Sparseness-Optimized Feature Importance

Isel Grau[1,2][0000−0002−8035−2887] and Gonzalo Nápoles[3][0000−0003−1936−3701]⋆

[1] Information Systems, Eindhoven University of Technology, The Netherlands
[2] Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, The Netherlands
[3] Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands
i.d.c.grau.garcia@tue.nl; g.r.napoles@uvt.nl

Abstract. In this paper, we propose a model-agnostic post-hoc explanation procedure devoted to computing feature attribution. The proposed method, termed Sparseness-Optimized Feature Importance (SOFI), entails solving an optimization problem related to the sparseness of feature importance explanations. The intuition behind this property is that the model's performance is severely affected after marginalizing the most important features while remaining largely unaffected after marginalizing the least important ones. Existing post-hoc feature attribution methods do not optimize this property directly but rather implement proxies to obtain this behavior. Numerical simulations using both structured (tabular) and unstructured (image) classification datasets show the superiority of our proposal compared with state-of-the-art feature attribution explanation methods. The implementation of the method is available on https://github.com/igraugar/sofi.

Keywords: model-agnostic explainability · feature importance · sparse explanations.

This document is an encore abstract of the paper "Sparseness-Optimized Feature Importance" published at the 2nd World Conference on Explainable Artificial Intelligence (XAI 2024) [2].

1 Summary

Explainable Artificial Intelligence (XAI) is essential for increasing the transparency and interpretability of machine learning models, especially in areas where decisions have significant consequences. In this paper, we propose a novel model-agnostic post-hoc explanation method, termed Sparseness-Optimized Feature Importance (SOFI), specifically designed for computing sparse feature importance rakings. Unlike existing methods that rely on perturbations or gradient propagation for computing attributions, SOFI directly addresses the sparseness of the explanations by defining the problem of computing the feature importance

⋆ Equal contribution.

ranking as an optimization problem. This approach is grounded on the intuition that marginalizing the most important features significantly impacts model performance while marginalizing the least important ones has minimal effect. As a result, the feature importance rankings obtained by SOFI are also deemed more correct and faithful to the behavior of the black box model since they show the steepest performance degradation curves in the experiments. In the literature, the sparseness and correctness properties derived from incrementally deleting features continue to be the mainstream approach to compare feature attribution post-hoc methods, regardless of their inner working principles [4]. Fig. 1 summarizes the overview of the proposed SOFI method.

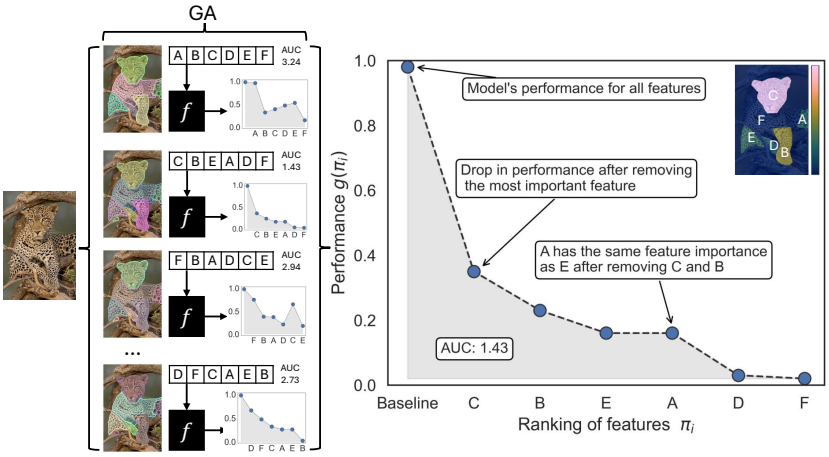

Fig. 1: Overview of SOFI method: Based on the input data, the ranking of features is optimized using a GA and evaluating a fitness function based on the area under the performance curve and a penalization factor for avoiding picks in the curve. The best permutation is returned as the resulting ranking of feature importance and optionally mapped to the instance for visualization purposes.

When tested on structured classification tasks, the SOFI method continuously reported the smallest area under the performance curve compared to PFI [1] and SHAP [3] methods while being the second fastest algorithm after PFI. Similarly, SOFI reported sparser explanations than SHAP, Grad-CAM [6], and RISE [5] methods when tested on image datasets instances. In the image classification task, segmentation annotations were used as super-pixels. This approach is not perceived as a relevant limitation since there are efficient algorithms for image segmentation that allow obtaining meaningful super-pixels, thus conveniently reducing the granularity of explanations. Future work will be focused on extending our approach to multi-objective optimization settings where more than one desiderata for explanations can be optimized, especially after the automated semantic segmentation of unstructured data such as images.

References

1. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
2. Grau, I., Nápoles, G.: Sparseness-optimized feature importance. In: World Conference on Explainable Artificial Intelligence. pp. 393–415. Springer (2024)
3. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
4. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys **55**(13s), 1–42 (2023)
5. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
6. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision **128**(2), 336–359 (2019). https://doi.org/10.1007/s11263-019-01228-7