Risk stratification of cardiac patients by utilizing additional unlabeled examples

Aleš Papič
University of Ljubljana
Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
ales.papic@fri.uni-lj.si

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

ABSTRACT

In the paper we address the challenge of cardiac patient risk stratification using the additional unlabeled data. The motivation for using unlabeled data comes from the field of semi-supervised learning (SSL), which has shown that additional unlabeled data can improve accuracy of supervised learning models. In addition to traditional SSL, we propose three new approaches that are based on active learning (AL), fuzzy learning (FL), and supervised clustering (SC). We evaluate them on the UCI ML heart disease dataset and with four different classification models. The results show that our approaches increase the inductive performance compared to the learning algorithms trained exclusively on labeled data. The most favorable performance was achieved with the fuzzy learning approach that utilizes a reliability estimate for selection of the most beneficial additional examples.

Keywords

risk stratification, cardiovascular diseases, machine learning, knowledge transfer, semi-supervised learning, active learning, fuzzy learning, supervised clustering, unlabeled examples

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of morbidity and death worldwide, together with cancer and chronic respiratory diseases. To prevent them, people with increased risk need early identification and medical guidance. In the past two decades, researchers have invested a lot of effort to develop clinical decision support systems for risk assessment of CVDs, but only some are included in clinical guidelines [2].

In this paper we tackle the issue of developing a patient risk stratification model, which classifies patients into levels for having a serious cardiac event. To improve the predictive performance we utilize ideas from the field of semi-supervised learning, which have shown that utilization of additional unlabeled data can improve accuracy of supervised learning models. Applying unlabeled data has advantages, such as not relying on expertise to label examples, which saves time, effort and reduces cost [12].

To perform knowledge transfer from unlabeled examples to supervised learning, we apply our implementation of the following four approaches: traditional semi-supervised learning as the baseline (SSL), active learning (AL), fuzzy learning (FL), and supervised clustering (SC). In the latter three approaches we pay special attention on how to perform the *instance knowledge transfer* to properly *select the right examples for the training data.*

The paper is structured as follows. Related work is presented in Section 2. Our approaches are described in Section 3. The evaluation and results are given in Section 4. The last section concludes the paper and gives directions for future work.

2. RELATED WORK

The rapid development of predictive models for cardiac and cardiovascular disease diagnostics happened between years 2000 and 2013 [6]. In European guidelines, Systematic COronary Risk Evaluation (SCORE) estimates the ten-year risk for a fatal cardiovascular event, such as heart attack, stroke, aneurysm of the aorta, by stratifying patients into four risk groups: low, moderate, high, and very high. Different databases contain different risk level definitions. To facilitate an initial approach to the problem, we have chosen a simpler public Cleveland database with two levels of risk. The most commonly used machine learning (ML) algorithms for heart disease diagnosis are: Support Vector Machine (SVM), Naive Bayesian classifier (Naive Bayes), Artificial Neural Network (ANN) and Decision tree (DT). Parthiban et al. [10] used SVM with RBF kernel and Naive Bayesian classifier for diagnosis of heart disease in diabetic patients. The accuracy of their approach was 94.6% and 74%, respectively. Dangare et al. [3] propose using ANN with an extended feature set for heart diseases. They included information about obesity and smoking as the risk factors for coronary heart disease. Das et al. [4] used the ensemble approach of three ANN with a tangent sigmoid function, single hidden layer, and 14 neurons. The experimental results gained 89% classification accuracy for heart disease diagnosis.

Lately, knowledge transfer (transfer learning) has become popular in the field of machine learning [9, 7]. The reasons for transferring knowledge are often associated with a lack of learning data for the target problem or with the time it takes to learn a new model. Partially labeled training data have shown to improve performance in machine learning [8]. Such data are often also easier and cheaper to obtain. Knowledge transfer approaches are also found in different medical fields: pneumonia risk assessment with multi-task learning, lifelong inductive learning in the field of heart disease and sequential inductive a model for knowledge transfer in the field of coronary artery disease diagnosis [7]. A problem that can occur during the transfer of knowledge is the so-called *neg-*

ative transfer, which harms the learning success for the new domain [9].

3. UTILIZATION OF UNLABELED EXAMPLES

In this section we present four approaches for learning from partially labeled data using knowledge transfer. Each approach uses a learning algorithm to derive the knowledge from a small portion of labeled data. This knowledge is then used to classify unlabeled examples, which afterwards supplement the original training data set to increase the final prediction accuracy.

3.1 Semi-supervised learning

A well-known approach of SSL is called self-learning [12]. It first trains on labeled examples, then classifies the unlabeled examples and combines the latter with the initially labeled data. The extended dataset is used for further supervised learning. This method allows us to build on top of it and serves as a baseline in our experiments.

3.2 Active learning

Active learning (AL) can similarly work with partially labeled data [11]. Its main goal is to find unlabeled examples that have the greatest potential to improve performance and present them to the teacher (oracle) who does the annotation process. The labeling of example is thus done iteratively rather than for all unlabeled dataset at once.

When selecting examples, we desire such that are labeled the most reliably. To estimate this reliability, we apply two metrics: (1) posterior class probability and (2) local modeling of prediction error with estimate CNK [1]. The posterior class probability for a given an example is provided by the learning algorithm. The reliability measure CNK estimates the reliability of the prediction by observing the local prediction error. In this work we adapt the original CNK estimate (designed for regression) for classification and compute it for each query example as:

$$CNK = \frac{\sum_i^k P_R(C_i)}{k}$$

where where C_i is an example from the local neighborhood $\{C_1, ..., C_k\}$ of our query example, and $P_R(C_i)$ is the posterior probability that the neighbor is classified into R, which is the class into which our query example is also classified. To summarize, CNK measures the average posterior probability for the classification into query example's class within its neighborhood. Such CNK estimate is defined on the interval $[0, 1]$, where 0 or 1 indicate unreliable or reliable classifications, respectively. In our experiments, we applied the size of the neighborhood $k = 5$, as used in the authors' original work.

The algorithm stops when any of the following three stopping criteria is reached. The first criterion defines the maximum number of iterations (N), which can be useful for large data sets. The second criterion stops the algorithm when there are no examples with reliable classifications. The third criterion is fulfilled when all the unlabeled examples have been utilized.

3.3 Fuzzy learning

Our Fuzzy learning (FL) approach labels examples with probabilities for belonging to all possible classes. We further use these probabilities to assign class probabilities to unlabeled examples as weighted class probabilities of the nearest neighbors using the locally weighted regression (LWR). We observe each class separately and assign a fuzzy class probability to each unlabeled example. A fuzzy class probability is derived from the class probabilities of local neighbors, which are weighted with the distance to the observed example and then summed up.

The weighted probabilities are afterwards calibrated to scale up to 1, to ensure probabilistic interpretation. We use measures of reliability, such as posterior class probability and local modeling of prediction error (CNK, as already described), to select examples which we include into the training set across multiple iterations. Finally, the learning algorithm is trained on the combined training set.

The algorithm stops either when a maximum number of iterations is reached, it runs out of data, or when labeling is not reliable enough to extend the training set.

3.4 Supervised clustering

Supervised clustering (SC) differs from classical clustering methods by considering class values during the clustering process [13]. Our approach looks for representative examples in the available data. We assign each example a class of the closest representative. All examples are then used to train the learning algorithm.

To find the representatives, we apply the iterative SRIDHCR algorithm [5]. The algorithm first constructs a random set of representatives, which represents the current solution. In each iteration, a single non-representative is added and another single representative is removed, generating two new candidate sets of representatives. Next, the algorithm evaluates each generated candidate set X using the fitness function $q(X)$:

$$q(X) = impurity(X) + \beta * \begin{cases} \sqrt{\frac{|K|-c}{N}} & |K| > c \\ 0 & |K| \leq c \end{cases}$$

which minimizes cluster impurity and punishes large number of clusters. The punishment is controlled with the input parameter β and takes effect if the number of clusters |K| is higher than the number of classes c. The candidate set, which improved the current best solution, is saved. In the next iteration, the procedure repeats itself until the algorithm cannot find a better set of representatives. The algorithm also utilizes a parameter S_{size}, which controls the number of candidate sets generated in each iteration. A higher value increases the probability of finding a better set of representatives because the algorithm performs more permutations.

Finally, we use the set of representatives from SRIDHCR to label the unlabeled data using the nearest neighbor approach. The final model is then trained on the combined data.

4. EVALUATION AND RESULTS

We evaluated our approaches on the Cleveland heart disease data from the UCI ML repository. The dataset has 297 patients of which 54% belong to low-risk (healthy) class and 46% to high-risk class. Two thirds of of patients are male with the average age of 54 years, and the remaining are female with the average age of about 56 years. We measured the performance using the Area Under a ROC Curve (AUC), which summarizes the overall performance of the model and reflects the discriminating ability to diagnose patients with and without the disease.

At the beginning of the evaluation process, we randomized the data. Next, the data was split using the 5-fold cross-validation into training and test sets. Since our experimental data set does not contain unlabeled examples, we split the training set into the labeled and unlabeled set (simulated, by hiding examples' classes). The ratio between the labeled and unlabeled set is controlled with an input parameter. In the experiments, we limited the AL and FL to 10 iterations and set their threshold to select examples with reliability at least 0.8. The SC generates 10 candidate sets for the representatives. The penalty (β) is set to 1.0 to prevent large numbers of representatives. The `SRIDHCR` algorithm performs restarts 30 times during the search process.

We used four different learning algorithms – Decision Tree (DT, using information gain and minimum number of 20 examples in leaves), K-Nearest Neighbors (KNN, with $k = 5$ and using Euclidean distance), Naive Bayesian classifier (NB) and Support Vector Machine (SVM, with linear kernel, regularization weight of 1 and termination criterion of 0.001). For each combination of approach and the learning algorithm we computed the *transductive* and *inductive* performance. The former reflects the quality of *transfer learning* by measuring labeling accuracy only of unlabeled examples prior their inclusion into the original learning data set. The inductive performance measures the final performance of the model that was built on the extended data set.

The transductive performance is shown in Table 1 and the inductive performance in Table 2. The first column displayes the percentage of labeled examples that were used in the experiments. Approaches AL and FL are displayed twice (with the use of posterior probability and with the use of the CNK estimate). The results of the transductive analysis show that on the average FL with CNK selection method obtains the highest AUC. Using 20% initially labeled examples, the obtained AUC is equal to 0.85 ± 0.03 and increases to 0.90 ± 0.05, with 80% of labeled examples. The AL approach shows low-averaged performance in combination with SVM, which in some cases learned to predict the opposite classes. The SC obtained the lowest transductive performance of 0.65 ± 0.14 on 20% and 0.75 ± 0.13 on 80% of labeled examples.

The inductive evaluation resembles the results of the transductive evaluation. The results of all approaches have decreased compared to transductive for about 0.05. FL obtained the best results followed by AL. The significant difference compared to the transductive results can be seen for SC. The predictive models obtain comparable performance, even though many misclassified examples were introduced. Using the 80% of initially labeled examples, the SC obtained AUC of 0.81 ± 0.05, which is equal to the other approaches.

5. CONCLUSIONS

We experimented with four different approaches for including unlabeled examples into risk stratification. The obtained results are comparable to the results in related works. The utilization of additional examples shows promising results, especially with the fuzzy learning approach that utilizes reliability estimate CNK.

In the future, we shall evaluate the methodology on databases with more complex risk levels. Secondly, we shall also analyze the performance of supervised learning algorithms and impact of their parameters. Thirdly, neural networks and deep learning are opening promising directions also in medical problems. Due to the limitation of resources, we did not include them in this work, but shall also include them in the work to follow.

6. REFERENCES

[1] Z. Bosnić and I. Kononenko. Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering*, 67(3):504–516, 2008.

[2] J. A. Damen, L. Hooft, E. Schuit, T. P. Debray, G. S. Collins, I. Tzoulaki, C. M. Lassale, G. C. Siontis, V. Chiocchia, C. Roberts, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *bmj*, 353:i2416, 2016.

[3] C. S. Dangare and S. S. Apte. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10):44–48, 2012.

[4] R. Das, I. Turkoglu, and A. Sengur. Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, 36(4):7675–7680, 2009.

[5] C. F. Eick, N. Zeidat, and Z. Zhao. Supervised clustering-algorithms and benefits. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 774–776. IEEE, 2004.

[6] M. Fatima and M. Pasha. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

[7] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[8] N. Nguyen and R. Caruana. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–559. ACM, 2008.

[9] S. J. Pan and Q. Yang. A survey on transfer learning. 10(22):1345–1359, 2010.

[10] G. Parthiban and S. Srivatsa. Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, 3:2249–0868, 2012.

[11] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[12] I. Triguero, S. García, and F. Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2):245–284, Feb 2015.

[13] N. Zeidat, C. F. Eick, and Z. Zhao. *Supervised clustering: algorithms and applications.* University of Houston, 2005.

LABELS	METHOD	DT	KNN	NB	SVM	$\bar{x}$	p-value
20%	SSL	0.71 ± 0.05	0.80 ± 0.02	0.77 ± 0.06	0.79 ± 0.04	0.77 ± 0.04	-
	AL	$\underline{0.75 \pm 0.06}$	$\underline{0.87 \pm 0.03}$	0.79 ± 0.05	$\underline{0.42 \pm 0.42}$	$\underline{0.70 \pm 0.14}$	0.017
	AL (CNK)	$\underline{0.83 \pm 0.04}$	$\underline{0.82 \pm 0.03}$	$\underline{0.84 \pm 0.03}$	$\underline{0.68 \pm 0.20}$	$\underline{0.79 \pm 0.07}$	< 0.001
	FL	$\underline{0.81 \pm 0.05}$	0.81 ± 0.05	$\underline{0.81 \pm 0.05}$	0.81 ± 0.05	$\underline{0.81 \pm 0.05}$	< 0.001
	FL (CNK)	$\underline{0.85 \pm 0.03}$	$\underline{0.85 \pm 0.03}$	$\underline{0.85 \pm 0.03}$	$\underline{0.85 \pm 0.03}$	$\underline{0.85 \pm 0.03}$	< 0.001
	SC	0.64 ± 0.15	$\underline{0.64 \pm 0.13}$	$\underline{0.67 \pm 0.14}$	$\underline{0.66 \pm 0.13}$	$\underline{0.65 \pm 0.14}$	< 0.001
	$\bar{x}$	0.77 ± 0.06	0.80 ± 0.05	0.79 ± 0.06	0.70 ± 0.15	-	-
50%	SSL	0.75 ± 0.05	0.82 ± 0.02	0.79 ± 0.05	0.81 ± 0.02	0.79 ± 0.03	-
	AL	$\underline{0.82 \pm 0.06}$	$\underline{0.88 \pm 0.03}$	$\underline{0.84 \pm 0.05}$	0.66 ± 0.39	$\underline{0.80 \pm 0.13}$	< 0.001
	AL (CNK)	$\underline{0.87 \pm 0.03}$	$\underline{0.85 \pm 0.03}$	$\underline{0.87 \pm 0.03}$	0.74 ± 0.27	$\underline{0.83 \pm 0.09}$	< 0.001
	FL	$\underline{0.86 \pm 0.03}$	$\underline{0.86 \pm 0.03}$	$\underline{0.86 \pm 0.03}$	$\underline{0.86 \pm 0.03}$	$\underline{0.86 \pm 0.03}$	< 0.001
	FL (CNK)	$\underline{0.88 \pm 0.02}$	$\underline{0.88 \pm 0.02}$	$\underline{0.88 \pm 0.02}$	$\underline{0.88 \pm 0.02}$	$\underline{0.88 \pm 0.02}$	< 0.001
	SC	0.70 ± 0.13	$\underline{0.72 \pm 0.12}$	$\underline{0.73 \pm 0.12}$	$\underline{0.72 \pm 0.13}$	$\underline{0.72 \pm 0.13}$	< 0.001
	$\bar{x}$	0.81 ± 0.06	0.84 ± 0.04	0.83 ± 0.05	0.79 ± 0.14	-	-
80%	SSL	0.77 ± 0.06	0.81 ± 0.05	0.81 ± 0.07	0.82 ± 0.06	0.80 ± 0.06	-
	AL	$\underline{0.82 \pm 0.07}$	$\underline{0.90 \pm 0.04}$	$\underline{0.86 \pm 0.05}$	0.46 ± 0.46	$\underline{0.76 \pm 0.15}$	< 0.001
	AL (CNK)	$\underline{0.88 \pm 0.05}$	$\underline{0.87 \pm 0.05}$	$\underline{0.88 \pm 0.05}$	0.56 ± 0.43	$\underline{0.80 \pm 0.15}$	< 0.001
	FL	$\underline{0.87 \pm 0.06}$	$\underline{0.87 \pm 0.06}$	$\underline{0.87 \pm 0.06}$	$\underline{0.87 \pm 0.06}$	$\underline{0.87 \pm 0.06}$	< 0.001
	FL (CNK)	$\underline{0.90 \pm 0.05}$	$\underline{0.90 \pm 0.05}$	$\underline{0.90 \pm 0.05}$	$\underline{0.90 \pm 0.05}$	$\underline{0.90 \pm 0.05}$	< 0.001
	SC	0.75 ± 0.13	$\underline{0.76 \pm 0.11}$	$\underline{0.72 \pm 0.13}$	0.75 ± 0.16	$\underline{0.75 \pm 0.13}$	0.003
	$\bar{x}$	0.83 ± 0.07	0.85 ± 0.06	0.84 ± 0.07	0.73 ± 0.20	-	-

Table 1: Transductive AUC performance for different percentages of labeled examples, labeling approaches and four classifiers. Statistically significant differences to the baseline (SSL) approach are underlined.

LABELS	METHOD	DT	KNN	NB	SVM	$\bar{x}$	p-value
20%	BASE	0.72 ± 0.07	0.80 ± 0.05	0.78 ± 0.06	0.80 ± 0.06	0.77 ± 0.06	-
	SSL	0.72 ± 0.07	0.80 ± 0.05	0.79 ± 0.06	0.80 ± 0.06	0.77 ± 0.06	0.938
	AL	0.72 ± 0.08	0.83 ± 0.05	$\underline{0.76 \pm 0.06}$	$\underline{0.66 \pm 0.16}$	$\underline{0.74 \pm 0.09}$	0.032
	AL (CNK)	0.72 ± 0.08	0.81 ± 0.05	0.80 ± 0.05	0.80 ± 0.07	0.78 ± 0.06	0.066
	FL	$\underline{0.76 \pm 0.07}$	0.79 ± 0.05	0.77 ± 0.05	0.78 ± 0.06	0.78 ± 0.06	0.881
	FL (CNK)	$\underline{0.78 \pm 0.05}$	0.80 ± 0.04	0.79 ± 0.05	0.80 ± 0.06	$\underline{0.79 \pm 0.05}$	0.016
	SC	$\underline{0.65 \pm 0.15}$	0.67 ± 0.16	$\underline{0.77 \pm 0.08}$	$\underline{0.68 \pm 0.15}$	$\underline{0.69 \pm 0.14}$	< 0.001
	$\bar{x}$	0.72 ± 0.08	0.78 ± 0.06	0.78 ± 0.06	0.76 ± 0.09	-	-
50%	BASE	0.73 ± 0.07	0.80 ± 0.03	0.78 ± 0.06	0.81 ± 0.05	0.78 ± 0.05	-
	SSL	0.73 ± 0.07	0.81 ± 0.04	0.79 ± 0.05	$\underline{0.80 \pm 0.06}$	0.78 ± 0.06	0.394
	AL	0.74 ± 0.08	0.81 ± 0.04	0.79 ± 0.05	0.80 ± 0.05	0.79 ± 0.06	0.993
	AL (CNK)	$\underline{0.76 \pm 0.07}$	0.81 ± 0.05	0.80 ± 0.05	0.81 ± 0.05	0.79 ± 0.05	0.071
	FL	$\underline{0.79 \pm 0.05}$	$\underline{0.77 \pm 0.05}$	$\underline{0.81 \pm 0.05}$	0.82 ± 0.05	0.80 ± 0.05	0.087
	FL (CNK)	$\underline{0.82 \pm 0.04}$	$\underline{0.78 \pm 0.05}$	$\underline{0.81 \pm 0.05}$	0.81 ± 0.05	$\underline{0.81 \pm 0.05}$	0.002
	SC	0.71 ± 0.10	0.77 ± 0.08	0.80 ± 0.05	$\underline{0.75 \pm 0.12}$	0.76 ± 0.09	0.102
	$\bar{x}$	0.75 ± 0.07	0.79 ± 0.05	0.80 ± 0.05	0.80 ± 0.06	-	-
80%	BASE	0.77 ± 0.05	0.83 ± 0.04	0.81 ± 0.05	0.83 ± 0.06	0.81 ± 0.05	-
	SSL	0.77 ± 0.05	0.83 ± 0.04	0.82 ± 0.05	0.82 ± 0.05	0.81 ± 0.05	0.851
	AL	0.78 ± 0.04	0.83 ± 0.04	0.82 ± 0.05	0.82 ± 0.05	0.81 ± 0.05	0.334
	AL (CNK)	0.77 ± 0.04	0.83 ± 0.04	0.82 ± 0.05	0.83 ± 0.05	0.81 ± 0.05	0.680
	FL	$\underline{0.81 \pm 0.05}$	$\underline{0.78 \pm 0.07}$	$\underline{0.83 \pm 0.05}$	0.83 ± 0.06	0.81 ± 0.06	0.651
	FL (CNK)	$\underline{0.80 \pm 0.03}$	$\underline{0.78 \pm 0.05}$	$\underline{0.83 \pm 0.05}$	0.83 ± 0.05	0.81 ± 0.05	0.899
	SC	0.78 ± 0.05	0.83 ± 0.06	0.82 ± 0.06	0.82 ± 0.05	0.81 ± 0.05	0.818
	$\bar{x}$	0.78 ± 0.05	0.82 ± 0.05	0.82 ± 0.05	0.83 ± 0.05	-	-
100%	BASE	0.79 ± 0.04	0.82 ± 0.04	0.82 ± 0.05	0.83 ± 0.03	0.82 ± 0.04	-

Table 2: Inductive AUC performance for different percentages of labeled examples, labeling approaches and four classifiers. Statistically significant differences to the baseline (BASE - a model trained on initially labeled data) approach are underlined.