Data compression for improved explanation estimation

Anna Semik, Mateusz Biesiadowski, Paulina Kaczyńska

The goal of the project is to examine if Kernel Thinning algorithms can be used to speed up calculating global explanations of machine learning models.

Variables, which effects on the compression are going to be checked, are:

- explanation method: SHAP, PVI, PDP/ALE,
- dataset (size): Friedman dataset[1] (sklearn), Bank Marketing Dataset[2] and COVID-19 dataset[3],
- model: XGBoost and K-nearest neighbours,
- compression parameters (kernel function).

In order to appropriately compare experiments' results, we established a few metrics: the accuracy of the explanation is measured as the L1 distance of the explanation on compressed data from the explanation on uncompressed data; the accuracy of the compressed distribution is measured with the Wasserstein distance; the time performance is rated as the difference between time of compression + time of explanation on compressed data and time of explanation on uncompressed data.

What we have done:

- the code refactorization;
- initial analysis of the data collected by H. Baniecki;
- experiments with XGBoost on Friedman dataset;
- replication of part of the experiments on the Bank Marketing Dataset.

Basing on the initial data analysis, the future experiments are narrowed down to the Compress++ alghoritm with gaussian kernel.

Next steps:

- experiments on the COVID-19 dataset;
- analysis of the kernel function impact;
- conducting experiments with KNN model;
- complex data analysis.

Difficulties:

- long computation time on the last dataset;
- gaining access to the Entropia cluster.

---

[1] J. Friedman, "Multivariate adaptive regression splines", The Annals of Statistics 19 (1), pages 1-67, 1991.

[2] Moro, Sérgio; Cortez, Paulo; Rita, Paulo (2014). "A data-driven approach to predict the success of bank telemarketing". *Decision Support Systems*. 62: 22–31. doi:10.1016/j.dss.2014.03.001. hdl:10071/9499.

[3] https://www.kaggle.com/datasets/meirnizri/covid19-dataset