Topologically-based Variational Autoencoder for Time Series Classification

Rodrigo Rivera-Castro [1] **Samir Moustafa** [1] **Polina Pilyugina** [1] **Evgeny Burnaev** [1]

Abstract

Topological Data Analysis is an approach to analyze data using different techniques from topology. These techniques aim to extract fundamental qualitative properties, such as shape and connectivity in data. In this work, we propose a universal approach for time series classification with variational autoencoders. It is built on extracted features from the persistent homology theory. Compared to standard classification approaches, the proposed methodology enables the classification of time series, which have different recurrent behavior in the reconstructed phase space. Multiple experiments with time-series datasets confirm that the method makes classification more robust to noisy and high-dimensional data and favors datasets in which shape has meaning.

1. Introduction

Topological Data Analysis (TDA) is becoming popular among the data sciences. The crucial impact of TDA is the opportunity to extract important properties of qualitative nature from data. It can help to obtain structural information, such as shape. In this work, we show that it facilitates the classification of time series. However, this is not the only case that benefits from a topological perspective. For instance, TDA has seen success in a wide variety of fields ranging from detecting significant local structural sites in proteins (Sacan et al., 2007) to finance (Rivera-Castro et al., 2019a), and marketing (Rivera-Castro et al., 2019b). TDA-based approaches are more general and robust than constructing a hyperplane in some metric space, which causes dependencies on the metric chosen, which, for instance, can be corrupted by noise in the data (Tan et al., 2016). In this work, we concentrated on extracting features from various topological summaries obtained with the help of persistent homology. The theoretical motivation for using persistent homology is provided in the preliminaries section. The main contributions are as follows:

[1]Skoltech. Correspondence to: Rodrigo Rivera-Castro <rodrigo.riveracastro@skoltech.ru>.

LatinX in AI Research at ICML 2020, Vienna, Austria, 2020.

(1) A general pipeline for extracting various topological features for time series classification datasets from UCR Time Series Archive (Dau et al., 2018).

(2) An extensive imputation study of which type of features are better for which particular dataset from the UCR Time Series Archive.

(3) A performance ranking of six supervised classification algorithms with and without the deep generative model of Variational AutoEncoder (Kingma & Welling, 2013) as well as a shallow artificial neural network against the benchmark algorithm, the 1-NN Euclidian distance.

2. Experiments and Results

To test the proposed methodology, we run all of the 128 univariate time-series classification datasets from the UCR repository through all stages of the pipeline depicted in Figure 1. The UCR collection contains a variety of both multiclass and binary time-series classification problems with sizes ranging from forty time-series to over 50000. For training, we used an Nvidia Titan RTX with 80 cores.

To evaluate the impact of our approach, we considered four different classification algorithms with and without the Variational Autoencoder and one shallow neural network with VAE as an additional component after the encoding part. Among those four are CatBoost, XGBoost, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) classifier. We used accuracy as the metric over which we optimized the classifiers' hyperparameters and recorded accuracy for each possible pair: classifier - dataset.

2.1. Imputation study

How to estimate for a given dataset if topological features will show excellent performance on it? From the imputation study, we can propose the following conclusion. The potential performance of topological features highly depends on the initial data structure and point clouds that it can produce if point clouds have distinct shapes and different combinations of holes and loops. We can expect topological features for different classes to provide good quality as they will capture those differences. A metric-based method will probably outperform the topological-based approach if those forms are not distinguishable, and most of the difference lies in

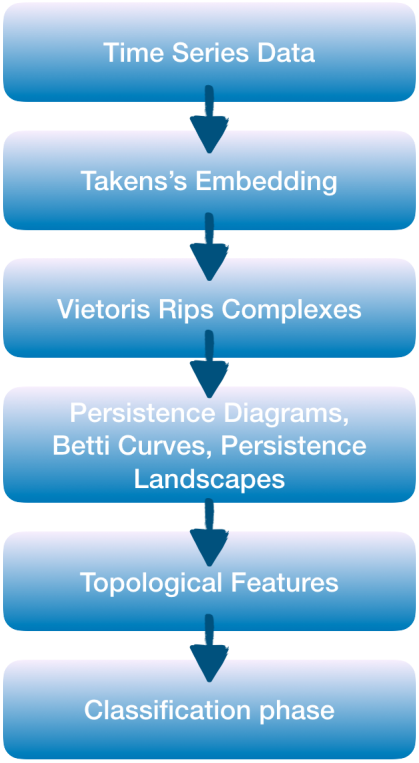

Figure 1. TDA pipeline for time-series classification

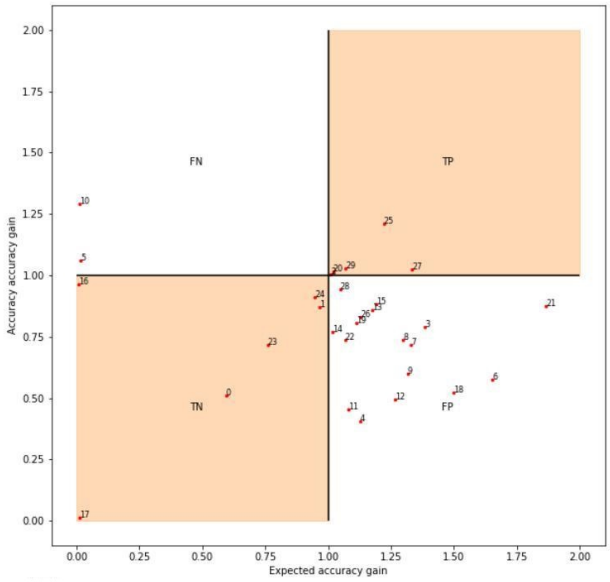

Figure 2. Expected accuracy gain calculated on training data versus actual accuracy gain on testing data

scales and distances.

2.2. Results

We present the results in Figure 2. For this, we built a Texas Sharpshooter using the following formula applied to the obtained train and test accuracy,

$$\text{Gain} = \frac{\text{Obtained Accuracy}}{\text{Accuracy of 1-NN Euclid}}$$

.

Some cases lie in the false-positive area. Data points of that region represent cases where we thought we could improve accuracy, but did not. In some cases, we improved accuracy by distinguishing peculiarities that were identified by a model. A successful model and hyperparameters selection also helped to outperform the baselines.

Similarly, we use a critical difference (CD) diagram to rank all algorithms. CD diagrams are an established methodology in the time-series classification literature to evaluate and compare multiple classifiers. The technique is based on the Wilcoxon-Holm method to detect pairwise significance, and we depict it in Figure 3. Thick horizontal lines show a group of classifiers that are not significantly different in terms of accuracy. The proposed auto-encoder performs as good as other classifiers.

Finally, we use Dolan-More curves (Dolan & Moré, 2001), in Figure 4, to compare the performance of the method against traditional techniques in the time-series classification literature such as Euclidean 1-Nearest Neighbor and Dynamic Time Warping 1-Nearest Neighbor. The higher the curve is on the plot, the better the quality is obtained by the corresponding algorithm. Overall, the Dolan-More curves show for each method a proportion of datasets, on which the method is worse than the best one not more than β times. We can see that using the proposed method with Variational Autoencoders leads to superior results than no using it as well as using standard methods from the literature.

3. Conclusion

This work aims to study and implement topological features for time series classification and apply a proposed pipeline to the UCR collection of datasets. The proposed features utilize topological summaries such as Persistence Diagram, Betti Curves, Persistence Landscape, and Persistence Entropy. The features are suited for identifying properties such as shape in the data. We notice that for point clouds with structural differences, topological features lead to excellent performance.

To validate the approach, we compared this method against multiple classifiers commonly used in machine learning as well as standard methods for time-series classification in the literature. The results show that the proposed technique is superior to the established practices in the time-series re-

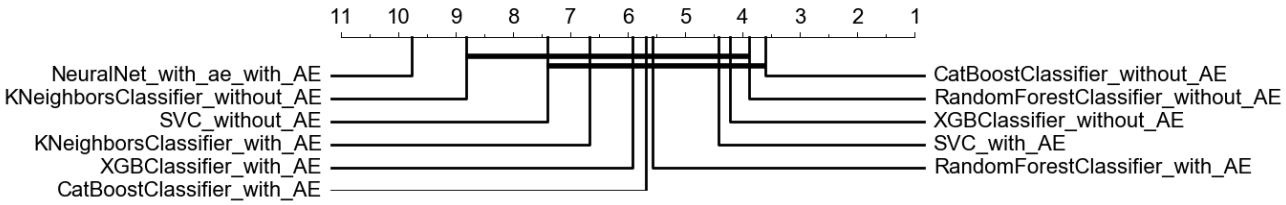

Figure 3. Critical Distance Diagram

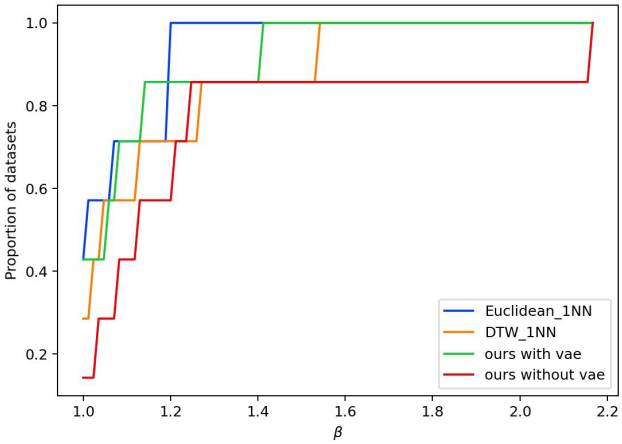

Figure 4. The Dolan-More curves show for each method a proportion of datasets, on which the method is worse than the best one not more than β times

search and is as equally good as standard machine-learning classifiers. However, the performance varies greatly depending on the dataset evaluated.

We can, therefore, conclude that the proposed approach's success depends on the structure of the time series and its corresponding point cloud. Thus, a future work line is to generate more representative point clouds tailored for time series tasks.

References

Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Dolan, E. D. and Moré, J. J. Benchmarking optimization software with performance profiles. February 2001.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. 2013. URL http://arxiv.org/abs/1312.6114.

Rivera-Castro, R., Pilyugina, P., and Burnaev, E. Topological data analysis for portfolio management of cryptocurrencies. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 238–243, November 2019a.

Rivera-Castro, R., Pletnev, A., Pilyugina, P., Diaz, G., Nazarov, I., Zhu, W., and Burnaev, E. Topology-Based clusterwise regression for user segmentation and demand forecasting. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 326–336, October 2019b.

Sacan, A., Ozturk, O., Ferhatosmanoglu, H., and Wang, Y. LFM-Pro: a tool for detecting significant local structural sites in proteins‡. *Bioinformatics*, 23(6): 709–716, 01 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl685. URL https://doi.org/10.1093/bioinformatics/btl685.

Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to data mining*. Pearson Education India, 2016.