A framework for machine learning of surrogate models with an application to Sentinel 5P

Michał Artur Szlupowicz
m.szlupowicz@gmail.com
Warsaw University of Technology
Faculty of Physic
plac Politechniki 1
Warsaw, Poland

Jure Brence
jure.brence@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Sašo Džeroski
saso.dzeroski@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

Surrogate models are computationally efficient approximations of computationally expensive simulations or models. In this paper we report improvements of a framework for learning surrogates on input and output spaces with reduced dimensionality. We present nonlinear embeddings and feature importance as additional methods for dimensional analysis and reduction. The choice of models for prediction is extended with two types of ensembles of decision trees. The performance of the additions is evaluated and compared with the original approaches on a dataset, generated by RemoTeC, a complex radiative transfer model.

KEYWORDS

spectral data, neural network, ensemble, surrogate model, dimensionality reduction

1 INTRODUCTION

The **TROPO**spheric **M**onitoring **I**nstrument (TROPOMI) is an on board satellite instrument on the Copernicus Sentinel-5 Precursor satellite [9]. Its main objective is to provide accurate observations of atmospheric parameters, as the concentrations of atmospheric constituents. Those can be used to obtain better air quality forecasts and to monitor global trends. However, the retrieval of interesting attributes involves running a retrieval algorithm, such as RemoTeC [2, 8], based on "optimal estimation methods" that tend to be computationally very expensive [7].

Machine learning techniques can be used to learn surrogate models that approximate the outputs of intensive simulations and are much faster at making predictions [13]. A framework for learning surrogates of radiative transfer models has been developed [1]. Due to the high dimensionality of both input and output spaces, the framework employs dimensionality reduction - methods that find low-dimenensional projections (embeddings) of data that preserve as much information as possible [4]. Predictive models are learned on input and output spaces with reduced dimensionality.

Despite promising results, the existing framework for learning surrogates is limited to simple feed-forward neural networks for the task of prediction, while offering a choice between PCA and autoencoders to reduce dimensionality [4, 6, 3]. In this paper we present an extension of the framework with two types of ensembles of decision trees for prediction [4], as well as an evaluation of the performance and utility of three additional algorithms for dimension analysis and dimension reduction: t-SNE [11], UMAP [12] and feature importance based on random forests [10].

The RemoTeC dataset is described in section 2, followed by a presentation of the involved algorithms for prediction and dimensionality reduction in section 3. The experiment, described in section 4, consists of dimension analysis and predictive performance comparisons. The impact of the results, as well as further work, are discussed in section 5.

Information Society 2020, 5–9 October, 2020, Ljubljana, Slovenia

2 DATASET

The training dataset was generated using RemoTeC tool and in total consists of 50000 samples. Each input state vector contains a set of atmospheric parameters: **s**olar **z**enith **a**ngle (SZA), albedo, temperature, pressure, aerosols and profiles of CH4, CO, H20 gases (in total 125 dimensions). The sampling of the data ensures that the data covers the entire range of conditions that S5P/TROPOMI is expected to encounter. Exploratory data analysis reveals three dimensions with zero variance. Removing them results in a dataset with a 122-dimensional input space.

The output training data was created using the RemoTeC RTM in the S5P/TROPOMI Shortwave InfraRed (SWIR3) band. Each target vector consists of an infrared spectrum with 834 dimensions.

3 SURROGATE MODELS

The framework for learning surrogates is capable of learning both forward and backwards models. The former predict spectra, given atmospheric parameters. The latter reverse this process and learn to approximate atmospheric parameters that produce a given spectra, which is useful for optimizing parameters of the RemoTeC simulation. Surrogates are generally predictive models that map between input and output data of a simulation or computationally expensive model. They offer much faster predictions at the cost incurring a prediction error. However, when the data is high dimensional and contains many samples, the computational cost of training and prediction can still be non-trivial. In such cases methods of dimensionality reduction can offer not only time savings, but also improvements in predictive performance. In our framework we employ dimensionality reduction to atmospheric parameters, as well as the spectral space. Predictive models learn to map between reduced spaces. An inversion of dimensionality reduction is performed on predictions in the reduced space to obtain predictions in the original output space. For that reason dimensionality reduction algorithms must provide an inverse transformation in order to be useful as a component of a surrogate model in our framework.

3.1 Dimensionality reduction

A high number of dimensions makes a problem much harder for many machine learning algorithms due to the curse of dimensionality. For this reason we have tried a range of dimensionality reduction (DR) methods on our data before performing training on them. DR methods are unsupervised algorithms which try to find a projection of the data to lower dimension that preserve as much information as possible.

A lower number of dimensions helps reduce computation time and often even improves the predictive performance of models. Furthermore, DR methods can also be used to visualize high dimensional data by finding an informative projection into two dimensions that is understandable to humans. Some algorithms, such as t-SNE or UMAP, serve especially this purpose.

Principal **C**omponent **A**nalysis (PCA) is one of the most popular dimension reduction methods [4]. PCA finds linear projections to a lower-dimensional subspace so that variance in the data is maximized. Visualizing the ratio of variance, covered by individual principal components is a way of assessing the intrinsic dimensionality of the data, as shown on figure 1. We see that for the 122-dimensional atmospheric parameter space, we need:

- 23 dimensions to explain 95% of the variance,
- 45 dimensions to explain 99% of the variance,
- 73 dimensions to explain 99.9% of the variance,

and for the output 834-dimensional spectral space:

- 1 dimensions to explain 95% of the variance,
- 2 dimensions to explain 99% of the variance,
- 9 dimensions to explain 99.9% of the variance.

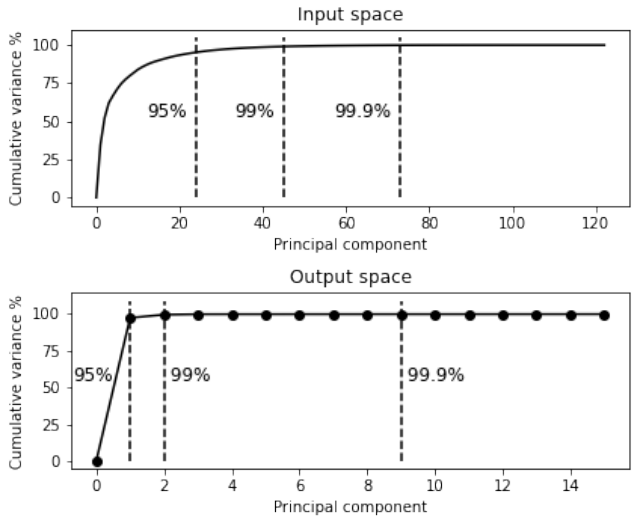


Figure 1: Dependence of the cumulative relative variance on the number of principal components for both the input and the output space.

Auto**e**ncoders (AE) [3] are a type of artificial neural network used to learn low dimensional representations. AE are trained to reproduce input data on the output of the network after passing through a bottleneck in the network architecture. To prevent autoencoders from memorizing the training dataset, a variety of regularization techniques can be employed. One of options is adding artificial noise to the input data, which forces the network to generalize.

In our framework, we employ this kind of autoencoder, often referred to as a denoising autoencoder, by adding Gaussian noise with mean 0 and standard deviation 0.1 to input data during the training process. A more thorough investigation of the effect of this technique on the predictive power can be found in [1]. For both atmospheric parameters and the spectral space we used the same 7 layers architecture with an appropriate size of input and output layers. The architecture can be summarized as:

- input layer of size $N_0 + Gaussiannoise$
- dense layer of size $N_1 < N_0$ and ReLu activation
- dense layer of size $N_2 = \frac{1}{2}N_1$ and ReLu activation
- dense embedding layer of size N_3 and linear activation
- dense layer of size N_2 and ReLu activation
- dense layer of size N_1 and ReLu activation
- output layer of size N_0 and linear activation

The **t**-Distributed **S**tochastic **N**eighbor **E**mbedding (t-SNE) [11] is non-linear unsupervised technique for high dimension data visualization which can model complex, non-linear dependencies. t-SNE places points that are similar in the original space close together in the embedding layer with a high probability, while placing dissimilar points close together with only a low probability. Since t-SNE is stochastic and non-parametric method there is no way to perform a reverse transformation from the embedding space to the original space. This excludes the method from use as part of the surrogate modelling process. It can, however, be useful for visualizing the dataset. Another disadvantage of t-SNE is its high computational complexity.

Uniform **M**anifold **A**pproximation and **P**rojection (UMAP) [12] is another dimension reduction technique used for dataset visualizations, constructed from a theoretical framework based in Riemannian geometry and algebraic topology. UMAP preforms similarly to t-SNE, but preserves more of the global data structure with superior run time performance. As is the case with t-SNE, UMAP does not allow for reverse transformations, which means we can not use it to learn surrogates. However, visualizations using UMAP allowed us to gain useful insights into the structure of our dataset.

3.2 Prediction models

One of the predictors we used in our experiment was a feed-forward **n**eural **n**etwork (NN). We have chosen an architecture, consisting of 2 hidden full connected layers with ReLu activation functions and linear activation on the output layer [6].

Random **F**orest (RF) is an ensemble learning technique suited for both regression and classification problems. It uses sample bagging and feature bagging methods to train a set of decision trees. Prediction is performed by averaging over predictions from the individual regression trees. The main advantage of RF over a simple decision tree is much better generalization. We decided to use this kind of predictor because it is capable of performing multi target regression [10].

Extra Random **T**rees (ET) is technique very similar to random forest, with two main differences. First, it uses the whole dataset for training individual trees instead of using bagging. Second, it uses random cuts for each split, instead of using the most optimal one (in case of Gini or Entropy reduction). It has been shown to perform better than random forests for some problems [5].

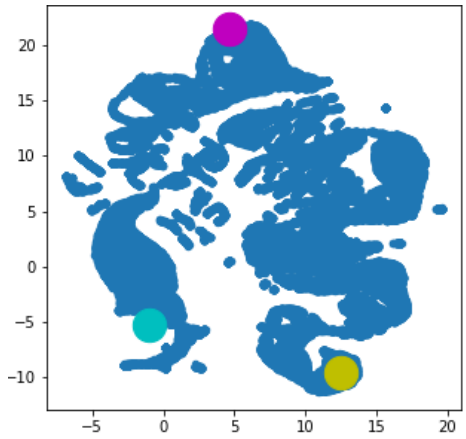
4 EXPERIMENT

Our experiment is composed of three parts. In the first two, we employ methods of dimensionality reduction as a way to gain

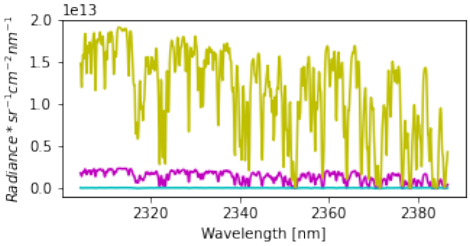
insight and understanding about our dataset and problem. The third part is an empirical evaluation of combinations of methods for dimensionality reduction and prediction, aiming to identify the one that offers the best predictive performance on unseen data.

4.1 Visualization

We applied the UMAP and t-SNE visualization techniques to both atmospheric parameters and spectrum data. As expected, both two methods showed clusters in the atmospheric parameters data. In cases of the spectrum data space, UMAP indicated a structure in the data, depicted in figure: 2. A comparison of the data points sampled from different clusters shows a large difference in the scale of individual data points. This is likely one of the reasons why such a high variance is concentrated in the first principal component (as seen on image 1).



(a) UMAP visualization



(b) Comparison of data points

Figure 2: UMAP visualization of the spectrum data.

4.2 Feature importance

The main advantage of using tree-based models over neural networks is their interpretability. While the ability to be understood by a human is lost when moving to an ensemble from a single tree, random forests can be very useful for estimating the importance of individual features for prediction. We trained a random forest predictor on the full dataset and visualized feature importance values in figure 3. We see that 70% of feature importance is accumulated just in two dimensions. This corresponds well to the PCA estimate of most variance being encompassed by two principal components. Only about half of the features are assigned a non-negligible importance. The features identified by this approach warrant further investigation by domain experts.

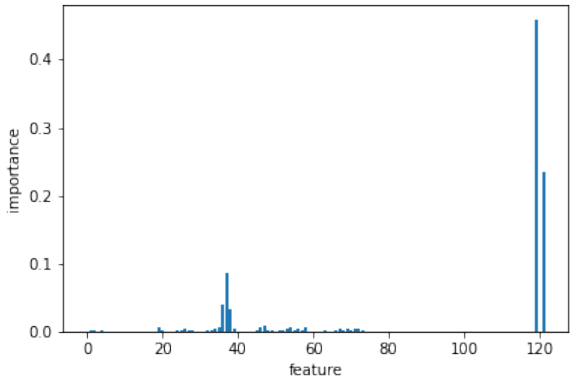


Figure 3: Random forest predictor importance of atmospheric data features.

4.3 Regression

To compare different regressors and methods of dimensions reduction we performed forward and backward predictions using neural network, random forest and extra random trees for both autoencoders and PCA embeddings. We reduced the dimensionality of the input space from 123 to 73 and the dimensionality of the output space from 834 to 9. These values correspond to 99.9% explained variance when using PCA. The noise level of the autoencoder was set to $\sigma = 0.1$. A more thorough study of these parameters can be found in [1]. We compare the predictive power of various combinations of either AE or PCA for dimension reduction, and either neural network, random forest or extra trees as a predictive model, using 10-fold cross validation. In table 1 we compare the results, using coefficient of determination as the evaluation metric [4]:

$$R^2 = 1 - \frac{MSE(\text{model})}{variance(\text{training set})}.$$

Table 1: Coefficient of determination for various combinations of dimensionality reduction method and predictive models, estimated by 10-fold cross validation.

	forward		backward	
model	**AE**	**PCA**	**AE**	**PCA**
NN	0.9995	**0.9998**	0.8454	0.9206
RF	0.8931	0.9937	0.9267	0.9311
ET	0.9228	0.9958	0.9370	**0.9510**

For the forward model, the best performance of $R^2 = 0.9998$ is achieved by a neural network, mapping between spaces reduced by PCA. For the backward model, the best performing model are extra trees, paired with PCA, achieving an $R^2 = 0.9510$. Both represent very satisfactory and promising models to employ as surrogates for radiative transfer modeling. From table 1 we can also see that PCA outperformed autoencoders in all cases, while also being much faster to compute. The comparison of predictive models is not as simple. For the forward model, the neural network is the best, but only by a small margin. For the backward model, the differences are bigger, with the neural network performing the worst. The performance of random forest was

between the performances of the other two predictive models for both the forward and the backward problem.

Since one of the main uses for surrogate models is speeding up computation, time complexity is an important consideration. The main disadvantage of neural networks is the computational complexity required for both training and prediction. An autoencoder takes about ten times as long to transform a data point to the embedding space than PCA. For predictive models, the neural network used in this study needed approximately three times as long to make a prediction than random forests and extra trees, which had a similar time complexity. Nonetheless, making predictions for a test set of 5000 points using any of the described surrogates takes up to one second, while running the full RemoTeC simulation requires several hours of computation.

When comparing with the evaluation results reported for the original framework in [1], the performances in this paper are slightly worse. The reason is the fact that the original study reduced the dimensions of the input space to 102 and the output space to 50 dimensions. In this study we focused on further reducing the dimensions and reduced the dimension of the input space to 73 dimensions and the output space to 9 dimensions. It is an interesting observation that for different dimensionalities, the best performance is achieved by different algorithms.

5 DISCUSSION AND FURTHER WORK

The original framework for learning surrogates on input and output spaces with reduced dimensionality showed high predictive and computational performance on the RemoTeC dataset. The results were very promising for applications in data analysis for Earth Observation missions as a way to dramatically speed up computation without sacrificing much accuracy. However, no single model and approach is the best for every dataset and application, which made the limited scope of options in the original framework a potential downside. With the work presented in this paper, the range of methods available has been extended. Since the choice of algorithms for dimensionality reduction on the input and output spaces, as well as the choice of prediction model for both the forward and the backward model are all independent from each other, the number of combinations of algorithms available is considerable. Furthermore, the dimension analysis enabled by UMAP, t-SNE and feature importance represents a new practical way of assessing intrinsic dimensionality and making a more informed choice of target dimension.

The paper presents an evaluation of the performance of various included methods on the RemoTeC dataset. However, each of the analyzed algorithms is defined by a number of hyperparameters, which is especially true for neural networks and autoencoders. Furthermore, the dimensions of the reduced input and output spaces can also be consider hyperparameters of the framework. For the presented evaluation we chose the hyperparameters based on values, reported in previous work and to some degree optimized them manually. A more rigorous study is required that employs automated hyperparameter optimization in order to compare the available algorithms fairly and arrive at a reliable conclusion of what is the best approach to modeling the RemoTeC simulation.

Finally, in this study we touched on the subject of estimating feature importance using random forests in order to gain insight about the data. However, feature importance can also be used to compute feature rankings and perform feature selection, which can be considered as another method of dimensionality reduction. In further work it might be worthwile to investigate this approach further and include it as an option in the framework for learning surrogates.

6 ACKNOWLEDGEMENTS

We thank Jennifer Adams and Edward Malina from the European Space Agency for providing the dataset and helping us understand and contextualize the problem. We thank Jovan Tanevski for his initial work on the project, as well as his ideas and help in further work.

REFERENCES

[1] Jure Brence, Jovan Tanevski, Jennifer Adams, Edward Malina, and Sašo Džeroski. 2020. Learning surrogates of a radiative transfer model for the sentinel 5p satellite. In *International Conference on Discovery Science*, in submission.

[2] A Butz, André Galli, O Hasekamp, J Landgraf, P Tol, and I Aben. 2012. Tropomi aboard sentinel-5 precursor: prospective performance of ch4 retrievals for aerosol and cirrus loaded atmospheres. *Remote Sensing of Environment*, 120, 267–276.

[3] David Charte, Francisco Charte, Salvador García, María J del Jesus, and Francisco Herrera. 2018. A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. *Information Fusion*, 44, 78–96.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Number 10. Volume 1. Springer series in statistics New York.

[5] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63, 1, 3–42.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

[7] Otto P Hasekamp and J Landgraf. 2002. A linearized vector radiative transfer model for atmospheric trace gas retrieval. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 75, 2, 221–238. ISSN: 00224073. DOI: 10.1016/S0022-4073(01)00247-3.

[8] Haili Hu, Otto Hasekamp, André Butz, André Galli, Jochen Landgraf, Joost Aan de Brugh, Tobias Borsdorff, Remco Scheepmaker, and Ilse Aben. 2016. The operational methane retrieval algorithm for tropomi. *Atmospheric Measurement Techniques (AMT)*, 9, 11, 5423–5440.

[9] IPCC. 2014. Fifth Assessment Report - Impacts, Adaptation and Vulnerability. (2014). Retrieved 06/12/2017 from http://www.ipcc.ch/report/ar5/wg2/.

[10] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2, 3, 18–22.

[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9, Nov, 2579–2605.

[12] Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction, (December 2018). https://arxiv.org/abs/1802.03426.

[13] J Tanevski, S Džeroski, and T Todorovski. 2019. Meta-model framework for surrogate-based parameter estimation in dynamical systems. *IEEE Access*, 99.