Data compression for improved explanation estimation

Paulina Kaczyńska PM.KACZYNSKA@STUDENT.UW.EDU.PL
Mateusz Biesiadowski MB406097@STUDENTS.MIMUW.EDU.PL
Anna Semik AS406153@STUDENTS.MIMUW.EDU.PL
University of Warsaw, Poland

Abstract

In this project we aimed to check whether data compression with different kernels will prove useful when comparing the time of calculations and the accuracy of global explanations against baseline uniform sampling. We compared the performance of explanation methods on three datasets with two different model architectures. We determined that for SHAP the time gain is significant even on smaller datasets, but the accuracy of explanation is not better than alternative uniform sampling. For PVI, PDP and ALE the time gain was significant when the dataset was big. Moreover, for PVI, PDP and ALE, the accuracy of explanation was better both in terms of the median performance of explanations and smaller variance of their accuracy.

1. Introduction

Global explanations give us important insight into how machine learning models make predictions. However, their usefulness is sometimes limited by the high computation cost of such explanations. This problem becomes evident on bigger datasets, since time of explanations grows with the size of dataset.

For models trained on tabular data, one way to deal with this problem is to use a data subset for estimation. It is usually uniformly sampled from the validation dataset. However, it makes the explanation prone to being disproportionally influenced by outliers, if they happen to be included in the sampled subset.

An alternative solution could be to use data compression algorithms, which can improve the subset's distribution at a low computational overhead and make it more representative of the whole data. One such algorithm is Kernel Thinning introduced by Dwivedi and Mackey (2022).

In this project we try to estimate the usefulness of the latter approach. We compare explanations made on uniform subsamples of datasets and subsets sampled by the Kernel Thinning algorithm in order to see in what conditions using the KT algorithm is computationally beneficial and least detrimental to the obtained explanations.

2. Methodology

2.1 Compression methods

General kernel thinning algorithm was introduced by Dwivedi and Mackey (2022). It is formulated as follows: given initial input points $S_{in} = (x_i)_{i=0}^{n}$ from the probability distribution $\mathbb{P}$, it returns S_{out} of $\sqrt{n}$ output points with comparable integration error across a reproducing kernel Hilbert space. Compress++ (Shetty et al. (2022)) reduces the runtime of generic thinning algorithms (e.g. kernel thinning) with a minimal loss in accuracy.

We compress the validation dataset with the Compress++ algorithm (Shetty et al. (2022)).

For comparing accuracy, we sample uniformly without replacement the same number of samples as we did with the Kernel Thinning.

Each compression is repeated for the same number of times in order to estimate the variation in performance. We check how different kernel functions perform, and on this basis, choose kernel function.

2.2 Metrics

In order to judge whether explanation accuracy cost is a fair trade-off for the time gain, we measure two metrics.

1. Firstly, we measure the time difference between calculations of explanations on the whole validation data versus on the validation data compressed with Kernel Thinning. It is compared to the time that the compression process took.

2. Secondly, we measure the Wasserstein distances between empirical distributions of explanation values for the given feature on the whole and compressed dataset. The sum of Wasserstein distances over all features is the estimate of how much the explanation on compressed dataset diverged from the explanation on the uncompressed dataset.

Compression and explanation on the compressed dataset are repeated multiple times.

2.3 Datasets and models

We conducted experiments on three datasets of variable sizes:

- a synthetic classification dataset created with sklearn (5000 points, with validation dataset of 1024 points).

- Bank Marketing Dataset (Moro et al. (2014)),

- Covertype dataset (Blackard and Dean (1999)) for classification of forest covertype.

The datasets are cleaned from all categorical variables.

For each experiment, we trained two models: XGBoost (Chen and Guestrin (2016)), and k-nearest neighbours (Fix and Hodges (1951)).

2.4 Methods of explanation

We test four different methods of global explanation:

1. SHapley Additive exPlanations (SHAP) (Lundberg and Lee (2017)). It is important to note that for the XGBoost model we use TreeShap and for the K-nearest neighobours we use basic SHAP.

2. Permutation Variable Importance (PVI) (Breiman (2001)),

3. Partial Dependence Plot (PDP) (Hastie et al. (2009)) caluclated for the most important variable according to PVI, and

4. Accumulated Local Effects (ALE) caluclated for the most important variable according to PVI (Apley and Zhu (2020)).

3. Experimental results

Initially, we analyze the performance of the kernel thinning algorithm on different kernel functions on the Bank Marketing Dataset. The Wasserstein distances between respective empirical distributions of points can be seen on 1 On this base we decide to perform the rest of the experiments with gaussian kernel function with parameter equal to 1.

3.1 The synthetic dataset

The time difference results can be seen on the 2 figure. The red line is the mean time kernel thinning compression took. Time gain of the KNN model is bigger than the time cost of the explanation for each explanation model, but only for SHAP the difference is considerably bigger. For XGBoost, only the time gain for SHAP slightly exceeds the time cost of compression. The other explanation methods' time differences are below the time of compression and PVI explanation takes similar time on the dataset before compression and on the compressed dataset.

Moreover, Wasserstein distance of explanation made on KT-compressed dataset is similar to the one made on uniformly sampled dataset both in terms of the median value and the variance of predictions, as can be seen in figure 4.

3.2 The Bank Marketing Dataset

Despite the validation dataset being bigger than the one of synthetic dataset by an order of magnitude, the effects of compression are fairly similar to the effects on synthetic dataset. As can be seen on 4a, time gains for SHAP for both models significantly exceed computation cost. The variable profile methods' time differences are still below or equal to the computation time for the XGBoost model. PVI time differences are slightly above the line for both models, but the effect is hardly significant.

Wasserstein distance for the SHAP explanation is comparable and equally small for both uniform sampling and kernel thinning. The accuracies of other explanations are bigger than of SHAP for both models, but there is no significant difference between the two methods of compression.

Overall, it can be concluded that the time gain and the accuracy trade-off on SHAP are very promising. However, kernel thinning has not demonstrated the advantage over uniform sampling. Compression for other methods on these datasets has not yet been proven profitable.

3.3 The Covertype Dataset

The experiments were conducted for the three methods that did not yield any advantage on the previous datasets: PVI, PDP and ALE. The results can be seen on the figure 5. For this dataset, time gain exceeds compression cost on both models for all three explanation methods.

Accuracy measured by Wasserstein distance is now better for the Kernel Thinning than uniform sampling. The median of experiment repetitions is for every method smaller in case of kernel thinning than uniform sampling. Moreover, the variance of uniform explanations is significantly bigger. This confirms the outlier effect, where in case of uniform sampling the choice of the outlier can significantly impact accuracy of the explanation.

This demonstrates, that for certain datasets (especially bigger ones) use of kernel thinning algorithm can improve the time of explanation and give better explanation accuracy than uniform sampling of datapoints. Since explanations are usually not repeated multiple times, decreasing outlier effects that is done by kernel thinning is especially important.

4. Conclusion

In order to determine if kernel thinning compression can improve time cost of explanation without too big of an accuracy loss, the experiments were performed on three datasets with two model architectures and different global explanation methods. SHAP explanation time gain exceeded time cost of compression even on smaller datasets, but performed similarily well as uniform sampling in terms of accuracy of explanation. Other explanations did not show significant time gain with the use of compression, nor better accuracy than uniform sampling on two smaller datasets. However, on the biggest of three datasets, they exhibited both time gain compared to the whole dataset explanation and accuracy advantage over uniform sampling. Moreover, the Wasserstein distance of KT explanation from the explanation on the full validation dataset had lower variance, which means that one time explanation is far more representable of the model when performed with the kernel thinning compression than with uniform sampling.

Summing up, especially explanations on big datasets can gain from the use of kernel thinning. Compression is profitable more for the SHAP method than PVI, ALE and PDP.

Appendix A. Plots

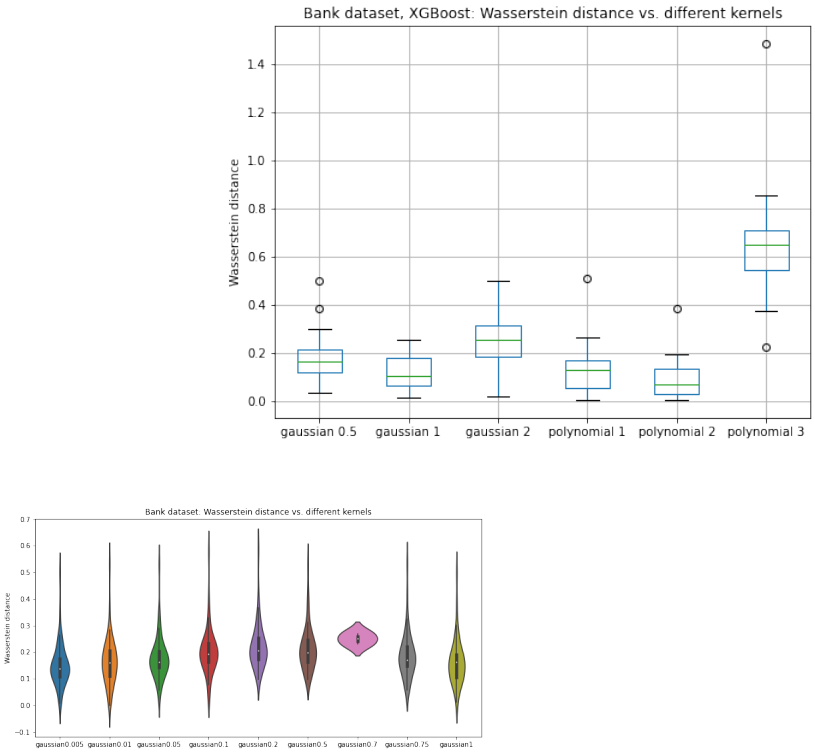


Figure 1: Wasserstein distance on the Bank dataset for different kernels

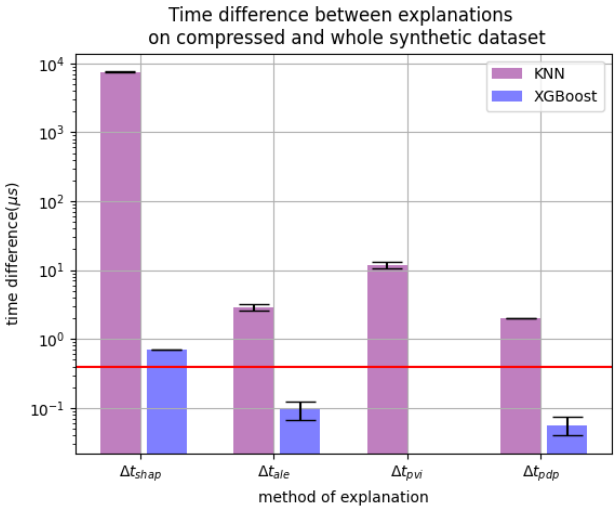


Figure 2: Time gained by computing different explanation methods with both models on the compressed dataset vs. the uncompressed dataset. The red line signifies the mean time of the kernel thinning compression calculation

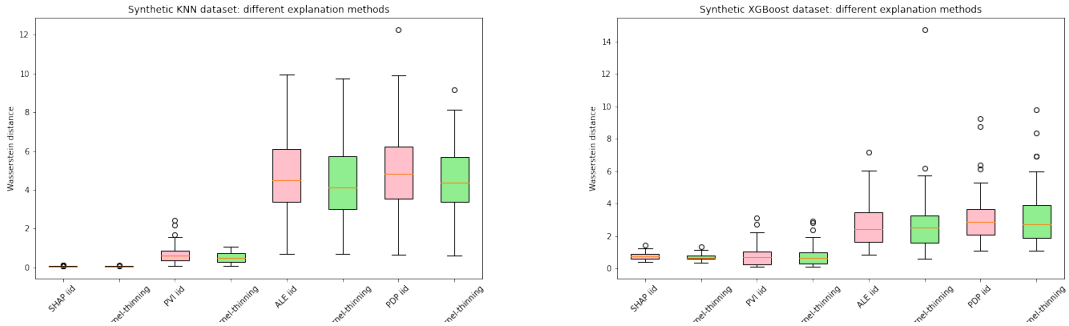
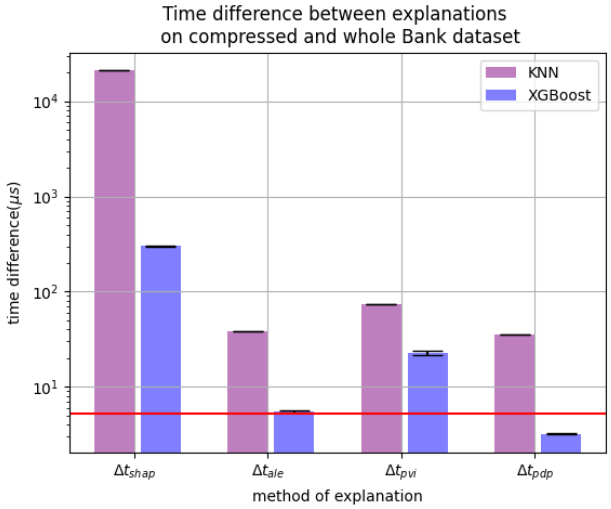
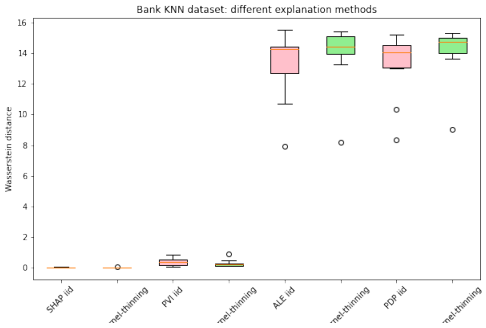


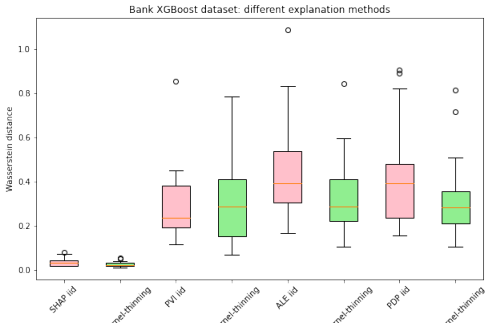
Figure 3: The Wasserstein distance between explanations on the full validation dataset and dataset compressed with kernel thinning (green) or uniformly sampled (red)



(a) Time difference between explanations on the whole test set and the one compressed with Kernel Thinning for different explanation methods

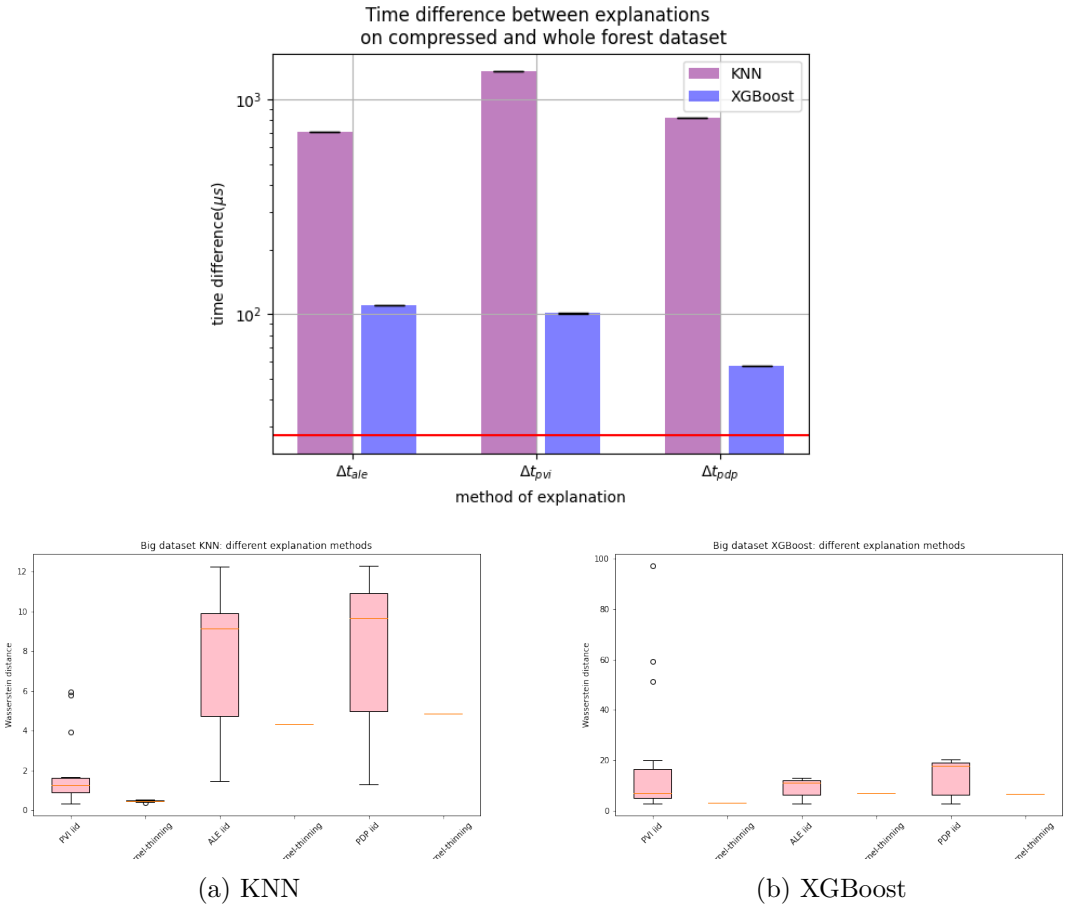


(b) KNN



(c) XGBoost

Figure 4: Time and Wasserstein distance of different explanation methods on the Bank dataset



(a) KNN (b) XGBoost

Figure 5: Time and Wasserstein distance of different explanation methods on the big dataset

References

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(4):1059–1086, September 2020. ISSN 1369-7412. doi: 10.1111/rssb.12377.

Jock Blackard and Denis Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24:131–151, 12 1999. doi: 10.1016/S0168-1699(99)00046-0.

Leo Breiman. Random forests. 45:"5–32", 2001. URL `https://doi.org/10.1023/A:1010933404324`.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. *ICLR*, 2022.

Evelyn Fix and Joseph Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. pages 41(128)–31, 1951.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 06 2014. doi: 10.1016/j.dss.2014.03.001.

Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022.