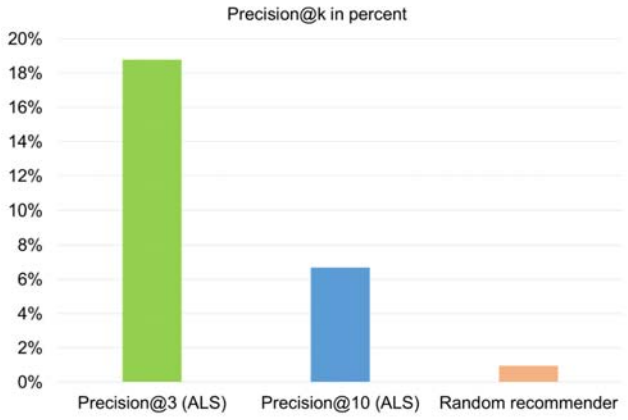


Figure 2: Precision@k comparison of the developed system with random recommender.

7. CONCLUSION

We developed a recommender system as a service for eco-friendly accommodations booking site. It was deployed in a cloud by using cloud-based machine learning server Prediction.IO. Data was collected in real time in user-action-item format by leveraging Prediction.IO built-in Event Server API. It offers recommendations for a user based on his past actions by building a latent factor model using matrix factorization with Alternating Least Squares learning algorithm. It also offers similar accommodations based on accommodation properties by computing Jaccard and Euclidian similarities.

The developed system was evaluated using the Precision@k performance measure and compared with a random recommender. It achieved 19% Precision@k which is far superior to 1% achieved by the random recommender.

MLaaS platform has shown to be an efficient tool for developing real-world machine learning applications, with a gentle learning curve. Due to good evaluation results we expect improved business results and improved user experience.

At the time of writing, a graphical representation of recommended accommodations is being implemented on the web portal. We further plan to use an algorithm developed for implicit datasets and to use other attributes about accommodations, especially price groups, since these have high influence on potential customers.

8. REFERENCES

[1] Bennett, J. and Lanning S. "The Netflix prize." *Proceedings of KDD cup and workshop*. 2007.

[2] Koren, Y., Bell R. and Volinsky C. "Matrix factorization techniques for recommender systems." *Computer* 8: 30-37, 2009.

[3] https://docs.prediction.io/. Accessed 20th September 2015.

[4] C. D. Manning, P. Raghavan, and H. Schutze. "Introduction to Information Retrieval." Cambridge University Press, New York, NY, USA, 2008.

[5] Slapničar, G. "Recommending accommodations using machine learning provider in a cloud." EngD thesis, University of Ljubljana, 2015.

[6] Adomavicius, G. and Tuzhilin, A. "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions." in Knowledge and Data Engineering, IEEE Transactions on, vol.17, no.6, pp.734-749, 2005

PREPOZNAVANJE BOLEZNI NA PODLAGI VPRAŠALNIKA IN MERITEV S SENZORJI VITALNIH ZNAKOV

Maja Somrak [1,2], Anton Gradišek [1], Mitja Luštrek [1, 2], Matjaž Gams [1, 2]
1 Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija
2 Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, 1000 Ljubljana, Slovenija
{maja.somrak, anton.gradisek, mitja.lustrek, matjaz.gams}@ijs.si

POVZETEK
V prispevku predstavljamo metodo, ki je bila razvita za prepoznavanje vrste bolezni na podlagi podatkov senzorjev, ki merijo vitalne znake, in vprašalnika o simptomih, na katerega odgovarja uporabnik. Metoda temelji na algoritmih strojnega učenja na podatkih, ki smo jih zbrali z vprašalniki ter s strokovno zdravniško pomočjo. Diagnostično metodo smo testirali na virtualnih in pravih bolnikih.

Ključne besede
Medicinska diagnostika, vprašalnik

1. UVOD
V zadnjih letih je napredek na področju senzorjev in informacijske tehnologije odprl vrata razvoju naprav in aplikacij s področja medicine, ki bodo namenjene domači uporabi. Osnovna ideja tako imenovanega m-zdravja je, da lahko uporabnik sam spremlja svoje zdravstveno stanje s pomočjo ustreznih naprav. Pri tem gre lahko za preventivni pristop (opozarjanje na morebiten pojav zdravstvenih težav, še preden te postanejo resne) ali za pomoč bolnikom s kroničnimi boleznimi, kot so sladkorna bolezen, kronično srčno popuščanje itd. Pristop m-zdravja koristi tako uporabniku, ki lahko bolje spremlja svoje zdravje, kot tudi zdravstvenemu sistemu, saj omogoča učinkovitejše in hitrejše obravnavanje posameznih bolnikov, krajšanje čakalnih vrst in nižanje stroškov.

Leta 2012 je bil objavljen natečaj Qualcomm Tricorder XPRIZE [1], katerega cilj je razviti napravo, ki bo sposobna spremljati vitalne znake posameznika ter pravilno napovedati serijo različnih bolezni. Na tekmovanju je sodelovala tudi slovenska ekipa MESI Simplifying diagnostics, v okviru katere je skupina z Instituta Jožef Stefan razvila pametno diagnostično metodo [2]. Ta na podlagi meritev vitalnih znakov in vprašalnika o simptomih, na katerega odgovarja uporabnik, določi, kakšen diagnostični test mora uporabnik narediti, da lahko potrdi ali ovrže sum na določeno diagnozo.

V tem prispevku predstavljamo strukturo metode ter trenutne rezultate testiranja na pravih in virtualnih bolnikih. Na koncu predstavimo tudi možne izboljšave pri nadaljnjem delu.

2. METODE
Diagnostična metoda je v obliki aplikacije na voljo uporabniku kot orodje za začetno diagnozo v domači uporabi. Diagnostična metoda temelji na kombinaciji vprašalnika in meritev s senzorji, ki merijo vitalne znake. Primarni vhod za diagnostično metodo (Slika 1) sestoji iz treh različnih vrst podatkov, ki jih aplikacija zajame ob samem začetku diagnostičnega procesa in vključujejo:

(1) razpoznane dejavnike tveganje na podlagi uporabnikovega profila v aplikaciji (npr. prekomerna teža, kajenje, visoka starost itd.),

(2) simptome, razpoznane na podlagi odstopanj izmerjenih vrednosti vitalnih znakov od pričakovanih vrednosti (npr. visok krvni tlak, povišana telesna temperatura itd.), in

(3) lokalizirane bolečinske simptome, ki jih uporabnik označi na anatomskem grafičnem prikazu (npr. glavobol, bolečina v prsih itd.)

Primarni vhodni podatki (dejavniki tveganja in obe vrsti simptomov) se v nadaljnjih korakih diagnostične metode obravnavajo enolično, t.j. kot simptomi. Diagnostična metoda lahko operira izključno s simptomi iz nabora vnaprej definiranih 60 simptomov. Vsak simptom se pri določenem uporabniku obravnava kot *znan* ali *neznan*, pri čemer je vsak simptom, katerega prisotnost je znana, označen kot *prisoten* ali kot *odsoten*.

Iz primarnih vhodnih podatkov diagnostična metoda pridobi informacije (znana prisotnost ali odsotnost) o 37 simptomih. Prisotni simptomi tvorijo (4) začetni nabor prisotnih simptomov.

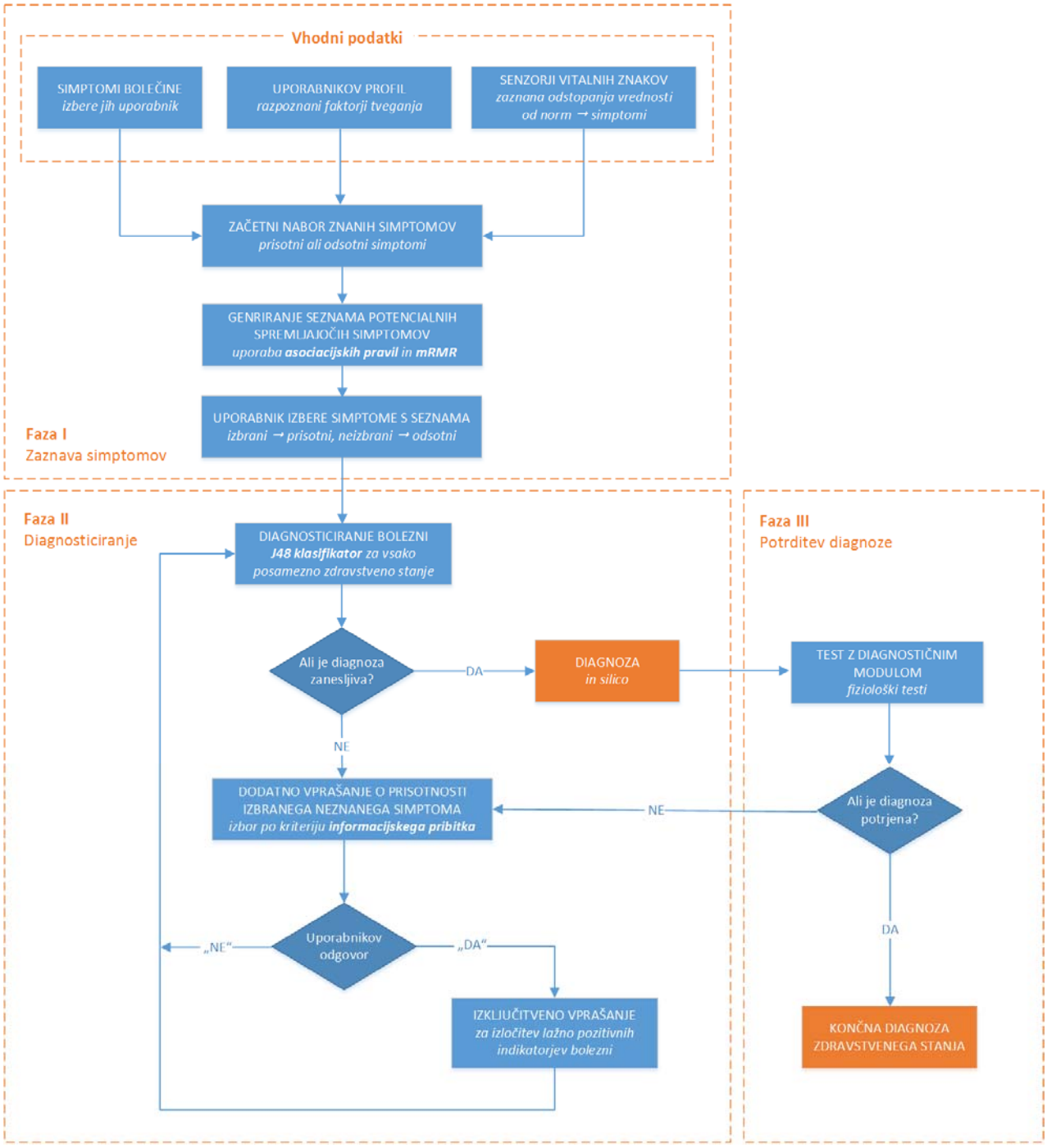
Nabor prisotnih simptomov se v naslednjem koraku diagnostičnega procesa uporabi za generiranje (5) seznama potencialnih spremljajočih simptomov. Na dotični seznam se lahko uvrstijo le simptomi iz množice neznanih simptomov. V kolikor je v začetnem naboru prisotnih simptomov vsaj en simptom, metoda za izbor potencialnih spremljajočih simptomov daje prednost tistim neznanim simptomom, ki se pogosteje (z večjo verjetnostjo) pojavljajo v kombinaciji s simptomi v trenutnem naboru prisotnih simptomov. Pri tem izboru se iščejo asociacijska pravila tipa 'simptom $j \rightarrow$ simptom i' [3–5], kjer je simptom j katerikoli simptom iz nabora prisotnih simptomov in simptom i katerikoli simptom izmed neznanih. Na podlagi pravil z najvišjo zanesljivostjo in zagotovljeno minimalno podporo je med neznanimi simptomi izbrano poljubno število potencialnih spremljajočih simptomov.

Poleg asociacijskih pravil se pri izboru potencialnih simptomov uporablja metoda minimalne redundance in maksimalne relevance (mRMR) [6]. Pri uporabi te metode se izmed neznanih simptomov izbere več simptomov, za katere velja, da:

a) prinesejo največ informacije o končnem razredu, t.j. bolezni (maksimalna relevanca) in

b) so obenem čim manj korelirani z že znanimi simptomi (minimalna redundanca).

Izbor potencialnih simptomov iz nabora neznanih simptomov (Ωs) po metodi mRMR je mogoč tudi v primeru, ko je začetni nabor prisotnih simptomov (S^+) prazen (med znanimi simptomi (S) so torej vsi simptomi odsotni; nabor odsotnih simptomov S^- je v tem primeru enak S).



Slika 1. Diagram poteka celotnega diagnostičnega procesa

Diagnostični proces poteka v treh fazah, pri čemer sta prvi dve fazi, (1) zaznava simptomov in (2) *in silico* diagnosticiranje, del razvite pametne diagnostične metode. V zadnji fazi diagnostičnega procesa poteka, (3) potrditev diagnoze, se izvedejo fiziološki testi z namenskimi diagnostičnimi moduli.

V sklopu pametne diagnostične metode je prisoten tudi korak z izključitvenim vprašanjem. Le-ta je namenjen izključno resničnim uporabnikom (in ne testiranju z virtualnimi), saj je njegov namen zmanjšanje subjektivnosti pri odgovarjanju. Izključitvena vprašanja so za vsak simptom točno določena vnaprej in so zasnovana za prepoznavanje ter izločitev najpogostejših lažnih indikatorjev bolezni. Primer izključitvenega vprašanja za simptom '*kri v urinu*' je *"Ali ste nedavno uživali rdečo peso?"* – saj se zaužita rdeča pesa izloča v urin, pri čemer ga obarva temno rdeče (nepatološko), kar uporabniki pogosto zmotno zamenjujejo s krvjo v urinu (patološko).

Uporabljeno pravilo mRMR za izbor potencialnega simptoma *i* je definirano na podlagi razlike v medsebojni informaciji [6] (alternativno bi bil lahko uporabljen tudi kvocient) po enačbi:

Enačba 1.

$$\boldsymbol{i}, \textbf{kjer}: \max_{\boldsymbol{i} \in \boldsymbol{\Omega s}} [\boldsymbol{I}(\boldsymbol{i}, \boldsymbol{h}) - \frac{\mathbf{1}}{|\boldsymbol{S}|} \sum_{\boldsymbol{j} \in \boldsymbol{S}} \boldsymbol{I}(\boldsymbol{i}, \boldsymbol{j})]$$

Izbira potencialnega simptoma z metodo mRMR se izvede nad naborom neznanih simptomov Ω*s*. Med temi je izbran tisti simptom, za katerega je vrednost funkcije (Enačba 1) največja. Pri tem je *I(i,h)* medsebojna informacija med simptomom *i* iz nabora neznanih simptomov ter boleznijo *h*, končnim razredom. *I(i,j)* je medsebojna informacija med neznanim simptomov *i* in simptomom *j* iz nabora znanih simptomov *S*, ki vsebuje |*S*| simptomov.

Ko je z enačbo mRMR izbran nov potencialni simptom *i*, bo le-ta odstranjen iz množice Ω*s* in dodan v množico potencialnih simptomov *P* (slednja je pred prvo iteracijo metode mRMR prazna). Za izbor še enega ali več dodatnih potencialnih simptomov pa simptoma oz. simptomov v množici *P* ne smemo več obravnavati kot neznane(ga), temveč kot tiste, ki so že (oz. še bodo) znani. Vse nadaljnje iteracije metode mRMR lahko torej opišemo z naslednjo enačbo:

Enačba 2.

$$i, \text{kjer}: \max_{i \in \Omega s} [I(i,h) - \frac{1}{|S| + |P|} \sum_{j \in S \cup P} I(i,j)]$$

Za izbiro vsakega dodatnega potencialnega simptoma se celotna iteracija z metodo mRMR ponovi (Enačba 2) nad naborom preostalih neznanih simptomov Ω*s*, dokler nabor potencialnih simptomov *P* ne doseže želenega števila elementov – simptomov. Zbran nabor potencialnih simptomov *P* se nato v obliki seznama ponudi uporabniku, da označi prisotne simptome (neoznačeni so posledično interpretirani kot odsotni).

V naslednjem koraku se informacije o vseh znanih – tako prisotnih kot tudi odsotnih – simptomih uporabijo za določanje bolezni oz. zdravstvenega stanja (6). Verjetnosti so izračunane za 15 vnaprej definiranih zdravstvenih stanj (14 bolezni in 'zdrav'). Za računanje teh verjetnosti se uporablja množica klasifikatorjev J48, po eden za vsako zdravstveno stanje.

Določena sta dva pragova za srednjo in višjo verjetnost prisotnosti nekega zdravstvenega stanja, ki sta empirično določena na 40% (srednji prag) in 80% (višji prag). V primeru, da verjetnosti vseh zdravstvenih stanj padejo pod srednji prag (zdravstveno stanje je malo verjetno) ali se dvignejo nad višji prag (zdravstveno stanje je zelo verjetno), se diagnoza obravnava kot zanesljiva in *in silico* diagnostični proces se konča (8). V realnem sistemu pri tem ne gre za končno diagnozo, temveč za usmeritev na najprimernejši fiziološki diagnostični test z dodatnimi diagnostičnimi moduli, s katerimi lahko diagnozo dokončno potrdimo ali ovržemo. Če denimo metoda napove, da ima uporabnik krvno bolezen (npr. anemijo), ga napoti na krvni test, v primeru kronične obstruktivne pljučne bolezni pa na test, pri katerem se posluša dihanje.

V primeru, ko se verjetnosti enega ali več zdravstvenih stanj nahajajo v območju med obema pragovoma (40 – 80%), v t.i. sivem območju, se diagnoza obravnava kot nezanesljiva. V tem primeru je potrebna informacija o prisotnosti ali odsotnosti vsaj še enega dodatnega neznanega simptoma. Kot dodatni simptom je med preostalimi neznanimi simptomi izbran tisti simptom *i*, ki ima največji informacijski pribitek *IG(h,i)*, po enačbi:

Enačba 3.

$$\text{IG}(h, i) = \text{H}(h) - \text{H}(h|i)$$

Pri tem je *H(h)* entropija končnega razreda in *H(h|i)* entropija končnega razreda v primeru, da bi bil simptom *i* znan.

Pri naši diagnostični metodi se informacijski pribitek simptomov izračuna na uteženih učnih podatkih, kjer so težje obteženi tisti primeri, katerih razredi sovpadajo s tistimi zdravstvenimi stanji, katerih verjetnosti so v sivem območju. Na ta način izbran simptom je v obliki vprašanja predstavljen uporabniku, ki potrdi njegovo prisotnost ali odsotnost (7).

Pri tem je zaradi uporabniške izkušnje pomembno, da se zanesljiva diagnoza postavi s čim manj dodatnimi vprašanji. Načeloma bi uporabnika lahko povprašali o vseh simptomih s seznama, vendar bi bilo to zamudno in večje število vprašanj nam ne bi nujno prineslo dodatnih informacij o bolezenskem stanju.

Izbor na podlagi informacijskega pribitka omogoča manjše število potrebnih vprašanj, kot če bi bila le-ta fiksno določena vnaprej. Namen uteževanja učnih podatkov je hitrejša konvergenca verjetnosti izven sivega območja in s tem še dodatno zmanjšanje števila potrebnih vprašanj za postavitev zanesljive diagnoze (8).

3. EKSPERIMENTI IN REZULTATI

Z uporabo ekspertnega znanja smo strukturirali tabelo, ki korelira 15 izbranih zdravstvenih stanj s 60 simptomi, določenimi s strani medicinskih strokovnjakov. Tabela je bila uporabljena za generiranje učnih podatkov s 15.000 virtualnimi bolniki in testnih podatkov s 1.500 virtualnimi bolniki. V obeh podatkovnih množicah so bile posamezne bolezni enakomerno zastopane; med učnimi podatki je bilo po 1000 bolnikov z vsako izmed bolezni, med testnimi podatki pa po 100 bolnikov. Celotni podatki so bili uporabljeni za učenje 15 različnih J48 klasifikatorjev, po eden za vsako izmed zdravstvenih stanj.

Testi so pokazali visoko senzitivnost in specifičnost. Na primer, pri 99% bolnikov s hipertenzijo je bolezen tudi ustrezno razpoznana oz. diagnosticirana. Med tistimi, ki pa so diagnosticirani s hipertenzijo, jih 88% tudi zares ima to bolezen. Najslabše je diagnosticiranje zdravega človeka (pri katerem so lahko sicer tudi prisotni simptomi), pri čemer je senzitivnost 61% in specifičnost 62%. Povprečna vrednost senzitivnosti preko vseh zdravstvenih stanj je 88,4%, povprečna specifičnost 88,6% in povprečna točnost 88,3% [7].

4. DISKUSIJA

Rezultati testiranj diagnostične metode na virtualnih bolnikih kažejo visoko senzitivnost, specifičnost in klasifikacijsko točnost (vse nad 80%) diagnostične metode [7]. Vrednosti so najverjetneje preveč optimistične, tudi glede na mnenja strokovnih medicinskih sodelavcev. Glavni izmed razlogov je uporaba ekspertne tabele za generiranje tako učnih kot testnih podatkov. Prav tako je število možnih zdravstvenih stanj zgolj 15, veliko manj kot sicer v praksi. Večina izmed 14 izbranih bolezni se med seboj bistveno razlikuje v simptomih, zaradi česar je med njimi lažje ločiti (višja klasifikacijska točnost). Izjeme so pljučne bolezni (pljučnica, tuberkuloza, kronična obstruktivna pljučna bolezen, spalna apneja), ki so si med seboj tudi bolj podobne po simptomih. V realističnem sistemu je ključno, da diagnostična metoda prepozna, da gre za sum

na pljučno bolezen, saj potem uporabnika napoti na ustrezni test – ta pa nato poda specifično diagnozo.

V nadaljevanju bi bilo zanimivo preučiti, kako se diagnostična metoda obnaša ob dodajanju dodatnih bolezenskih stanj in dodatnih simptomov.

5. LITERATURA

[1] Qualcomm Tricorder XPRIZE, http://www.qualcommtricorderxprize.org/

[2] Somrak M., Gradišek A., Luštrek M., Mlinar A., Sok M., in Gams M.: *Medical diagnostics based on combination of sensor and user-provided data*. (NetMed, ECAI 2014)

[3] McCormick T.H., Rudin C. in Madigan D.: *A Hierarchical Model for Association Rule Mining of Sequential Events: an Approach to Automated Medical Symptom Prediction*. Annals of Applied Statistics. (2012) .

[4] Soni S. in Vyas O.P.: *Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining.* International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), (2012) .

[5] Soni S. in Vyas O.P.: *Using Associative Classifiers for Predictive Analysis in Health Care Data Mining*. International Journal of Computer Applications (2010)

[6] Peng H., Ding C.: *Minimum Redundancy Feature Selection from Microarray Gene Expression Data.* Journal of Bioinformatics and Computational Biology (2005)

[7] Somrak M., Luštrek M., Sušterič J., Krivc T., Mlinar A., Travnik T., Stepan L., Mavsar M., Gams M.*: Tricorder: Consumer Medical Device for Discovering Common Medical Conditions*. Informatica 38, Ljubljana (2014)

Projekt DysLex

Tomaž Šef
Institut "Jožef Stefan"
Jamova cesta 39
1000 Ljubljana
+386 1 477 34 19
tomaz.sef@ijs.si

POVZETEK

Slepim in slabovidnim ter dislektikom so na razpolago različna orodja za delo z računalnikom, učenje in samopomoč. Najbolj razširjena programska oprema že omogoča prilagoditve njihovim specifičnim potrebam. Najpomembnejši pripomoček, odvisen od jezika, je sintetizator govora. Pri tem naletimo na dva problema: nezadostna kvaliteta umetno generiranega govora in (ne)podpora različnim (mobilnim) platformam oz. napravam na ravni operacijskega sistema.

E-storitev DysLex omogoča vključitev vseh slovenskih sintetizatorjev govora (tako brezplačnih kot komercialnih) v sam operacijski sistem mobilnih naprav. Zasnovana je v obliki strežnika v oblaku in pripadajoče mobilne aplikacije. Rešitev omogoča povezljivost sintetizatorja govora s poljubnim programom in se je izkazala kot najbolj smotrna za dislektike ter slepe in slabovidne, ker odpravlja potrebo po dragem razvoju različnih specifičnih aplikacij.

Ključne besede

Govorni bralnik besedil, sinteza slovenskega govora, slepi in slabovidni, disleksija.

1. UVOD

Disleksija je motnja sposobnosti branja ali razumevanja prebranega, poleg ohranjene senzorne in splošne sposobnosti. Je motnja veščin branja in pisanja, pogosto s tendenco, da se pomeša med seboj črke ali besede med branjem ali pisanjem, ali da se ne opazi določenih črk ali besed. Strokovnjaki ugotavljajo da je primerov te motnje vedno več in se je odstotek v zadnjem desetletju kar precej zvišal (5−10 % otrok). Disleksija nikakor ni bolezen. Organski izvor disleksije je sicer v možganih, vendar z njimi ni nič narobe, le malo drugače delujejo. Znanstveniki ugotavljajo, da imajo osebe z disleksijo nenavadne povezave med nevroni, saj so jih našli na neobičajnih mestih v možganih; poleg tega so ugotovili, da niso urejeni enako kot pri možganih oseb brez disleksije. Večina ljudi uporablja bolj levo polovico možganov in se lotijo reševanja problemov postopno, linearno in z logičnim sklepanjem. Pri večini ljudi so sposobnosti – talenti porazdeljeni enakomerno. Osebe z disleksijo drugače rešujejo probleme kot ostali, saj razmišljajo z desno polovico možganov. Zadev se lotijo nekoliko drugače, vendar pri večini problemov je to lahko zelo koristno, saj nas je drugačen način razmišljanja ponesel v sedanjost in nas bo pripeljal v prihodnost. Albert Einstein, Leonardo da Vinci, Pablo Picasso, Hipokrat, Galilejo Galilej, Isac Newton, William Shakespeare, Hans Christian Andersen, Walt Disney, Agatha Christie, Whoopi Goldberg, Tom Cruise, Cher, Steven Spielberg, John F. Kennedy, Steve Jobs in Richard Branson so med drugimi tudi imeli oz. imajo disleksijo, zato je vredno to motnjo sprejeti in jo obrniti sebi v korist.

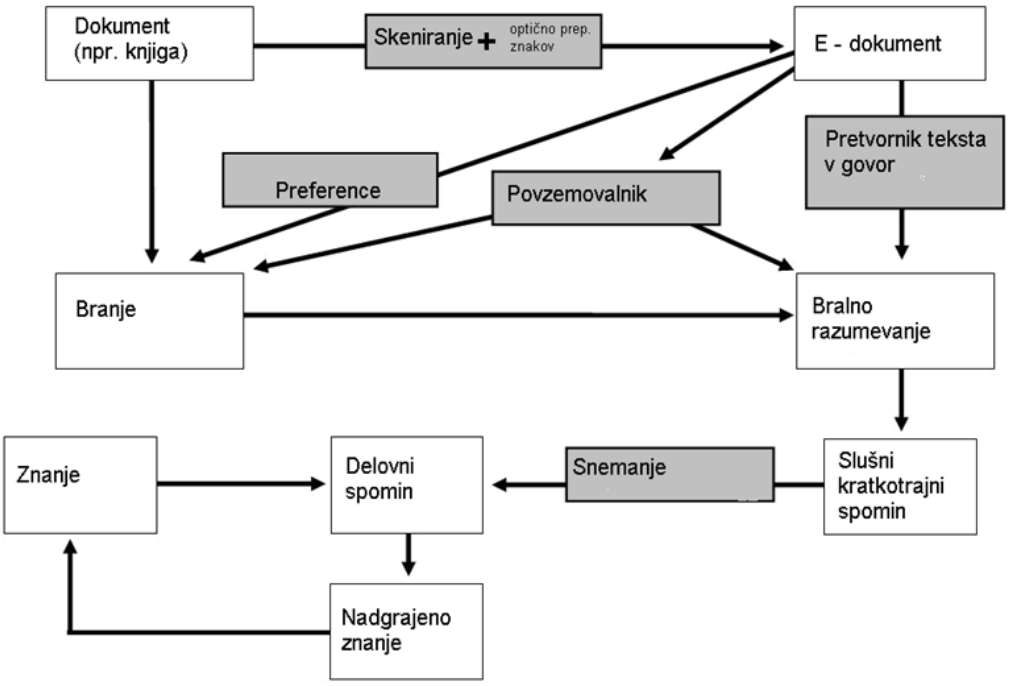
Osebe z disleksijo pri branju delajo značilne napake, ki se odražajo tudi pri črkovanju. Najbolj tipične napake pri branju so:

- počasno branje, pri katerem se otrok vidno muči,
- glasno branje je počasno, po odrezanih sekvencah (ni tekoče in gladko),
- otrok pri glasnem branju pogosto ne upošteva ločil,
- otrok je po kratkotrajnem branju vidno utrujen,
- razumevanje prebranega je pogosto slabše, ker otrok vso energijo usmeri k pravilno prebranim besedam; slušno razumevanje je običajno veliko boljše,
- otrok pogosto obrača, zamenjuje ali premešča glasove v besedi: mak – kam, zima – miza, tri – tir,
- besede, ki so videti podobno, nadomešča z drugimi, čeprav lahko spremenijo celoten pomen povedi: zahod – zavod, prt – vrt, sveča – sreča, leva – lega …,
- pri branju povedi ali zgodbe besedo nadomesti z novo sopomenko, čeprav si na pogled nista podobni: potovanje – izlet, jokanje – vekanje, deček – fant, čaša – kozarec …,
- otrok izpušča ali dodaja kratke besede: od, smo, k, za, pri …

Dislektikom so v današnjem času na razpolago **različni (brezplačni) paketi za e-učenje in orodja za samopomoč**. Najkoristnejši pripomočki so prenosni računalnik in različne mobilne naprave z ustrezno podporno programsko opremo, ki med drugim pomaga pri učinkovitem pridobivanju, zapisovanju in organizaciji informacij. Glavni elementi dostopanja do besedila so: optično branje / skeniranje (OCR), pretvornik besedila v govor (bralnik zaslona) / sintetizator govora, povzemanje, snemanje. Pri pripravi besedila pa so v pomoč: miselni vzorci, popravljanje napak (lektoriranje), pretvorba govora v besedilo (prepoznavanje govora) in razni »učitelji« tipkanja. Diagram na sliki 1 prikazuje glavne elemente dostopanja do besedil v postopku učenja [1]

Veliko tehnologij je na srečo neodvisnih od jezika. **Najpomembnejši pripomoček, odvisen od jezika, je pretvornik besedila v govor** (bralnik zaslona) **oz. sintetizator govora**, ki ga dislektiki lahko uporabljajo za različne namene, kot so:

- izgovorjava posamezne besede,
- branje elektronskih izročkov,
- dostopanje na internet,
- izgovorjava posamezne besede,
- branje elektronskih izročkov,
- dostopanje na internet,
- lektoriranje lastnega dela,
- branje skenirane knjige, ipd.



Slika 1. Glavni elementi dostopanja do besedil v postopku učenja [1].

Pri uporabi sintetizatorja govora je zelo pomembno, da slednji podpira čim več tipov elektronskega gradiva (npr. tudi pdf in internetne vire) ter čim več že obstoječe programske opreme. Umetno generirani govor mora zveneti naravno in biti prijeten za poslušanje. Pomembne so tudi nastavitve za hitrost branja in jakost zvoka ter možnost uporabe različnih glasov. Z namensko razvitimi programi (in v njih vgrajenih sintetizatorjih govora) tem zahtevam ni moč zadostiti. Edina smiselna rešitev je podpora oz. storitev vgradnje (slovenskega) sintetizatorja govora neposredno v operacijski sistem. Za okolje Windows na osebnih računalnikih je za to poskrbljeno s podporo Microsoftovemu govornemu vmesniku SAPI. Uporabnik lahko v meniju operacijskega sistema izbere slovenski glas enako kot katerega koli drugega za tuj jezik, ki ga je že serijsko vgradil Microsoft. In potem »znajo« vse Windows aplikacije avtomatsko »govoriti« tudi v našem domačem jeziku. Povsem drugačna je situacija pri mobilnih napravah, kjer takšne podpore za slovenske sintetizatorje govora še ni. Nekateri proizvajalci sistemske programske opreme (npr. Google z Androidom) v zadnjih različicah svojega mobilnega operacijskega sistema možnost takšne vgradnje že predvidevajo, drugi (Apple na iOS-u in Microsoft na Windows Phone-u) bolj odprte ter do uporabnikov in razvijalcev prijaznejše rešitve še pripravljajo.

V sledečih poglavjih si bomo podrobneje pogledali e-storitev in mobilno aplikacijo DysLex ter problematiko izdelave kvalitetnega sintetizatorja slovenskega govora.

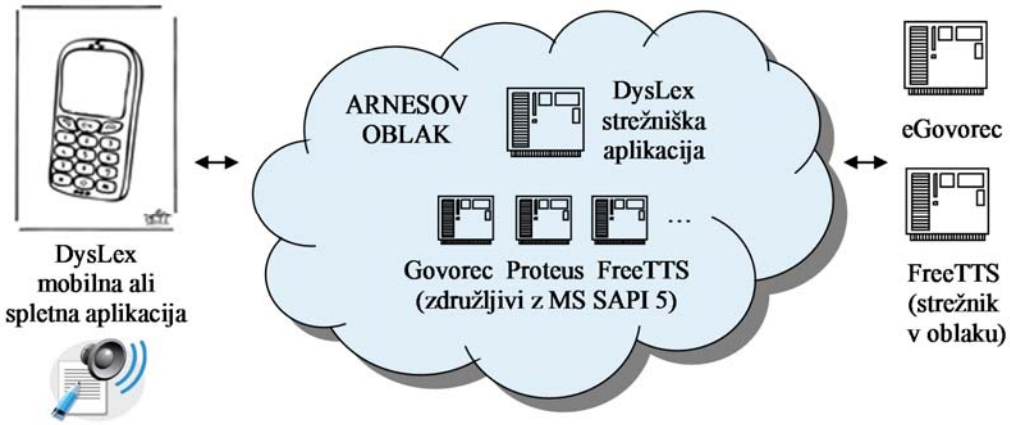
2. E-STORITEV DYSLEX

E-storitev in mobilna aplikacija DysLex omogoča vgradnjo poljubnega Microsoft SAPI kompatibilnega (slovenskega) sintetizatorja govora v operacijski sistem mobilnih naprav. E-storitev je sestavljena iz strežnika in mobilne aplikacije.

Strežnik na eni strani omogoča povezavo s sintetizatorjem govora, na drugi strani pa z mobilno napravo. Sintetizator govora je lahko naložen bodisi na istem strežniku kot sama strežniška aplikacija bodisi se poveže z nekim oddaljenim strežnikom z nameščenim sintetizatorjem govora (npr. s strežnikom eGovorec). Sintetizator govora lahko razvijalci sproti posodabljajo, ne da bi uporabniki za to sploh vedeli oz. jim za uporabo najnovejše različice ni potrebno narediti čisto nič (ni nobene potrebe po prenosu sintetizatorja govora, saj se ta posodobi na samem strežniku).

Mobilna aplikacija omogoča vgradnjo glasov vseh v strežnik prijavljenih sintetizatorjev govora v operacijski sistem mobilne naprave. V nastavitvenem meniju mobilne naprave se pri govornem bralniku avtomatsko pojavijo še slovenski glasovi. Ta funkcija deluje na vseh mobilnih napravah z operacijskim sistemom Android. Za ostale naprave, ki delujejo na operacijskem sistemu iOS ali Microsoft Phone ter na platformi HTML pa so podprte različne funkcije govornega bralnika v slovenskem jeziku, ki slepim in slabovidnim ter dislektikom olajšajo učenje in omogočajo poslušanje umetnega govora tudi na teh mobilnih napravah. Takoj, ko bodo razvijalci teh preostalih operacijskih sistemov ponudili možnost vgradnje drugih govornih sistemov v njihov operacijski sistem, bomo mobilno aplikacijo posodobili tudi s to funkcijo.

Arhitektura razvite e-storitve DysLex s strežniško aplikacijo v Arnesovem oblaku in pripadajočo mobilno aplikacijo (odjemalcem) na mobilni napravi je prikazana na sliki 2. Sintetizatorji govora se nahajajo bodisi v Arnesovem oblaku (MS SAPI 5 združljivi) bodisi na samostojnem strežniku.



Slika 2. Arhitektura e-storitve DysLex.

Vgradnja govornega bralnika v sam operacijski sistem mobilne naprave zagotavlja največjo uporabniško dostopnost in najboljšo uporabniško izkušnjo. Avtomatsko začnejo delovati vse z operacijskim sistemom podprte funkcionalnosti govornega bralnika mobilne naprave tudi v slovenskem jeziku. Takšne funkcionalnosti so npr.:

- branje izbranega besedila v poljubni aplikaciji,
- branje zaslona mobilne naprave,
- avtomatsko branje samodejnih popravkov in velikih začetnic,
- branje menijev v načinu za slepe in slabovidne oz. drugače prizadete.

Slovenski glasovi delujejo v vseh aplikacijah, ki uporabljajo govorni bralnik operacijskega sistema; vključno z vsemi aplikacijami za dislektike ter slepe in slabovidne.

E-storitev DysLex se zelo preprosto nadgrajuje in posodabljala. Podpora različnim platformam je izvedena na enem samem centralnem mestu (strežniku v Arnesovem oblaku). Zelo preprosto je dodati podporo novim platformam. Prilagoditi ali na novo prekodirati je potrebno le mobilnega odjemalca. Sintetizatorje govora lahko razvijalci posodabljajo sproti, ne da bi uporabniki za to sploh vedeli. Slednji zgolj opažajo, da je govor čedalje bolj razumljiv in naraven ter da se v sistemu pojavljajo novi glasovi v slovenskem jeziku. Namesto, da bi razvijalci večino sredstev in svojega časa namenjali podpori različnim platformam, se lahko osredotočajo na kvaliteto umetnega govora.

Aplikacija DysLex uporablja različne napredne funkcije. Vgrajeni senzor prenosa podatkov na mobilnih napravah omogoča spremljanje in nadzor nad porabo podatkovnega prenosa. Rešitev je horizontalno skalabilna, s čimer je omogočeno optimalno prilagajanje strojnih virov dejanskim potrebam oz. rasti števila uporabnikov in vsebin; omogočena je hitra vzporedna obdelava na večprocesorskih računalnikih, podprta je možnost namestitve na grozd računalnikov (v oblaku). Sistem je načrtovan tako, da na nobeni točki ne vsebuje kritične točke odpovedi.

Spoštovane so usmeritve o dostopnosti spletnih rešitev WCAG 2.0. Ne le, da aplikacija dosledno spoštuje navedene usmeritve (način prikazovanja, opis vsebin, razumljivost uporabe, enostavna navigacija in možnost iskanja, brez časovnih omejitev, povečava teksta, branje teksta, detekcija vnesenih napak, združljivost z obstoječimi tehnologijami), ki predvsem služijo podpori ljudem s posebnimi potrebami, brez takšne ali podobne rešitve pomembnega dela usmeritev na mobilnih napravah sploh ni mogoče izpolniti. Mobilne naprave oz. operacijski sistemi, ki tečejo na njih, takšno podporo v dobršni meri že nudijo (podporo se nastavi v nastavitvenem meniju operacijskega sistema, pod »Dostopnost« oz. »Accessibility«), vendar je bila do sedaj v pomembnem delu dosegljiva le za nekatere izbrane jezike. Nepogrešljiv del te podpore namreč predstavlja v operacijski sistem vgrajeni govorni bralnik, ki je sedaj na razpolago tudi za slovenski jezik.

Rešitev lahko koristno uporabljajo vsi državljani in državljanke naše države. Uporabna je lahko kot izobraževalni oz. učni pripomoček, delovni pripomoček, pa tudi med prosto časovnimi aktivnostmi. Dislektiki, slepi in slabovidni ter drugače prizadeti imajo na razpolago celo vrsto namensko razvitih aplikacij (pisanih in prilagojenih za globalni trg), ki so bile do sedaj zaradi odsotnosti podpore slovenskega govora zanje neuporabne.

3. E-GOVOREC

E-govorec ponudnikom najrazličnejših e-vsebin omogoča dinamično podajanje informacij v govorni obliki ter v domačem slovenskem jeziku. Jedro eGovorca predstavlja brezplačen sintetizator slovenskega govora (njegova kvaliteta pa še ni takšna, da bi bila sprejemljiva za večino ljudi).

Razvija se povsem novi sintetizator slovenskega govora, ki bo dokončan do konca letošnjega leta. Poleg obsežnih govornih korpusov bo vključeval tudi namensko razvite slovarje (morfološki, fonetični, pomenski) za slovenski jezik, ki omogočajo razvoj napredne avtomatske besedne, stavčne in pomenske analize besedil [2, 3]. Ob naravno zvenečem umetnem govoru bo slednje še posebej prišlo do izraza, saj bi napačno naglašene besede ipd. bile toliko bolj opažene in moteče.

Prvi prototip novega korpusnega sintetizatorja in z njim generirani demo posnetki so spodbudni in so celo presegli pričakovanja [4].

Slika 3. DysLex na platformi Windows Phone.

4. ZAKLJUČEK

Našo civilizacijo zaznamuje večinoma logično linearno mišljenje in tako je zasnovana tudi sodobna šola, ki je lahko neprijazna do oseb z disleksijo. Temeljne veščine v šoli so branje, pisanje in računanje, pot do teh veščin pa je postopna, linearna, temelji na zaporedju in na odnosu vzrok-posledica. Ravno ta utemeljenost pa je pravo nasprotje tistega, kar bi otrok z disleksijo potreboval. Znajde se pred goro problemov, ki se jih mora lotevati vsak dan v šoli in vsak popoldan doma. Pogosto so ti problemi preveliki za otrokove posebnosti, on potrebuje drugačne pogoje za svojo rast in razvoj: učenje z raziskovanjem, z osebno izkušnjo, umetniškim ustvarjanjem, drugačne pristope pri opismenjevanju, **široko uporabo posebnih tehnoloških pripomočkov in podpornih tehnologij**…

Predstavljena e-storitev in mobilna aplikacija DysLex omogoča vgradnjo sintetizatorja slovenskega govora v operacijski sistem Android, kar pomeni, da lahko umetno generirani govor poslušamo v vseh aplikacijah (npr. PDF reader, brskalnik), ki sintezo govora podpirajo in so naložene v teh napravah. Za operacijska sistema iOS in Windows Phone (slika 3) takšna vgradnja žal še ni možna, ker proizvajalca te funkcionalnosti zaenkrat ne omogočata. Na mobilnih napravah z Androidom lahko posledično dislektiki ter slepi in slabovidni uporabljajo vse (učne) pripomočke, izdelane za globalni trg ali že standardno vgrajene v napravi.

Naravnost in razumljivost novega sintetizatorja slovenskega govora sta primerljiva s sintetizatorji govora za druge jezike. Poslušanje takšnega govora ni več naporno, zato je sintetizator primeren za najširši krog potencialnih uporabnikov.

5. ZAHVALA

Operacijo delno sofinancira Evropska unija iz Evropskega sklada za regionalni razvoj ter Ministrstvo za izobraževanje, znanost in šport. Operacija se izvaja v okviru Operativnega programa krepitve regionalnih razvojnih potencialov za obdobje 2007-2013, razvojne prioritete: Gospodarsko razvojna infrastruktura; prednostne usmeritve Informacijska družba.

6. LITERATURA IN VIRI

[1] *Bravo, društvo za pomoč otrokom in mladostnikom s specifičnimi učnimi težavami.* DOI= http://doi.acm.org/10.1145/332040.332491http://www.drustvo-bravo.si

[2] Taylor, P 2009. *Text-to-Speech Synthesis*, Cambridge University Press.

[3] Šef, T. 2001. *Analiza besedila v postopku sinteze slovenskega govora*, doktorska disertacija, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

[4] Šef, T., Romih, M 2011. Zasnova govorne zbirke za sintetizator slovenskega govora Amebis Govorec, *Zbornik 14. mednarodne multikonference Informacijska družba*, zvezek A, str. 88-91, 2011.

Modeling biofilm respiration in Kamniška Bistrica riverbed sediment using decision trees

Mateja Škerjanec[1]
+386 1 425 40 52
mateja.skerjanec@fgg.uni-lj.si

Nataša Mori[2]
+386 59 23 27 39
natasa.mori@nib.si

Tatjana Simčič[2]
+386 59 23 27 38
tatjana.simcic@nib.si

Barbara Debeljak[2]
+386 59 23 27 33
barbara.debeljak@nib.si

Tjaša Kanduč[3]
+386 1 588 52 38
tjasa.kanduc@ijs.si

David Kocman[3]
+386 1 588 52 18
david.kocman@ijs.si

Primož Banovec[1]
+386 1 425 40 52
primoz.banovec@fgg.uni-lj.si

ABSTRACT
In this contribution, we aim to deepen the knowledge of the interactions between reach- and catchment-scale drivers affecting Kamniška Bistrica River metabolism and their relation to biofilm respiratory activity by applying one of the machine learning methods, namely induction of decision trees.

Categories and Subject Descriptors
E.1 [**Data structures**]: *Trees*, H.2.8 [**Database management**]: Database applications – *Data mining*, I.2.6 [**Artificial intelligence**]: Learning – *Induction*

General Terms
Measurement, Experimentation, Human Factors

Keywords
Decision trees, Kamniška Bistrica River, respiration, river metabolism, WEKA

1. INTRODUCTION
Riverbed biofilm can process high amounts of organic matter and nutrients originating from the contributing catchment, riparian zone or autochthonous river production [1]. Efficient carbon and nutrient transformations in wetted sediments can prevent river eutrophication as well as a decrease in groundwater quality. The key process in this biogeochemical cycling is respiration, which has been lately widely used for the evaluation of freshwater ecosystem functioning [2].

Within our research, we wanted to assess the ecosystem functioning of the Kamniška Bistrica River (Slovenia), which is strongly polluted (especially in its lower parts) due to the dense settlements, intense agriculture and variety of industrial activities. The most important environmental issues within the Kamniška Bistrica River catchment are wastewaters from households and industry, water leaching from agricultural soils and hydro-morphological alterations. These are especially important as the river recharges several aquifers (porous and fissured) that are used for domestic water supply and are thus highly vulnerable to contamination.

The main objective of our research was to find out which environmental factors (i.e., reach- and catchment-scale drivers) affect the biofilm respiratory activity (i.e., respiratory electron transport activity - ETS) in the Kamniška Bistrica riverbed sediment. To this end, we gathered all the available information on physicochemical properties of water, composition of riverbed sediment, channel morphology, catchment land use and nutrient inputs. The obtained dataset was used for the induction of decision trees.

Decision tree induction [3] belongs to the group of machine learning methods used to learn the dependencies between the inputs and the outputs of the observed system from the measured data by employing recursive data partitioning techniques. The method is frequently used in modeling tasks where interpretability is one of the key requirements, e.g. for explaining the interactions between different ecosystem variables. Decision trees have been successfully applied for habitat suitability modeling [4], predicting algal blooms [5], modeling lake zooplankton dynamics [6], analyzing impacts of exotic species on ecosystems [7], conducting ecological assessments [8], etc.

2. METHODS

2.1 Study area and sampling sites
For the purpose of our research, we studied the approx. 22 km long Kamniška Bistrica River stretch between the towns Spodnje Stranje and Videm (Slovenia). Between July 2013 and June 2014, we took water and sediment samples, measured temperature, oxygen, conductivity, hydraulic conductivity and vertical hydraulic gradient at six sampling sites along the selected river stretch (see Table 1).

At locations KB1, KB5 and KB6, we took 18 samples on three different dates corresponding to different seasons (summer, winter

[1] University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova 2, 1000 Ljubljana

[2] National Institute of Biology, Večna pot 111, 1000 Ljubljana

[3] Jožef Stefan Institute, Department of Environmental Sciences Jamova 39, 1000 Ljubljana

or spring). At the other locations (KB2, KB3 and KB4), we took 12 samples at two different dates, corresponding to either summer or winter season. The samples were taken from the riverbed surface and from the depth between 20 and 40 cm. Altogether, we took 90 water and sediment samples. The water samples were analyzed for physicochemical properties. For each sediment sample, we determined the content of fine sediments, grain size distribution, the content of particulate organic matter, and the biofilm respiratory activity.

Table 1. Sampling sites along the Kamniška Bistrica River

Code	Location	GKY	GKX
KB1	Spodnje Stranje	469904	122232
KB2	Perovo	469793	118424
KB3	Volčji Potok	469455	116392
KB4	Homec	469880	115061
KB5	Domžale	469861	111248
KB6	Videm	472251	103640

2.2 Data

Besides the measured data (physicochemical properties of water, sediment composition and biofilm respiratory activity), we applied GIS tools to determine the contributing catchment area for each sampling site and the percentage of different land uses in it. The percentage of land uses was also determined for the 250 m wide buffer stripe adjacent to the river segment located upstream of each sampling point. Additionally, GIS tools were used for the estimation of nitrogen and phosphorous surpluses based on modeled surpluses obtained from the Geological Survey of Slovenia (http://www.geo-zs.si/tiskaj.aspx?id=114). For the 50 meters long river segments, located upstream of each sampling site, we also evaluated the surface of wetted sediments. In order to estimate the carbon turnover within the sampled reaches, we set up a 1D hydraulic model by using LIDAR data, flow measurements at hydrometric stations operated by the Slovenian Environment Agency, and photographs taken at the same time as the field samplings were performed

Final dataset comprised 90 records (rows) corresponding to each field sample and 45 attributes (columns) representing different environmental factors (catchment- and reach-scale).

2.2.1 Introduction of qualitative measures

Because the focus of our research was on the qualitative measures of the respiratory activity in the riverbed sediment, we performed manual discretization of the measured biofilm respiratory activity. To this purpose, we introduced new discrete valued attribute replacing measured numeric attribute. One of three class values were assigned to the new attribute: “low”, “med” or “high”, corresponding to low, medium or high respiratory activity, respectively. The discretization was performed in order to ensure as equal representation of the three classes in the dataset as possible. Consequently, class values “low”, “med” and “high” were assigned to 30 records each.

2.3 Decision trees

Decision trees are hierarchical structures composed of nodes and branches, learned by splitting the source dataset into subsets based on attribute value tests. Each tree has three types of nodes: the root (the starting node) at the top of the tree, internal nodes, and the terminal nodes or leaves containing the predictions of the target variable. The root and the internal nodes contain tests on the input attributes. The branches are used to connect the nodes.

The dataset used for the induction of decision trees is typically organized in a spreadsheet table, where each row (i.e., data record) corresponds to an example and each column to an attribute (i.e., descriptor of the system). One column (usually the last one) represents a target (dependent) variable. Each example consists of measured attribute and target values.

There are two types of decision trees: classification and regression trees. Classification trees are used when the predicted (target) variable is a class. On the other hand, regression trees are used when the target variable is numeric or continuous. For the purpose of this study, we used classification trees.

One of the most popular algorithms for induction of classification trees is the C4.5 [9]. Java re-implementation of C4.5, the J48 algorithm, is a part of the machine-learning package WEKA [10], also used in this study. The J48 algorithm repeatedly partitions the original dataset into subsets, as homogeneous as possible (in terms of number of examples) with respect to the target variable. Thus, the most important task of the algorithm is to find the optimal splitting values of the measured attributes and to give the most accurate prediction of the target. The principle follows two basic steps (see Figure 1):

1. A given set of examples [S] is divided into subsets ([S1] and [S2] in Figure 1) according to the splitting value or the test of the "best" attribute. The best attribute is the one that splits the dataset [S] into the most homogeneous subsets, i.e., subsets with most homogeneous value of the target (class) variable. In Fig. 2, the data set [S] is first split into two subsets, one with all examples that have ATT 1_value $\leq$ value 1 and the other which contains all examples with ATT 1_value > value 1. Next, the subset [S1] is split into two subsets with respect to the values of the ATT 2.
2. Each subset [Si] represents a node in the tree. If all examples in a subset are of the same class as the target variable (or another stopping criterion is met) then a leaf is created, otherwise the splitting procedure is repeated for [Si]. In Figure 1, all the examples in the subset [S2] belong to the class 3, so a leaf is created. The subset [S1] is not homogeneous and is therefore further partitioned.

The algorithm repeats the above-mentioned procedure until all examples are correctly classified. However, this can result in a big tree with many branches, which is difficult to interpret. Another problem we can encounter is the problem of overfitting, i.e., when tested on new (unseen) data the model fails in its predictions.

2.3.1 Pruning

Pruning is a powerful technique to cope with tree complexity and overfitting. It improves the transparency of the induced trees by reducing their size, as well as enhances their classification accuracy by eliminating errors that are present due to noise in the data [11]. Generally, we distinguish between forward and post-

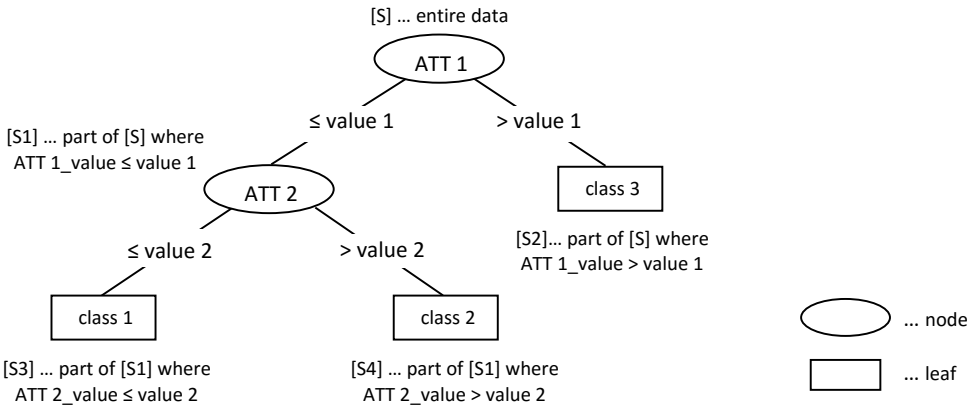


Figure 1. Classification tree – an example.

pruning. Forward pruning is applied during the tree construction procedure, by implementing some stopping criteria. An example of stopping criterion is every leave should contain a minimum number of examples or no branching is allowed. Utilization of post-pruning eliminates the least informative nodes of the tree, i.e. entire sub-trees, after the complete tree is constructed. The sub-trees are replaced by leaves based on their reliability, which is evaluated by comparing the classification error estimates before and after the replacement.

2.3.2 Evaluating decision tree models

After the tree is constructed from the training (learning) dataset, it is necessary to assess the model quality, i.e., the accuracy of prediction. This can be done by simulating the model on a testing dataset and comparing the predicted values of the target with the actual values. The model accuracy is expressed as a percentage of correctly classified examples. Another option is to employ cross-validation, where a given (training) dataset is partitioned into a chosen number of folds (n). In turn, each fold is used for testing, while the remaining n-1 folds are used for training. The final error is the averaged error of all the models throughout the procedure.

2.3.3 Attribute (feature) selection

To improve the modelling accuracy, we can employ one of the automatic attribute (also termed feature) selection techniques included in WEKA. The purpose of these techniques is to discard irrelevant or redundant attributes from a given dataset. Such techniques can be used to evaluate either individual attributes or subsets of attributes.

Information Gain Attribute Ranking [12] is one of the simplest and fastest methods for the evaluation of individual attributes. Because it can only operate on discrete valued attributes, the discretization has to be performed as a data pre-processing step in order to apply the method to numeric attributes. The method evaluates the worth of an attribute by measuring the information gain with respect to the class. Its only disadvantage is that it does not take into account attribute interaction.

The Correlation-based Feature Selection method (CFS; [13]) is used to evaluate subsets of attributes rather than individual attributes. The CFS algorithm takes into account the usefulness of individual attributes for predicting the class and the level of inter-correlation among them. It values subsets that correlate highly with the class value and have low correlation with each other.

3. RESULTS AND DISCUSSION

We examined which catchment- and reach-scale environmental factors govern the differences in respiratory activity between seasons, between (and within) sampling sites, and at different sediment depths (either at hyporheic or benthic zone). The dataset used to construct the classification tree model included 90 records corresponding to each respiration measurement and 45 attributes representing different environmental factors.

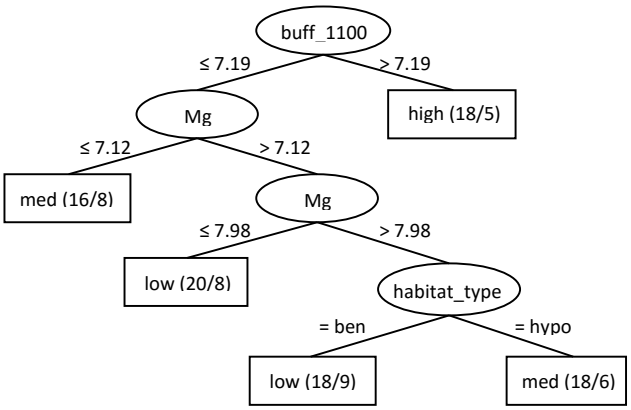


Figure 2. Modeling the factors influencing temporal and spatial distribution of respiratory activity (i.e., oxygen consumption per g of dry sediment). The values in the leaves indicate correctly/incorrectly classified examples.

The resulting tree (see Figure 2) suggests that the percentage of arable land within the catchment (attribute buff_1100) is the most important attribute affecting biofilm respiration, where over 7% of arable land in the catchment leads to a higher respiratory activity. This is not surprising, as the arable land within the catchment causes increased soil erosion and nutrient inputs [14] that consequently govern higher respiratory activity of the biofilm. The second most important attribute seems to be the concentration of Mg that depends on the catchment geology and the residence time of water. In general, when concentrations of ions in freshwater are reasonably close to the average, the biological consequences should not be very significant [15]. The

model has moderately high cross-validation (46%) and evaluation (60%) values, suggesting there are some other factors affecting the biofilm respiratory activity that are not explained by the model. However, the important result of the model is that the sediment depth (attribute habitat_type) determines the intensity of the respiratory activity. The latter seems to be higher deeper in the riverbed sediment (at the hyporheic zone, i.e., 20 to 40 cm in depth), indicating that the majority of stream metabolism is taking place there. This is in line with the findings of the previous researches [1].

In order to better understand the response of stream metabolism to different catchment- and reach-scale environmental factors, we examined the interaction between the selected environmental factors and the estimated total carbon (in mg C h^{-1}) processed within a 50 m long reach at each sampling site and during different seasons. The dataset used to construct the classification tree model included 20 records corresponding to the estimated total carbon processed within the riverbed sediments of the 50 m long river reaches and 23 attributes representing different environmental variables.

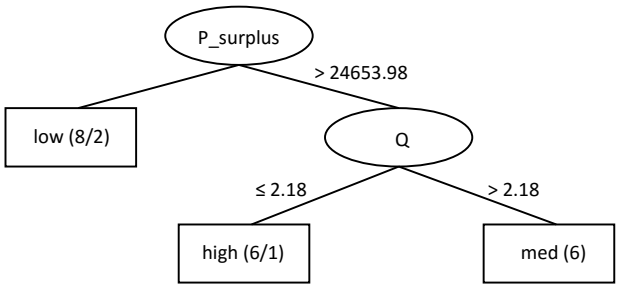


Figure 3. Modeling the factors influencing carbon processing at the river reach scale. The values in the leaves indicate correctly/incorrectly classified examples.

The constructed model (Figure 3) resulted in only two variables affecting the intensity of carbon processing: anthropogenic input of phosphorus from the catchment (P_surplus) and the river discharge (Q). Lower phosphorus input resulted in less intense carbon mineralization, while higher phosphorus contribution together with low discharge indicated higher mineralization rates. The model has relatively high cross-validation (65%) and evaluation (85%) values. By applying different attribute selection techniques, we achieved even better cross-validation values (70% when applying Information gain attribute ranking method and 85% when applying the CFS method), while the constructed trees remained the same. The modelling results indicate the importance of the hydrological regime on the intensity of the stream metabolism.

4. ACKNOWLEDGMENTS

This study was funded by the Slovenian research agency (ARRS) (L2-6778; P1-0255). Our thanks go to the colleagues from the National Institute of Biology for their assistance in performing field and laboratory analyses.

5. REFERENCES

[1] Naegeli, M. W. and Uehlinger, U. 1997. Contribution of the hyporheic zone to ecosystem metabolism in a prealpine gravel-bed river. *J. N. Am. Benthol. Soc.* 16 (Dec. 1997), 794-804.

[2] Doering, M., Uehlinger, U., Ackermann, T., Woodtli, M., and Tockner, K. 2011. Spatiotemporal heterogeneity of soil and sediment respiration in a river-floodplain mosaic (Tagliamento, NE Italy). *Freshwater Biol.* 56 (Jul. 2011), 1297-1311.

[3] Quinlan, J. R. 1986. Induction of decision trees. *Mach. Learn.* 1 (Mar. 1986), 81-106.

[4] Džeroski, S. 2009. Machine Learning Applications in Habitat Suitability Modeling. In *Artificial Intelligence Methods in the Environmental Sciences*, S.E. Haupt, A. Pasini, and C. Marzban, Eds. Springer, Berlin, 397-412.

[5] Volf, G., Atanasova, N., Kompare, B., Precali, R., and Ožanić, N. 2011. Descriptive and prediction models of phytoplankton in the northern Adriatic. *Ecol. Model.* 222 (Jul. 2011), 2502-2511.

[6] Gal, G., Škerjanec, M., and Atanasova, N. 2013. Fluctuations in water level and the dynamics of zooplankton: a data-driven modelling approach. *Freshwater Biol.* 58 (Apr. 2013), 800-816.

[7] Boets, P., Lock, K., and Goethals, P. L. M. 2013. Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. *Ecol. Inform.* 17 (Sept. 2013), 78-81.

[8] Dakou, E., D'heygere, T., Dedecker, A. P., Goethals, P. L. M., Lazaridou-Dimitriadou, M., and De Pauw N. 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol.* 41 (Sept. 2007), 399-411.

[9] Quinlan, J.R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann.

[10] Witten, I.H., Frank, E., and Hall, M.A. 2011. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[11] Bratko, I. 1989. Machine learning. In *Human and machine problem solving*, K. J. Gilhooly, Ed. Pelnum Press, New York and London, 265-287.

[12] Hall, M.A. and Holmes, G. 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE T. Knowl. Data En.* 15 (Nov. / Dec. 2003), 1437-1447.

[13] Hall, M.A. 1999. *Correlation-based Feature Subset Selection for Machine Learning*. Doctoral Thesis. University of Waikato.

[14] Allan, J. D., Erickson, D. L., and Fay, J. 1997. The influence of catchment land use on stream integrity across multiple spatial scales. *Freshwater Biol.* 37 (Feb. 1997), 149-161.

[15] Allan J. D. 1995. Stream Ecology: Structure and Function of Running Waters. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Hot water heat pump schedule optimization

Jernej Zupančič
"Jožef Stefan" Institut
Jamova cesta 39
Ljubljana, Slovenia
jernejzupancic@ijs.com

Žiga Gosar
Faculty of Mathematics and Physics
Jadranska 19
Ljubljana, Slovenia
gosar.ziga@gmail.com

Matjaž Gams
"Jožef Stefan" Institut
Jamova cesta 39
Ljubljana, Slovenia
matjazgams@ijs.com

ABSTRACT

The paper presents a multiobjective optimization approach to the heating schedules optimization for the hot water heat pump, and a modified on-off control that uses the dynamics of temperature recorded by one temperature sensor in order to estimate the amount of hot water available in the reservoir. The optimization task is to find schedules that optimally control the hot water heating according to two objectives: energy consumption and user discomfort. The problem is solved with a multiobjective evolutionary algorithm coupled with a numerical simulator of the hot water heat pump. The resulting solutions outperform the standard controls with respect to both criteria when a simulated standard hot water heater test is performed.

1. INTRODUCTION

Water heating is the second biggest energy consumer in average Slovenian household and the biggest electricity consumer [6]. While technical improvements for the heat pumps, heaters and water reservoirs are constantly developed and implemented, the potential for savings by smart scheduling is quite unexplored. Most water heating systems keep water temperature at a pre-set temperature throughout the day, with some offering some form of user-defined schedules. Smarter scheduling would find an optimised schedule for a specific household, providing energy savings and/or increasing user comfort. In this paper we utilize evolutionary algorithms to achieve smarter scheduling.

2. RELATED WORK

There has already been some research done on the topic of electric water heaters. Many focus on water heating system, when electricity tariff is dynamically changing in real time [4] or in combination with solar panels so that the effective electricity tariff is changing [7]. Other research focuses on reducing peak electricity usage and therefore the load on the electrical network [3].

All of the aforementioned research approach the problem with regards to only one objective, the costs or energy consumption of the system. Other objectives (e.g. user comfort) are either set to a constant (e.g. fully meeting all of the user requirements) or completely ignored (e.g. CO_2 emission). Various controls considered according to multiple objectives were presented [1], however, no optimization of the control parameters was performed.

In the presented work the goal is to develop intelligent scheduling algorithm for water heating according to multiple objectives. The first objective is energy consumption of the system and the second is the discomfort of the users as defined in [1].

Since the information from a single thermometer is not enough to describe the state of the water in the water reservoir, a method for the estimation of the amount of hot water remaining in the water reservoir was developed.

3. MODIFIED ON-OFF CONTROL

The modified on-off control is based on the standard temperature on-off control [8]. Given a lower boundary m and a difference δ as parameters, the heating body is activated when the measured value falls bellow m and is deactivated when the measured value exceeds $m + \delta$. The value that is compared to these boundaries in the modified on-off control is not the water temperature but the product of the water temperature and the estimated percent of hot water in the reservoir. Since hot water heat pump has two heating bodies, different parameters can be set for each of them.

3.1 Estimating the amount of hot water in a water reservoir

In our model, the reservoir has two regions, the upper region with hot water and the lower region with cold water, with a sharp border between. This roughly resembles the actual water reservoir, since they are designed in a way that the mixing of water is minimised. When drawing the water from the water reservoir, the border is moving upwards and the temperature on top is slowly decreasing, until the border reaches the top. Then the temperature as recorded by the sensors drops quickly.

When the measured temperature is decreasing, it is estimated how much energy is lost due to conduction and how much due to water draw offs. When the temperature is increasing, the amount of energy added to the system is known

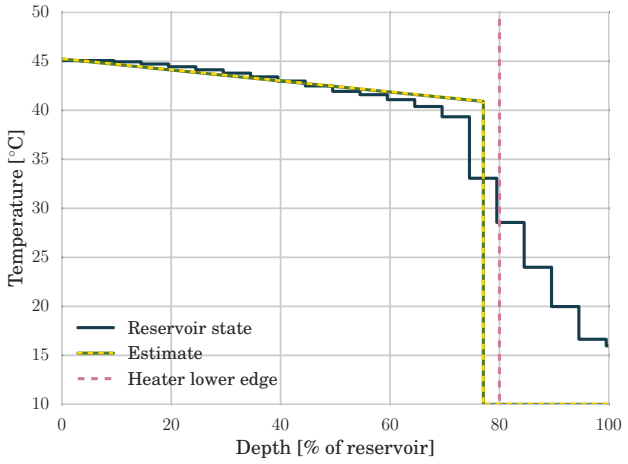


Figure 1: Simulated and estimated reservoir state.

from the heat pump operation history. Assumption is made in the estimation that only the upper region is heated. Using the information about the current temperature and temperature dynamics, energy consumed and previously estimated state of the water reservoir a new state is computed. Since the new estimated state is computed using the previous estimated state, the errors will increase with time. However, since the temperature in the upper region of reservoir is increasing quicker when all of the water is heated, we use this information to reset our estimate and eliminate the errors we have collected since the last reset. An example of a discrepancy between a simulated reservoir state and an estimated reservoir state is presented in 1.

4. EXPERIMENTAL SETUP

The above problem formulation was applied to the heat pump water heater testing using the "L" load profile [9] for a period of 28 days. As instructed in the [9], a random sequence of a selected load profiles and the load profile one below the selected load profile is generated. The probability of choosing either of the load profiles is 0.5. No distinction is made between the working days and the weekend days. As a simulator a numerical model for hot water heat pump [1] was used. The settings of the model were as follows: water reservoir 200 l, heat pump power 560 W, coefficient-of-performance 3.3, electric heater power 1500 W, heating body heats upper 80 % of the water in the reservoir. Evolutionary algorithm, developed in the DEAP [5] framework, was used for the optimization procedure.

Given the model settings, load profile, simulation duration and control strategy, the simulator numerically evaluates the heating and cooling process of the hot water heat pump and returns the values of output variables, namely average daily electric energy consumption and average daily user discomfort according to [1].

DEAP is a Python framework for evolutionary computation that implements various building blocks for designing the evolutionary algorithms. In this study a population-based algorithm for multiobjective optimization was used comprising the following operators:

- First generation of individuals is generated with uniform random sampling from valid intervals.
- Blend crossover; the child individual c is an affine combination of two parent individuals p_1 and p_2

  $$c = \gamma * p_1 + (1 - \gamma) * p_2, \gamma \in [-\alpha, 1 + \alpha],$$

  where α is a user defined parameter. In this study the α was set to 0.2.
- Gaussian mutation; random numbers sampled from Gaussian distribution are added to the the parent values p in order to obtain the mutant m

  $$m = p + \delta, \delta \sim \mathcal{N}(\mu, \sigma),$$

  where μ and σ are user defined parameters. In this study we used $\mu = 0$ and $\sigma = 1$
- Set to bounds repair function; each gene of the individual is checked for boundary constraints after the operators of crossover and mutation are applied, when the gene value exceeds the boundary value the gene value is set to boundary value.
- Non-dominated sorting procedure and the crowding distance metric known from NSGA-II [2] are used for new population selection.

Additional algorithm parameter values in this study were as follows: population size 200, number of solution evaluations 50 000, crossover probability 0.7, mutation probability 1, and gene mutation probability 0.05.

4.1 Individual representation

The optimization algorithm searches for the optimal heating schedule according to the selected load profile testing scenario. Each heating schedule can be can be encoded as a list of quintuples

$$\left(t_i, m^i_{EH}, \delta^i_{EH}, m^i_{HP}, \delta^i_{HP}\right),$$

where m^i_{EH}, δ^i_{EH}, m^i_{HP}, δ^i_{HP} are the parameters of an modified on-off control described in section 3 and t_i is the duration for which the modified on-off control with those parameters is used. The quintuples are chained together in order to produce an individual as shown in figure 2. Based on the preliminary experiments we have set the length of the individuals to 7 quintuples, which translates into 35 genes.

t_0	m^0_{EH}	δ^0_{EH}	m^0_{HP}	δ^0_{HP}	$\cdots$	t_m	m^m_{EH}	δ^m_{EH}	m^m_{HP}	δ^m_{HP}

Figure 2: Individual encoding

The m-th quintuple of parameters for the modified on-off control are active in the time interval

$$T_m = \left[\sum_{i=0}^{m-1} t_i, \sum_{i=0}^{m} t_i\right] \cap [0\,\mathrm{h}, 24\,\mathrm{h}].$$

If the active time interval is empty ($T_m = \emptyset$) then the controls are never active. In order to remove the individuals that have the same active time intervals and the same values of parameters for each active interval (have the same

phenotype) we also implemented a population truncation operator. This operator eliminates duplicate phenotype individuals, so that among duplicates only the individual with the smallest sum of its gene values is left in the population. This operator ensures the diversity of population.

5. RESULTS

This section presents and discusses the results of the performed optimization experiment. Figure 3a shows the value of the hypervolume indicator over 50 000 evaluations for four runs of the algorithm. Final hypervolume values are different for each run since the generation of a load profile testing scenario is a random process, therefore, different criteria function is used in each run. We can see that the algorithm converges rather quickly to the hypervolume values that are close to the best values achieved during the runs. Only minor improvements can be observed after 30 000 evaluations. This implies that the optimization could be stopped earlier without significantly deterioration the results and may even be desired in order to prevent the excessive overfitting to the randomly generated testing scenario. Note that the entire run (60 000 parallel evaluations using 6 processes) took approximately 4.5 days on a 3.4 GHz Intel Core i7-2600 CPU with 8 GB RAM.

The non-dominated solutions of the last generations mapped into the objective space are shown in figure 3b. Two knees can be observed in the figure, first one is where the discomfort values are at about 0.05 and the second one is at the discomfort values of about 0. When examining the solutions in the parameter space two distinct types of solutions were identified. Figure 4a shows a typical solution that belongs to the first knee. One can observe that mainly the heat pump was used in order to heat the water, since the parameter value m_{EH} is set too low (and therefore never reached) in order to turn on the electric heater. Figure 4b shows a typical solution that belongs to the second knee. One can observe that although usually the heat pump is used there is a time interval when the electric heater is turned on for a short period of time and it is set to heat the water significantly. When examining the used "L" load profile it was found that hot water at 55 °C is required at 12:45 and at 20:30, at other times lower water temperatures are required. The optimization correctly identified the hot water requirements and heated the water accordingly.

5.1 Results generalization

While (sub)optimal solutions were found for some randomly generated 28 days sequences of load profiles, we were also interested in how the solutions perform in another randomly generated 28 days sequence i.e. how the solution generalize. We have chosen the non-dominated solutions from each of the optimization runs and evaluated them on a newly generated 28 days sequence of load profiles. The performance of the individuals is shown in figure 5. Solutions of the first knee still outperform standard temperature on-off control that uses only heat pump (no electric heater) by up to 3 % in energy consumption and up to 30 % in the discomfort criteria. Solutions of the second knee, however, outperform standard temperature on-off control, which uses electric heater and heat pump, according to the energy consumption (1 to 45 %) while decreasing the discomfort by 25

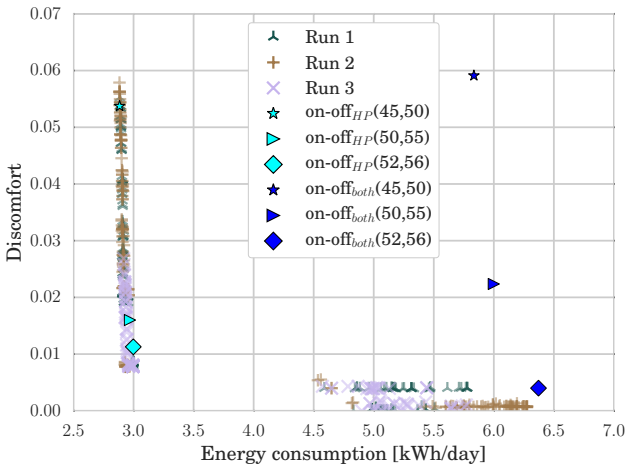


Figure 5: Test run performance compared to temperature on-off controls.

or even 100 % therefore meeting all of the comfort requirements of the "L" load profile.

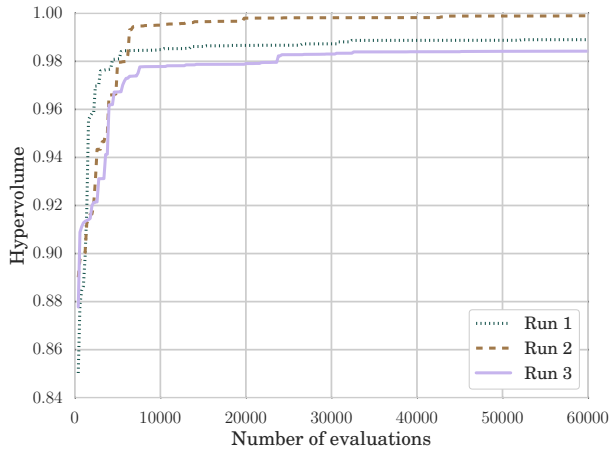
6. CONCLUSION

Several hot water heat pump manufacturers enable the user to set the heating schedule for their hot water heat pump according to his preferences. However, a well chosen parameters require the knowledge of the heat pump heating processes and the knowledge of hot water consumption patterns. While an expert could set an optimal schedule given enough information, the user looses the power to optimally adjust the operation of her heat pump according to her needs. In this study we implemented a multiobjective evolutionary algorithm that finds (sub)optimal schedules for hot water heat pump according to two objectives (energy consumption and discomfort). The optimization takes into account the user's hot water consumption pattern and heat pump characteristics in order to (sub)optimally adjust the schedules, which outperform traditional control strategies. User could the choose the schedule that best reflects her preferences according to the energy consumption and discomfort. Further, we developed a modified on-off control strategy that uses the estimate of the amount of hot water available in the reservoir rather than temperature for control decisions.

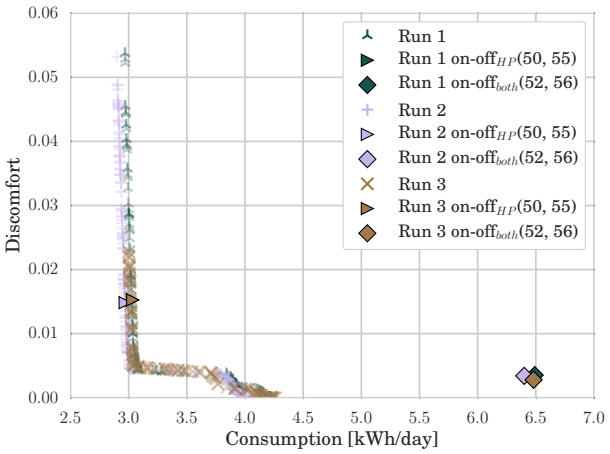
In the future work, schedules will be replaced by decision trees that will determine what parameters to choose for heating control depends on various conditions (including environmental parameters) and not on the hour only. This approach is expected to further improve the performance of the controller and generalize better. Different stopping criteria will be investigated in order to avoid overfitting. Further, machine learning methods will be applied in order to improve the estimates of the amount of hot water available based on temperature dynamics captured by a single temperature sensor.

7. ACKNOWLEDGMENTS

We thank Jure Brence and Vid Seražin for the help in the hot water heat pump simulator implementation.

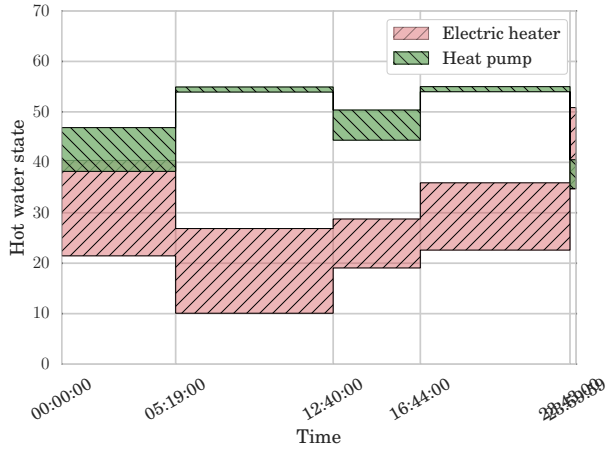


(a) Algorithm performance.

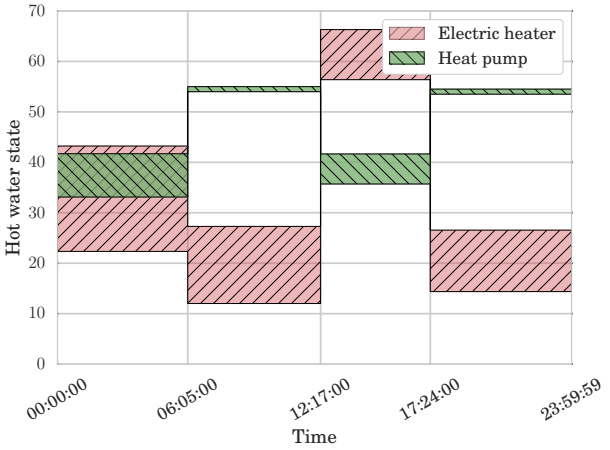


(b) Non-dominated solutions and temperature on-off controls.

Figure 3: Final results of the algorithm runs.



(a) A typical solution of the first knee.



(b) A typical solution of the second knee.

Figure 4: Solution representatives. The heating body is activated/deactivated if the hot water state falls below/rises above the coloured band.

8. REFERENCES

[1] J. Brence, Ž. Gosar, V. Seražin, J. Zupančič, and M. Gams. Multiobjective optimisation of water heater scheduling. In *Proceedings of the 17th International Multiconference INFORMATION SOCIETY 2014*, volume A, pages 5–8, October 2014.

[2] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, 2002.

[3] R. Diao, S. Lu, M. Elizondo, E. Mayhorn, Y. Zhang, and N. Samaan. Electric water heater modeling and control strategies for demand response. In *Power and Energy Society General Meeting, 2012 IEEE*, pages 1–8. IEEE, 2012.

[4] P. Du and N. Lu. Appliance commitment for household load scheduling. *Smart Grid, IEEE Transactions on*, 2(2):411–419, 2011.

[5] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, jul 2012.

[6] T. Rutar. Poraba energije in goriv v gospodinjstvih, slovenija, 2013 - končni podatki. `http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6564`. Accessed: 2015-09-09.

[7] S. Sichilalu, X. Xia, and J. Zhang. Optimal scheduling strategy for a grid-connected photovoltaic system for heat pump water heaters. *Energy Procedia*, 61:1511–1514, 2014.

[8] L. Sonneborn and F. Van Vleck. The bang-bang principle for linear control systems. *SIAM Journal on Control and Optimization*, 2(2):151, 1964.

[9] The European Commission. Commission delegated regulation (eu) no 812/2013. *Official Journal of the European Union*, 239:83–135, 2013.

Classification of fictional *What-if* ideas

Martin Žnidaršič, Jasmina Smailović
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
{martin.znidarsic, jasmina.smailovic}@ijs.si

ABSTRACT

The paper presents the process and the outcomes of machine learning the human evaluation model of textually represented fictional ideas. The ideas that we are targeting are generated artificially by various computational creativity systems. The study presented in this paper is one of the first that employ machine learning in the challenging domain of evaluating the artificially generated creative artefacts. Our results confirm the validity of the approach and indicate the differences in importance of the features used.

1. INTRODUCTION

The work presented in this paper is addressing the problem of machine learning the differences among good and bad fictional *What-if* ideas that are represented as human readable text. Fictional *What-if* ideas are ideas such as:

> What if there was a city built in the air?

that start with a "What if " proposition and describe a fictional situation, i.e., a situation that is not realistic or is at least not commonly considered plausible. The ideas of this kind are practically used in art, literature, entertainment industry (e.g., movie plots) and advertisement. In these domains the ideas are the main driving force of creative work, and their production, also of the *What-if* kind, is an essential part of the conceptualization phase of product development.

The *What-if* ideas that we are focused on are not only fictional, but also computer generated. Automated production of such creative artefacts belongs into domain of computational creativity, a subfield of artificial intelligence that is concerned with understanding, simulation and recreation of creative behaviours [1, 7, 8]. Although production of fictional ideas is deemed creative work and as such eludes solutions with computational means, there are now automatic *What-if* idea generators being developed [5, 9], mostly in the scope of the WHIM project[1]. The idea generators can be parameterized in various ways to generate large amounts of various flavours of *What-if* ideas. However, most of these automatically generated ideas are noisy and of low quality, particularly from the generators that are producing less restricted outputs, which despite a larger proportion of noise, usually produce the more interesting and valuable results. Automated evaluation of such results is very difficult, but it would also be very beneficial, if it could be made possible. Our aim is to create such an evaluation system with the means of machine learning. More specifically, based on human-labelled data we intend to create (components of) a human evaluation model of fictional *What-if* ideas. There are two questions that we are aiming to answer with our work: (I) can an evaluation model for *What-ifs* be constructed automatically, and (II) which features of the *What-ifs* are the most relevant in this respect?

Evaluation is difficult and controversial in the context of computational creativity systems [6, 4]. Namely, for the outputs of these systems there is often no common measure of appropriateness or value. We intend to create a general evaluation model, which might suppress some subjective views and should reflect the general population as much as possible. For this reason, we are using the crowdsourcing approach through an open online evaluation platform and crowdsourced questionnaires.

Results of our experiments indicate that, despite the challenging nature of the problem, we can address it with machine learning methods. Namely, despite relatively low classification performance results, the evaluation models improve upon the baseline, which is not an obvious result in such a challenging domain. Regarding the feature importance, the outcomes are not straightforward. Benefits of using the words from the *What-ifs* are evident, while the other features have a much less profound impact on performance.

2. DATA

The aim of our data gathering is the acquisition of human evaluations of the computer-generated *What-if* ideas. This data serves as the basis for the development of the audience evaluation model. We are using two kinds of opinion sources: (I) the assessments and opinions of anonymous visitors of our open online platform[2] and (II) the results from crowdsourced questionnaires with a specific experimental target.

[1] www.whim-project.eu
[2] http://www.whim-project.eu/whatifmachine/

Table 1: Label distribution for each generator.

Gen.	1	2	3	4	5	Sum
Labels in the open platform						
Alt	603	337	277	211	176	1,604
Dis	1,043	647	680	776	551	3,697
Kaf	315	188	168	189	154	1,014
Met	111	143	204	267	180	905
Mus	30	33	16	37	21	137
Uto	563	448	434	466	318	2,229
all	2,665	1,796	1,779	1,946	1,400	9,586
Labels in the questionnaires						
Alt	929	933	1,239	891	403	4,395
Dis	861	952	1,349	861	411	4,434
Kaf	923	880	1,239	920	454	4,416
Met	657	855	1,398	1,032	497	4,439
Uto	769	824	1,336	930	519	4,378
Test	1,527	745	935	632	413	4,252
all	5,666	5,189	7,496	5,266	2697	26,314

We will denote these opinion sources as *open* and *targeted* respectively. The first one is intended for gathering simple casual opinions from the general public, while the aim of the second is to gather data for specific experiments. The assessment procedure in the online platform is tailored to the online context and favors simplicity and clarity of the interface over thoroughness of the assessments. The questionnaires, on the other hand, are more controlled as we can decide to accept only fully completed questionnaires, we can ask for demographic data and we can evaluate more complex concepts (such as novelty, narrative potential, etc.), since the paid evaluators are expected to devote more effort to the task.

2.1 Datasets

Two datasets were prepared for the work presented in this paper as described in the following.

2.1.1 *Open: from the open online platform*

The first dataset is a collection of evaluated *What-if* sentences obtained using the open online platform. *What-ifs* on this platform were created by 6 different generators – Alternative Scenarios, Disney, Kafkaesque, Metaphors, Musicals, and Utopias and Distopias. The anonymous visitors rated the *What-ifs* of a selected generator on a 5-point Likert scale from 1 to 5. They could also report a sentence as offensive/incorrect or comment it. Using the online platform we acquired 9,586 labels for 5,908 different *What-if* sentences. The label distributions for each generator are shown in Table 1.

Since some *What-if* sentences were labeled multiple times by different people, we merged multiple labels of one sentence into one label by calculating their median value. We considered the *What-ifs* marked with labels 2 or less to belong to the negative class, those with labels 4 or more to the positive class and the others to the neutral class. The final label/generator distribution is shown in Figure 1. In the experiments and evaluation, we used only positively and negatively scored *What-if* sentences. There are 1,881 of the first kind and 2,600 of the latter. The majority class represents 58.02% of this dataset.

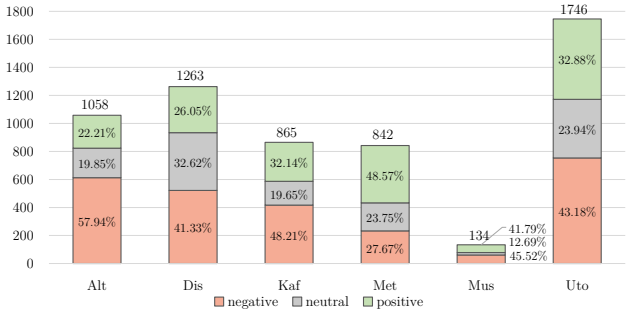


Figure 1: Label distribution for the open dataset.

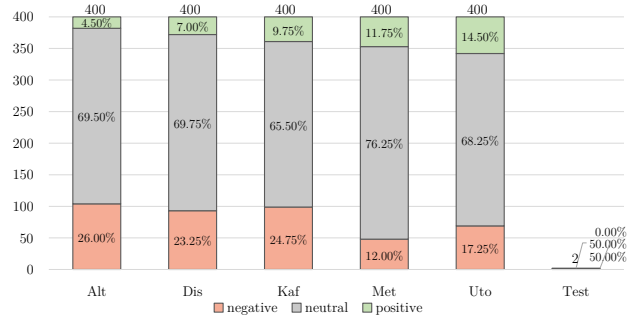


Figure 2: Label distribution for the targeted dataset.

2.1.2 *Targeted: from crowdsourced questionnaires*

The second dataset consists of a collection of *What-if* sentences labelled in crowdsourced[3] questionnaires. *What-ifs* were created by 5 different generators – Alternative Scenarios, Disney, Kafkaesque, Metaphors, and Utopias and Distopias. There were also two manually constructed *What-ifs* for testing purposes (denoted as Test). The annotators labeled *What-ifs* based on several criteria, however in this study we only analyze scores of the overall impression. The annotators rated the *What-ifs* on a 5-point Likert scale from 1 to 5 to express their overall impression about them. We acquired 26,314 labels for 2,002 *What-if* sentences. The label distributions for each generator are in Table 1.

As in the *open* dataset, we merged multiple labels into one (median) label, which was mapped into negative/neutral/positive class. The resulting label distribution is shown in Figure 2. The final *targeted* dataset contains all the unique positively scored (190) and negatively scored (414) examples, with the majority class covering 68.54% of items.

2.2 Data features

The data items that we collect consist of a textual representation of a *What-if* and an assigned label from the [1..5] interval. For the use in our experiments, we constructed a number of features to better represent each data item for the machine learning purposes. Merging of the labels was simple, as we have used a median value of the assessments in cases when there were more than one for a given item.

[3]We used the crowdsourcing platform CrowdFlower (`http://www.crowdflower.com/`)

Definition of the feature space was more complex and is described in the following, separately for the pure textual features and for some additional ones.

2.2.1 *n-grams*

In the process[4] of constructing textual features we first applied tokenization and stemming of the *What-if* sentences. Furthermore, we constructed n-grams of length 1 (i.e., unigrams) and 2 (i.e., bigrams), and discarded those whose number of occurrences in the dataset was lower than 2. Finally, we created feature vectors by employing the term frequency feature representation. However, as our initial experiments showed only a minor advantage of bigrams compared to unigrams (accuracy of 62.47% compared to 62.45% on the *open* dataset, and 74.95% compared to 74.92% on the *targeted* dataset), we have used only the latter, which allowed for much faster experimental work.

2.2.2 *Additional features*

In addition to n-grams, we have used the following additional features of each *What-if*:

- *length*: is the length of a *What-if* in terms of the number of characters;
- *ambiguity*: is an assessment of ambiguity according to the number of different meanings of a term in Wordnet[5]. It corresponds to the average number of Wordnet senses of known content words in the text;
- *rhyming*: denotes if there is a pair of words in the sentence that rhyme. It corresponds to the number of word combinations in the *What-if* that rhyme, given the rhyming level (number of ending syllables that have to match; we have used 2 in our experiments);
- *number of adjectives*: corresponds to the number of adjectives that appear in a *What-if*; and
- *number of verbs*: corresponds to the number of verbs in a *What-if* .

A feature that we currently have access to is also the *generator* - the algorithm that generated a specific *What-if* idea. The positive and negative *What-ifs* are similarly distributed with respect to generators (as shown in Figures 1 and 2), but some distinctive differences can be observed and they could contribute to classification performance. However, as the generators are constantly being updated and as we want our evaluation system to be general and not limited with respect to a set of specific generators, we did not use these features in our datasets.

3. EXPERIMENTS AND EVALUATION

In this section we present the methodology of experimental work and the results of evaluation on the *open* and *targeted* datasets.

3.1 Methodology

For classification we employed the SVM classifier[6]. Evaluation was performed using 10-fold cross validation. The experiments were performed 10 times – each time with different examples in folds. We experimented with using different feature sets: (I) only n-grams (words), (II) only 3-class additional features, where the borders between three classes were determined by discretization of the sorted feature data into 3 equal frequency partitions (oth(3c)), (III) both n-grams and 3-class additional features (w+oth(3c)), (IV) only 2-class additional features, where the border between two classes was the median feature value (oth(2c)), and (V) both n-grams and 2-class additional features (w+oth(2c)). The evaluation results and the average ranks of feature sets are shown in Tables 2 and 3.

Table 2: Experimental comparison of classifiers learned on the *open* dataset with various feature sets. Results of ten 10-fold CV experiments with different folds are shown. In bold is the average over all ten experiments and in the last row, the average ranks (according to ranking in each experiment).

words	oth(3c)	w+oth(3c)	oth(2c)	w+oth(2c)
62.57%	60.54%	62.33%	60.72%	62.35%
62.31%	59.99%	62.58%	59.83%	62.64%
62.53%	60.59%	62.58%	60.03%	62.58%
62.49%	59.87%	62.33%	60.61%	62.35%
62.40%	59.43%	62.46%	59.92%	62.46%
62.51%	59.65%	62.33%	60.72%	62.35%
62.55%	60.46%	62.51%	60.72%	62.53%
62.33%	60.37%	62.40%	60.68%	62.49%
62.44%	60.57%	62.46%	60.72%	62.51%
62.40%	60.57%	62.62%	60.68%	62.58%
62.45%	**60.20%**	**62.46%**	**60.46%**	**62.48%**
2.2	4.8	2.2	4.2	1.6

Table 3: Experimental comparison of classifiers learned on the *targeted* dataset with various feature sets. Results of ten 10-fold CV experiments with different folds are shown. In bold is the average over all ten experiments and in the last row, the average ranks (according to ranking in each experiment).

words	oth(3c)	w+oth(3c)	oth(2c)	w+oth(2c)
75.22%	68.89%	74.57%	67.42%	74.92%
74.85%	68.75%	73.19%	67.41%	73.36%
74.84%	68.86%	73.68%	67.88%	73.83%
74.84%	68.09%	74.69%	66.43%	74.69%
75.83%	68.72%	74.85%	66.54%	74.83%
74.65%	68.97%	74.81%	67.19%	74.98%
74.32%	67.36%	74.81%	65.89%	74.32%
75.30%	68.85%	74.30%	67.55%	73.80%
74.17%	69.43%	73.34%	68.07%	73.34%
75.16%	67.86%	74.82%	66.03%	75.50%
74.92%	**68.58%**	**74.31%**	**67.04%**	**74.36%**
1.45	4	2.4	5	2.15

3.2 Discussion of results

According to results in Tables 2 and 3, the machine-learned classifiers were able to distinguish (to a limited extend, of course) between the *What-ifs* that are generally considered good and the ones that are considered bad. Namely, all the classifiers that consider all the available features were able to beat the baseline of the given classification problem. Among

[4]The process was performed using the LATINO toolbox (`http://source.ijs.si/mgrcar/latino`)

[5]`https://wordnet.princeton.edu/`

[6]We used the wrapper around the SVM^{Light} [3] classifier in the LATINO library.

Table 4: The features that are relatively (in one class, compared to the other) the most represented in positively and negatively scored *What-ifs* for the *open* and the *targeted* dataset.

dataset	positive	diff	negative	diff
	'used'	204	'there'	-826
	'were'	176	'was'	-723
	'become'	159	'if'	-719
	'and'	99	'what'	-719
open	'their'	89	'?'	-719
	'by'	34	'a'	-693
	'embrace'	27	↓numvb	-642
	'beautiful'	24	'who'	-624
	'stars'	23	↑ambiguity	-511
	'engineer'	22	↓rhyming	-471
	'not'	31	'what'	-224
	'all'	31	'if'	-224
	'its'	30	'?'	-224
	'lies'	16	'there'	-222
targeted	'used'	14	'was'	-211
	'invent'	14	'a'	-209
	'stories'	11	↓numvb	-195
	'tell'	9	'who'	-179
	'talk'	8	↓numadj	-176
	'inherit'	7	↓rhyming	-172

the classifiers, there is a notable difference in performance of those that use the words of *What-ifs* as features and those that do not, as the former (denoted: words, w+oth(3c) and w+oth(2c)) consistently yield higher accuracies. Statistical analysis with the suggested [2] Friedman test indicates that the differences among these classifiers are not significant. It has to be noted though that the test is very conservative, particularly in situation of only two independent datasets. As there is no statistical test available for multiple classifiers on a single dataset and as repeated 10-fold CVs on the same dataset cannot be strictly regarded as independent, we report the average ranks of the experiments for all the classifiers. These ranks are nevertheless considered [2] as a fair and useful method of comparison.

Using the words as features is clearly beneficial, while the usefulness of the other features of the *What-ifs* is not so evident. In the *open* dataset it seems that the use of additional features is beneficial, as the average accuracies are higher, but this is not the case in the *targeted* dataset. However, in some non-reported experiments in which we set the discretization borders manually in an arbitrary fashion (the reason for ommiting them from Tables 2 and 3), the best results were gained with such a combined feature set, e.g.: 62.51% accuracy with a somewhat worse rank than the w+oth(2c) for the *open* dataset and 74.95% with best rank for the *targeted*. Although this does not affect the significance of the results, it indicates the importance of feature preprocessing (discretization method in this case). Another indication that the additional features have some impact is shown in Table 4, where we present the top ten features that are relatively the most represented in either one or the other class for each dataset. The features with an ↑ or ↓ sign represent high or low values of such an additional feature. The features in quotes are the n-grams (words). Appearance of additional features among the top most representative ones indicates that some of them do have a notable relation with the general human perception of *What-if* texts.

4. CONCLUSIONS

In the paper, we presented the data collection and modelling processes aimed at generating a data-based evaluation model for *What-if* ideas. According to results of experiments, the created models manage to reflect the human evaluation behaviour, but to a limited extend. Importance of the features remains an open question, as the results clearly indicate the benefit of using the words that appear in textual representations of *What-ifs*, but are not conclusive regarding the importance of the more complex additionally computed features. With more data which we expect in the future, we intend to build upon this work and shed more light on the characteristics of feature construction and selection in this challenging problem domain.

5. ACKNOWLEDGMENTS

This research was supported by the Slovene Research Agency and through EC funding for the project WHIM 611560 by FP7, the ICT theme, and the Future Emerging Technologies FET programme.

6. REFERENCES

[1] S. Colton and G. A. Wiggins. Computational creativity: the final frontier? In *ECAI*, volume 12, pages 21–26, 2012.

[2] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[3] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[4] A. Jordanous. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3):246–279, 2012.

[5] M. T. Llano, R. Hepworth, S. Colton, J. Gow, J. Charnley, N. Lavrač, M. Žnidaršič, M. Perovšek, M. Granroth-Wilding, and S. Clark. Baseline methods for automated fictional ideation. In *Proceedings of the International Conference on Computational Creativity*, 2014.

[6] G. Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99, 2007.

[7] R. Saunders. Towards autonomous creative systems: A computational approach. *Cognitive Computation*, 4(3):216–225, 2012.

[8] T. Veale. A service-oriented architecture for computational creativity. *Journal of Computing Science and Engineering*, 7(3):159–167, 2013.

[9] T. Veale. Coming good and breaking bad: Generating transformative character arcs for use in compelling stories. In *Proceedings of ICCC-2014, the 5th International Conference on Computational Creativity, Ljubljana, June 2014*, 2014.

Towards ABAC Policy Mining from Logs with Deep Learning

Decebal Constantin Mocanu
Eindhoven University of Technology
Dep. Of Electrical Engineering
Eindhoven, the Netherlands
d.c.mocanu@tue.nl

Fatih Turkmen
Eindhoven University of Technology
Dep. Of Mathematics and Computer Science
Eindhoven, the Netherlands
f.turkmen@tue.nl

Antonio Liotta
Eindhoven University of Technology
Dep. Of Electrical Engineering
Eindhoven, the Netherlands
a.liotta@tue.nl

ABSTRACT

Protection of sensitive information in platforms such as the ones offered by smart cities requires careful enforcement of access control rules that denote "who can/cannot access to what under which circumstances". In this paper, we propose our ongoing work on the development of a deep learning technique to infer policies from logs. Our proposal improves the state-of-the-art by supporting negative authorizations (i.e. denied access requests) and different types of noise in logs. A preliminary evaluation of the proposed technique is also presented in the paper.

Categories and Subject Descriptors

Computing methodologies [**Machine learning approaches**]: [Neural networks]; Security and privacy [**Security services**]: [Access control]

General Terms

Theory, Algorithms, Security

Keywords

Deep Learning, Boltzmann Machines, Density Estimation, Attribute-Based Access Control, ABAC Policy Mining

1. INTRODUCTION

Smart urban systems such as ACCUS[1] with tightly integrated components will collect and disseminate large amounts of sensitive information from/to citizens, government agencies and commercial/non-commercial organizations. While they will fundamentally improve citizen's life, contribute to preservation of environment and enable sustainability, they will also expose certain security vulnerabilities. Protecting sensitive information from unauthorized access in this context is one of the major obstacles in obtaining full benefit from their deployments. Example consequences of data leakage include financial loss, reputation damage or even social unrest. Authorizations systems ensure that access to sensitive information is strictly regulated through security policies which encode rules on "who can/cannot access to what under which circumstances".

There are different languages to specify security policies and these languages provide different constructs based on underlying access control model when modeling security requirements. In other words, the selection of the policy language and thus the model determine expressiveness in encoding rules and simplicity in administration. Among various models, attribute-based access control (ABAC) [3] has been shown to provide very expressive constructs and various tools have been developed to assist policy specifications with them [14, 13]. In ABAC, access rules are specified by using user and resource attributes. For instance, the "role" of a user in an organization or the "type" of a sensitive resource can be used to identify sets of users/resources in a single rule. The use of attributes not only allows compact encoding of permission assignment to users but also maintains readability.

In order to assist policy administrators when specifying ABAC policies, a particularly useful approach is to infer access control rules from existing logs. Among other information, these logs contain tuples denoting user, sensitive resource, the exercised right and a time-stamp. They may contain access logs that should have not have happened (i.e. under-assignments) in the case of an error in the enforcement or may contain only partial information about the permissions (i.e. over-assignments). To our knowledge, [15] is the first work that discusses policy mining in the context of ABAC. It presents a custom mining algorithm (referred as Xu-Stoller) and its extensive evaluation with both realistic and synthetic policies. However, the approach has two major limitations. On one side it considers only positive authorizations ("who can access to what") whereas many ABAC languages allow also negative authorizations ("who cannot access to what"). The logs may contain denied access requests as well due to auditing purposes which are negative examples in the mining process. On the other side, it provides no support for under-assignment (i.e. noise) case and only partial support for the over-assignment case.

To overcome the aforementioned limitations, this paper is

[1] http://www.projectaccus.eu/

an early report on how deep learning performs for ABAC policy mining in the case of only positive authorizations. The approach has two phases. The first phase consists in generalizing the knowledge from the logs yielding a good set of candidates rules in a binary vector format. The idea is to obtain insight about certain parameters that are significant in obtaining good rules. In the second phase the target will be to transform the set of candidate rules from the binary vector format to the format acceptable by Xu-Stoller and compare them. In this paper, we focus on the first phase, and we make use of the excellent capabilities of Restricted Boltzmann Machines (RBMs) as density estimators to propose a technique that is capable to produce a set of suitable candidate rules based on the knowledge extracted by the RBMs from the processed logs. More exactly, in our previous work we showed that the reconstruction error (i.e. the difference between a data point and its *reconstruction* made by the RBM model) of various type of RBMs may be used as an excellent similarity measure to find the closest clusters of Markov Decision Processes (MDPs) and we demonstrate the advantage of using it in transfer learning [1], or to assess in an objective manner the quality of impaired images [7, 9] (i.e. to estimate automatically how humans would perceive the degradation level of impaired images in comparison with their unimpaired version). Herein, we show that by using the generative power of RBMs and the previous mentioned reconstruction error, we may generalize and create a set of good candidates rules from a small amount of ABAC logs.

The remaining of this paper is organized as follows. Section 2 provides the problem definition and insights on ABAC policy mining for the benefits of the non-specialist reader, while Section 3 details our proposed method. Section 4 shows a preliminary evaluation of our approach, while Section 5 concludes the paper and presents further research directions.

2. PROBLEM DEFINITION

Given a set of users (U), resources (R) and operations (O) that users can perform on resources, an attribute-based access control system contains two types of attributes A_u and A_r for users and resources respectively. An attribute a of a user or resource x may have an empty value (denoted $\perp$) or a set of values from its domain D_a and this value is denoted by the attribute assignment relation $d(x, a)$.

An ABAC rule $\langle e, o\rangle$ specifies an expression that determines its applicability[2] and an operation. For instance, a rule

$$\langle role \in \{nurse, doctor\} \wedge resource \in \{PatRec\}, read\rangle$$

imposes that only nurses or doctors can read a patient record. For simplicity of presentation, an expression is considered to be a total mapping, $e : A \mapsto \{2^{D_A} \setminus \{\perp\}, \top\}$, that maps each attribute a to either a subset of values from its domain (D_a) or to a special value $\top$. The value $\top$ means that the expression does not specify any constraint on the value of the attribute. If the triple $\langle u, r, o\rangle$ denotes a request by user u to perform operation o on resource r and e is an expression then for each attribute a in A_u or A_r either $e(a) = \top$ or $(d(u, a) \subseteq e(a) \wedge d(r, a) \subseteq e(a))$ holds in order for e to be satisfied.

[2]We only consider positive authorizations in this paper and the order of rules in the policy does not matter.

An ABAC policy is a sequence of rules $\{r_1, \ldots, r_n\}$ that encodes authorizations of users over resources. The access control system receives a request, evaluates it against the policy and logs the permitted requests with a time stamp. Thus the problem we consider in this paper is formulated as follows:

Definition 2.1 (ABAC Policy Mining Problem). *Given a log L where each entry is of the form $\langle u, r, o, t\rangle$ denoting the fact that user u performed o on resource r at time t, users U, resources R, operations O, attributes A and attribute assignment relation d, an ABAC policy mining problem amount to finding a policy that maximizes a policy quality metric.*

There are various policy quality metrics that help to compare policies. An example metric, which is also considered in our work, is weighted structured complexity (WSC) [10] where the expression and the operation in a rule are given weights. More specifically, the complexity of a rule $\langle e, o\rangle$ is calculated as $WSC(e) + WSC(o)$ where the WSC of an expression or an operation is given as the number of atomic elements contained in them. The overall WSC of a policy is the sum of WSCs of all its elements.

Noise and Incompleteness in the Log. There are certain types of problems that a policy mining approach must handle. A log L may contain permissions that are not supposed to be granted (under-assignment). Another problem is the completeness of the log. A log may lack certain permissions that have not been exercised (over-assignments).

3. APPROACH

In this section, firstly, we present the mathematical details of RBMs, and secondly, we describe in details the proposed approach.

3.1 Restricted Boltzmann Machines

Restricted Boltzmann Machines(RBMs) [11] are energy-based models capable to perform unsupervised learning. These models are stochastic with stochastic nodes and layers, making them less vulnerable to local minima [12]. Furthermore, due to their architecture and neural configurations, RBMs and their variants posses excellent generalization capabilities [2, 8].

Formally, an RBM has a visible binary layer $\mathbf{v} = [v_1, .., v_{n_v}]$, where $\forall i, v_i \in \{0, 1\}$ and v_{n_v} represents the total number of visible neurons and a hidden binary layer $\mathbf{h} = [h_1, .., h_{n_h}]$, where $\forall j, h_j \in \{0, 1\}$ and v_{n_h} is the total number of hidden neurons, as shown in Fig. 1. The visible layer encodes the data directly, while the hidden one increases the learning capacity by enlarging the class of distributions that can be represented to an arbitrary complexity. The total energy (i.e. a cost function over all neurons and connections between them) of an RBM is given by equation 1.

$$E(v, h) = -\sum_{i,j} v_i h_j W_{ij} - \sum_i v_i a_i - \sum_j h_j b_j \qquad (1)$$

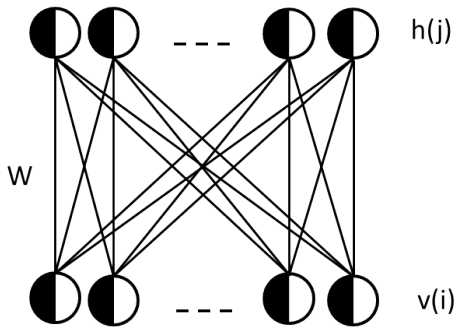


Figure 1: RBM general architecture, where $\mathbf{v}(i)$, $\mathbf{h}(j)$, and $\mathbf{W}$ represent the visible neurons, hidden neurons and the undirected connections between the neurons from different layers, respectively.

where W_{ij} denotes the connection between visible neuron i and hidden neuron j, a_i is the bias for visible neuron i and b_j is the bias for hidden neuron j. The term $\sum_{i,j} v_i h_j W_{ij}$ represents the total energy between neurons from different layers, $\sum_i v_i a_i$ represents the energy of the visible layer and $\sum_j h_j b_j$ the energy of the hidden layer. The inference for the hidden neuron j is done by sampling from a sigmoid function, $p(h_j = 1|\mathbf{v}, \mathbf{W}, \mathbf{b}) = sigm(b_j + \sum_i v_i W_{ij})$. The inference for the visible unit i is done by sampling also from a sigmoid function, $p(v_i = 1|\mathbf{h}, \mathbf{W}, \mathbf{a}) = sigm(a_i + \sum_j h_j W_{ij})$.

In order to maximize the likelihood of the model, the gradients of the energy function with respect to the parameters (i.e. weights, biases) have to be calculated. Unfortunately, in all types of RBMs the maximum likelihood can not be straightforwardly applied due to intractability problems. As a solution to these problems, Contrastive Divergence (CD) algorithm to train RBMs, was introduced by Geoffrey Hinton in [6].

In Contrastive Divergence, learning follows the gradient of:

$$CD_n \propto D_{KL}(p_0(\mathbf{x})||p_\infty(\mathbf{x})) - D_{KL}(p_n(\mathbf{x})||p_\infty(\mathbf{x})) \quad (2)$$

where, $p_n(.)$ is the resulting distribution of a Markov chain running for n steps, and $D_{KL}(.||.)$ represents the Kullback-Leibler divergence. To find the update rules for the parameters of RBMs we have to calculate the derivatives of the energy function from equation 1 with respect to those parameters (i.e. $\mathbf{W}$, $\mathbf{a}$ and $\mathbf{b}$). Since the visible units are conditionally independent given the hidden units and vice versa, learning can be performed using one step Gibbs sampling, which practically has two half-steps: (1) update all the hidden units, and (2) update all the visible units. Thus, in CD_n the weight updates are done as follows: $W_{ij}^{\tau+1} = W_{ij}^{\tau} + \alpha \left(\left\langle \langle h_j v_i \rangle_{p(\mathbf{h}|\mathbf{v};\mathbf{W})} \right\rangle_0 - \langle h_j v_i \rangle_n \right)$ where τ is the training epoch, α is the learning rate, and the subscript $_{(n)}$ indicates that the states are obtained after n iterations of Gibbs sampling from the Markov chain starting at $p_0(\cdot)$. For a more comprehensive discussion about RBM and CD, the interested reader is referred to [2].

3.2 Policy mining procedure

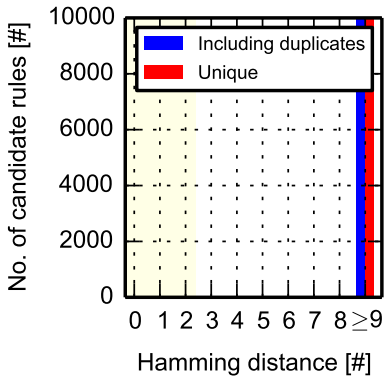
To obtain the generalized set of the ABAC candidate rules, the first step is to represent each entry of the logs L in a binary vector $\mathbf{l}$ to be easy understandable by RBMs. Let L^b denote this transformed set of binary logs. The elements of each binary vector obtained $\mathbf{l} \in L^b$ will have the following unveiled meaning $[l^1_{A^1_u}, .., l^{|D^1_u|}_{A^1_u}, l^1_{A^2_u}, .., l^{|D^2_u|}_{A^2_u}, .., l^1_{A^{|A_u|}_u}, .., l^{|D^{|A_u|}_u|}_{A^{|A_u|}_u}, l^1_{A^1_r}, .., l^{|D^1_r|}_{A^1_r}, l^1_{A^2_r}, .., l^{|D^2_r|}_{A^2_r}, .., l^1_{A^{|A_r|}_r}, .., l^{|D^{|A_r|}_r|}_{A^{|A_r|}_r}, l^1_O, .., l^{|O|}_O]$, where $l^1_{A^i_u}, .., l^{|D^i_u|}_{A^i_u}$, $\forall i$, represent each discrete value from the domain (D^i_u) of user attribute A^i_u, and $|A_u|$ is the total number of user attributes, $l^1_{A^i_r}, .., l^{|D^i_r|}_{A^i_r}, \forall i$, represent each discrete value from the domain (D^i_r) the resource attribute A^i_r, and $|A_r|$ is the total number of resource attributes, while $l^i_O, \forall i$, represent all the possible discrete values of the operations set and $|O|$ is the total amount of operations. Furthermore, for each binary vector $\mathbf{l} \in L^b$ each element $l_i \in \mathbf{l}, \forall i$ is set to 1 if the original entry in the logs L contains the corresponding attribute (or operation) value and it is set to 0 otherwise.

```
1  %% initialization;
2  Transform logs L in binary logs L^b;
3  Initialize RBM.W, RBM.a, RBM.b ← N(0, 0.01), learning rate;
4  Train RBM on L^b;
5  %% get the set of hidden configurations H^{L^b} from the binary logs;
6  %% find the maximum reconstruction error mHD on the binary logs;
7  Initialize the set H^{L^b} empty;
8  mHD=0;
9  for each l ∈ L^b do
10     RBM.v=l;
11     RBM.h=RBM.inferHiddenLayer(RBM.v, RBM.W, RBM.a, RBM.b);
12     Add RBM.h to H^{L^b};
13     RBM.v=RBM.inferVisibleLayer(RBM.h, RBM.W, RBM.a, RBM.b);
14     mHD=max(mHD,computeHammingDist(l,RBM.v));
15 end
16 %% generate the set of good candidate rules R^b with the RBM;
17 Initialize the set of candidate rules R^b empty;
18 for all trials do
19     select h randomly from H^{L^b};
20     RBM.h=h+N(0, σ); % σ = 0.7 seems to be a good choice (Section 4);
21     RBM.v=RBM.inferVisibleLayer(RBM.h, RBM.W, RBM.a, RBM.b);
22     possibleRule=RBM.v;
23     RBM.h=RBM.inferHiddenLayer(RBM.v, RBM.W, RBM.a, RBM.b);
24     RBM.v=RBM.inferVisibleLayer(RBM.h, RBM.W, RBM.a, RBM.b);
25     if (possibleRule ∉ R^b) then
26         if (computeHammingDist(possibleRule,RBM.v)≤mHD) then
27             Add possibleRule to R^b
28         end
29     end
30 end
```

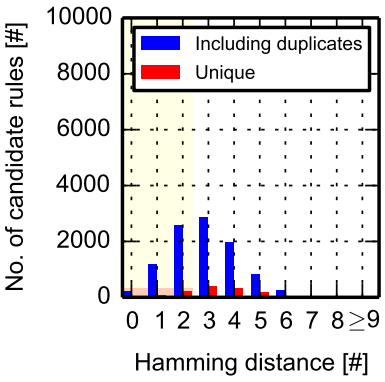
Algorithm 1: High level pseudo-code to generate the set of good candidate rules. The rules are generated from the true probabilities of the hidden neurons (inferred from the binary logs L^b) altered by a Gaussian noise $\mathcal{N}(0, \sigma)$.

The second step consists in generating good candidate rules using RBMs capabilities, as briefly explained in Alg. 1 and detailed below. Firstly, the set of binary logs is used to train an RBM (Alg. 1, lines 3-4). As a consequence, the trained RBM will learn to incorporate well the hidden distribution of the logs. The straight forward solution to obtain the set of good candidates rules from the trained RBM would be: (1) compute the maximum reconstruction error[3] obtained by any of the binary logs L^b passed through the trained RBM, and set it as a threshold (Alg. 1, lines 8-15); (2) pass all the possible combination of rules through the RBM and

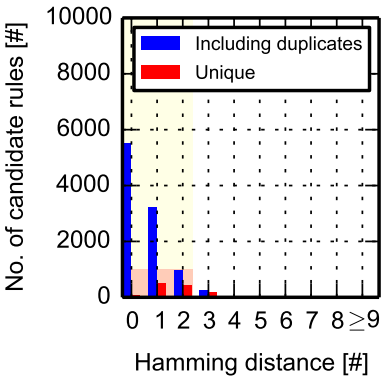
[3]For any data point the reconstruction error is represented by the difference between the data point under scrutiny and the expected values of the visible neurons, inferred from the expected values of the hidden ones, which initially are inferred from the visible neurons set to the data point itself.



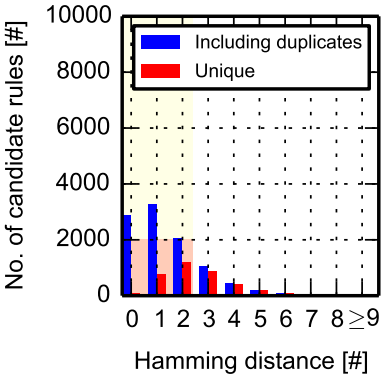
(a) Randomly generated candidate rules.



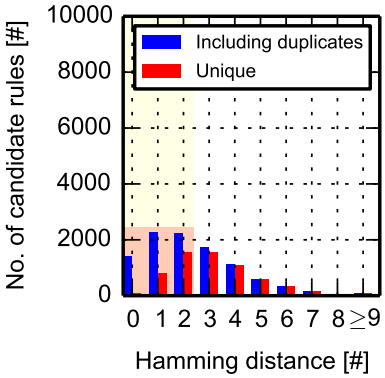
(b) Candidate rules generated directly from random generated probabilities of the hidden neurons.



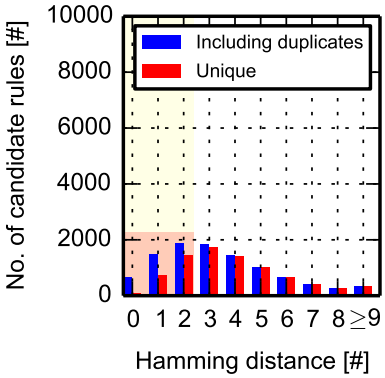
(c) Candidate rules generated from the hidden neurons set to the true probabilities of the hidden layer inferred from randomly chosen binary logs L^b plus noise $\mathcal{N}(0, 0.3)$.



(d) Candidate rules generated from the hidden neurons set to the true probabilities of the hidden layer inferred from randomly chosen binary logs L^b plus noise $\mathcal{N}(0, 0.5)$.



(e) Candidate rules generated from the hidden neurons set to the true probabilities of the hidden layer inferred from randomly chosen binary logs L^b plus noise $\mathcal{N}(0, 0.7)$.



(f) Candidate rules generated from the hidden neurons set to the true probabilities of the hidden layer inferred from randomly chosen binary logs L^b plus noise $\mathcal{N}(0, 0.9)$.

Figure 2: Histograms of rules generation successful rate. For each plot we made 10000 trials to generate rules. The light yellow area represents the target zone, while the light red area represents the number of unique good candidate rules for the method under scrutiny. The blue bar represents the total number of generated rules for which the reconstruction error is expressed by the x-axis, while the red bar shows the total number of unique generated rules for each possible value of the reconstruction error (i.e. x-axis). The highest number of good unique candidate rules is generated by the method from subplot (e).

consider as good candidates rules those ones for which the reconstruction error is smaller or equal with the previous computed threshold. For the sake of brevity, for a better understanding on how the reconstruction error may reflect if some data points belong to the hidden distribution incorporated by an RBM type model, the interested reader is referred to [1, 7]. However, due to the fact that the total number of possible combination of rules is 2^{n_v} (i.e. exponential with the number of visible neurons), enumeration of all candidate rules does not represent a feasible solution. As an alternative, we propose to make use of the generative properties of RBMs and to sample rules from the trained RBM in a controlled manner (Alg. 1, lines 19-22). This yields to a reduced amount of rules drawn from a distribution close enough to the one incorporated in the trained RBM, and which can be assessed further-on using the reconstruction error procedure (Alg. 1, lines 23-29). To sample in a controlled manner from a trained RBM, one can start from a specific configuration of the hidden neurons to infer the values of the visible ones. Depending on how the configuration of the hidden neurons is chosen the amount of generate rules may differ, as we will show in Section 4.

4. PRELIMINARY EVALUATION

In this section, we evaluate the effectiveness of the proposed approach to generate a set of good candidates rules. For this, we make use of the *healthcare* dataset from [15]. On this specific dataset, after we transformed the original logs in binary logs we obtained a vector of 46 binary values to cover all user attributes, resource attributes, and operations. Thus, in the RBM we set the number of visible neurons to 46, and the number of hidden neurons to 40 to ensure enough representational power. The learning rate was set to 0.001 and we trained the RBM model until it converges (i.e. 200 training epochs). After training, the RBM obtained was used to generate new candidates rules. In all experiments, we used the Hamming distance [5] to measure the reconstruction error. The biggest reconstruction error obtained by passing any of the binary logs $\mathbf{l} \in L^b$ through the trained RBM was 2.

To assess which is the best controlled manner to generate rules with a high chance of belonging to the same distri-

bution incorporated in the trained RBM, we have considered 6 slightly different scenarios, as depicted in Fig. 2. In one scenario we have generated rules completed randomly, while in the other 5 the rules were generated from the true probabilities of the hidden neurons (inferred from the binary logs L^b) altered by a Gaussian noise $\mathcal{N}(0, \sigma)$ with a scenario specific standard deviation, while keeping the constraint $0 \leq h_j \leq 1, \forall j$. For each scenario, we have made 10000 trials to generate rules. Because different trials may yield duplicate rules, the goal was to find the method which is capable to generate the highest number of unique rules which passed through the trained RBM have a reconstruction error (i.e. Hamming distance) smaller or equal with 2. As it is reflected in Fig. 2e, by setting the values of the hidden neurons in the generative phase to the true probabilities of the hidden layer, inferred with the trained RBM from randomly chosen binary logs, and adding them a random Gaussian noise with a standard deviation of 0.7, yields the highest amount of unique good rules (i.e. 2423). Remarkable, in the extreme case of generating rules completely randomly (i.e. Fig. 2a), none of them was successful.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we reported our initial findings about the use of RBM for ABAC policy mining when there are only positive examples. Our overall idea is to exploit certain features (e.g. inductive bias [4]) offered by neural networks to overcome limitations of existing approaches to policy mining problem in particular when the logs contain negative authorizations or large amount of noise with varying levels of incompleteness. As further work, we intend to implement the second phase of our approach where we compare the prediction accuracy of our technique to Xu-Stoller according to different policy quality metrics. This will involve evaluations with different real-world policies and supporting more complex policies that contain expressions of different types.

6. ACKNOWLEDGMENTS

This work has been supported by ARTEMIS project ACCUS (Adaptive Cooperative Control in Urban sub-Systems), Grant agreement n. 333020 (http:/projectaccus.eu/).

7. REFERENCES

[1] H. B. Ammar, E. Eaton, M. E. Taylor, D. C. Mocanu, K. Driessens, G. Weiss, and K. Tuyls. An automated measure of mdp similarity for transfer in reinforcement learning. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[2] Y. Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

[3] O. X. T. Committee, 2013. eXtensible Access Control Markup Language (XACML).

[4] M. Craven and J. W. Shavlik. Using neural networks for data mining. *Future Generation Comp. Syst.*, 13(2-3):211–229, 1997.

[5] R. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 26(2):147–160, 1950.

[6] G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, Aug. 2002.

[7] D. Mocanu, G. Exarchakos, H. Ammar, and A. Liotta. Reduced reference image quality assessment via boltzmann machines. In *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pages 1278–1281, May 2015.

[8] D. C. Mocanu, H. Bou-Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, and K. Tuyls. Factored four way conditional restricted boltzmann machines for activity recognition. *Pattern Recognition Letters*, 2015.

[9] D. C. Mocanu, G. Exarchakos, and A. Liotta. Deep learning for objective quality assessment of 3d images. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 758–762, Oct 2014.

[10] I. Molloy, Y. Park, and S. Chari. Generative models for access control policies: applications to role mining over logs with attribution. In *17th Symposium on Access Control Models and Technologies (SACMAT)*, pages 45–56, 2012.

[11] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, et al., editors, *Parallel Distributed Processing: Volume 1: Foundations*, pages 194–281. MIT Press, Cambridge, 1987.

[12] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Two distributed-state models for generating high-dimensional time series. *Journal of Machine Learning Research*, 12:1025–1068, 2011.

[13] F. Turkmen, J. den Hartog, S. Ranise, and N. Zannone. Analysis of XACML policies with SMT. In *4th International Conference on Principles of Security and Trust - (POST) Proceedings*, pages 115–134, 2015.

[14] F. Turkmen, S. N. Foley, B. O'Sullivan, W. M. Fitzgerald, T. Hadzic, S. Basagiannis, and M. Boubekeur. Explanations and relaxations for policy conflicts in physical access control. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, November 4-6, 2013*, pages 330–336, 2013.

[15] Z. Xu and S. D. Stoller. Mining attribute-based access control policies from logs. In *Data and Applications Security and Privacy XXVIII - 28th Annual IFIP WG 11.3 Working Conference, DBSec 2014, Vienna, Austria, July 14-16, 2014. Proceedings*, pages 276–291, 2014.

Indeks avtorjev / Author index

Konferenca / Conference

Uredili / Edited by

Inteligentni sistemi / Intelligent Systems

Rok Piltaver, Matjaž Gams