Explaining Change in Models and Data with Global Feature Importance and Effects

Maximilian Muschalik[1,2,*,†], Fabian Fumagalli[3,*,†], Barbara Hammer[3] and Eyke Hüllermeier[1,2]

[1] *LMU Munich, D-80539 Munich, Germany*

[2] *MCML, Munich*

[3] *Bielefeld University, D-33619 Bielefeld, Germany*

Abstract

In dynamic machine learning environments, where data streams continuously evolve, traditional explanation methods struggle to remain faithful to the underlying model or data distribution. Therefore, this work presents a unified framework for efficiently computing incremental model-agnostic global explanations tailored for time-dependent models. By extending static model-agnostic methods such as Permutation Feature Importance, SAGE, and Partial Dependence Plots into the online learning context, the proposed framework enables the continuous updating of explanations as new data becomes available. These incremental variants ensure that global explanations remain relevant while minimizing computational overhead. The framework also addresses key challenges related to data distribution maintenance and perturbation generation in online learning, offering time and memory efficient solutions like geometric reservoir-based sampling for data replacement.

Keywords

Explainable Artificial Intelligence, Interpretable Machine Learning, Online Learning, Concept Drift

1. Introduction

In applied machine learning, data often evolves over time, which necessitates changes to prediction models. Ensuring the reliability of such time-dependent models is increasingly important in high-stake applications, such as financial services [1], sensor [2, 3] and network [4] analysis. In recent years, eXplainable Artificial Intelligence (XAI) has targeted such time-dependent explanations of predictions that react to changes in the underlying data distributions and prediction models [5]. In extreme cases, where data is observed sequentially over time from a data stream, models are updated incrementally with each now observation, known as online learning or incremental learning [6]. In this context, re-computing XAI methods from scratch can become computationally infeasible, where incremental variants have been proposed [7, 8, 9, 10, 11].

In this work, we present a unified framework that allows to efficiently compute incremental variants of model-agnostic global explanations (MAGEs). We demonstrate that existing incremental XAI techniques are summarized in the incremental MAGE framework. Furthermore, static MAGEs cover a wide range of existing model-agnostic XAI methods, including Shapley interactions [12], which expand the range of efficient incremantal XAI techniques for interpretability of black-box online learning models.

2. Background

We first introduce background on model-agnostic global explanations (Section 2.1), as well as online learning from data streams (Section 2.2). We consider a trained black-box model $f : \mathcal{X} \to \mathcal{Y}$ with input domain $\mathcal{X}$ equipped with a d-dimensional feature representation $\mathcal{D} = \{1, \dots, d\}$, e.g. $\mathcal{X} = \mathbb{R}^d$, and

TempXAI@ECML-PKDD'24: Explainable AI for Time Series and Data Streams Tutorial-Workshop, Sep. 9th, 2024, Vilnius, Lithuania

[*] Corresponding author.

[†] These authors contributed equally.

Maximilian.Muschalik@lmu.de (M. Muschalik); ffumagalli@techfak.uni-bielefeld.de (F. Fumagalli)
0000-0002-6921-0204 (M. Muschalik); 0000-0003-3955-3510 (F. Fumagalli); 0000-0002-0935-5591 (B. Hammer); 0000-0002-9944-4108 (E. Hüllermeier)

output domain $\mathcal{Y}$. We do do not make any further assumption on the model architecture and instead only allow access to the model by predicting instances. This is known as model-agnostic explanations [13].

2.1. Model-Agnostic Global Explanations

A *global explanation* of a black-box model considers the behavior of f across a whole labeled dataset $(x_j, y_j) \in \mathcal{X} \times \mathcal{Y}$ with $j = 1, \dots, n$. *Global feature importance (FI)* is an instance of global explanations that outputs an importance score $\phi^{\text{FI}} : \mathcal{D} \to \mathbb{R}$ for every feature $i \in \mathcal{D}$ [14]. Global FI measures a change in a model's performance, if the model's access to this feature's information is restricted. *Permutation FI (PFI)* [15, 16] is computed by permuting the values of the target feature and measuring the change in performance across a dataset. By permuting the feature's value, the model's access to this information is limited and, thus, PFI yields an efficient way to compute global FI. However, a feature's information provided to the model's performance might strongly depend on other features. Therefore, perturbing a single feature's value in the presence of all remaining features is a limitation of PFI. Shapley additive global importance (SAGE) [14, 17] accounts for this limitation by computing the increase in loss across sampled permutations $\pi : \mathcal{D} \to \mathcal{D}$ over the feature space. For such a permutation π, each feature $i \in \mathcal{D}$ appears at a certain position. SAGE proposes to measure the average increase in loss for the preceding features of i with and without i. By sampling over several permutations, an approximation of the Shapley Value (SV) [18] is obtained, a concept from cooperative game theory that guarantees that the SAGE values *fairly* decompose the overall loss. While global FI quantifies the impact of individual features, it is limited in its expressivity.

To understand Feature Effects (FE), Partial Dependence Plots (PDPs) [19] visualize the effect of imputing a specified feature's value cross all observations and compute the average prediction of all observations, when this feature's value is set. The PDP visualizes this average prediction across a range of different values, which allows to globally interpret the effect of changing this feature's value on average [20]. Besides PDP, there exist other FE methods [20, 21, 22] with extensions to regional explanations [20, 23]. Another way of quantifying FEs is by using interaction indices that distribute contributions to all individuals and groups of features up to a maximum group size k. In recent work, several Shapley-based interaction indices have been proposed [24, 25, 26] as well as their efficient computation in a model-agnostic setting [12, 24, 25, 27, 28, 29]. Model-agnostic global explanations were widely applied in static environments [30], however, in practice, data is often of dynamic nature, where explanations become outdated when models are adapted over time.

2.2. Online Learning From Data Streams

In many real-world applications [2] data is observed sequentially over time. In an extreme setting, we observe a data stream $(x_0, y_0), \dots, (x_t, y_t)$, where at time t the data point (x_t, y_t) is observed. The goal of *online learning* [6] is to train a time-dependent model f_t by using the current observation (x_t, y_t) once to obtain an updated model f_{t+1}, i.e.

$$\texttt{IncrementalUpdate}(f_t, x_t, y_t) \longrightarrow f_{t+1}.$$

Prominent instances of online learning algorithms include Hoeffding adaptive trees [31] and adaptive random forests [32], where splits and tree-structures are replaced, if they become outdated. Other training schemes, such stochastic gradient descent, inherently allow for incremental updates [6]. Online learning is especially important, if the underlying data distribution changes over time. This phenomenon is known as concept drift and occurs in many applications [33]. Detecting concept drift and reacting adequately by updating the model accordingly is one of the major applications of incremental learning [6]. A common approach to detect concept drift is via accuracy-based drift detectors, where a sudden change in accuracy of the model indicates a change of distributions [33]. Recently, it was proposed to enhance such detection schemes using global FI methods [5]. However, the computation of such methods is a challenging problem that has been mainly considered in static scenarios.

3. A Unified Framework for Explaining Change in Models and Data

We now present unified framework that allows to efficiently explore incremental variants of model-agnostic global explanations in an online learning setting. In a static setting, a global explanation is typically computed for individual features (global FI) or groups of features (global FE), which we summarizes in the following definition.

Definition 1 ($\mathcal{E}$). *Global explanations are computed for every element in the* ***explanation domain*** *$\mathcal{E}$, which is a collection of features and interactions $\mathcal{E} \subseteq 2^{\mathcal{D}}$.*

Given an explanation domain, the explanation can be computed for each element in a static setting.

Definition 2 (Static MAGE). *A static* ***Model-Agnostic Global Explanation (MAGE)*** *$\phi_f : \mathcal{E} \to \mathbb{R}$ for a set of features S is*

$$\phi_f(S) := \frac{1}{n} \sum_{j=1}^{n} \lambda_f \left(x_j, y_j, S, \mathcal{P}_{x_j,S}\right).$$

Here, $\mathcal{P}_{x_j,S}$ is a set of data points and λ_f is a method-specific explanation function.

Typically, the perturbation data $\mathcal{P}_{x,S}$ is constructed by using a combination of the data point x and another sampled data point $\tilde{x}$, where the feature's values from S and $-S := \mathcal{D} \setminus S$ are taken from either x or $\tilde{x}$ [14]. Thereby, the sampling of $\tilde{x}$ may be done dependently or independently of x. Instantiations of static MAGEs include PFI [15] where $\mathcal{E}$ contains individual features, and λ_f measure the increase in loss. Therein, $\mathcal{P}_{x_j,S}$ includes a single data point constructed by the values of x_j for features in $-S$ and the values of S from another data point obtained from the dataset using a permutation. SAGE [14] is also covered in this framework by choosing λ_f as the average over sampled permutations over $\mathcal{D}$, as described in Section 2.1. Lastly, PDPs [19] are contained in this framework, where λ_f is chosen as the prediction of a combination of x_j and $\tilde{x} \in \mathcal{P}_{x_j,S}$, where $\mathcal{P}_{x,S}$ contains the data points for which the PDP is visualized.

Having established a unified view on static MAGEs, we now turn our focus to an *online learning* setting as described in Section 2.2. Using the observed data points at time t a naive way to compute MAGEs is via Definition 2 as

$$\phi_t(S) := \frac{1}{t} \sum_{s=0}^{t-1} \lambda_{f_t} \left(x_s, y_s, S, \mathcal{P}_{x_s,S}\right). \tag{1}$$

Re-computing Eq. 1 at every time step t is an exhaustive operative, since static MAGEs are already time-consuming when computed once [14]. Moreover, Eq. 1 requires to store the full data stream, which is typically considered infeasible. As a remedy, practitioners might restrict the computation to a time window of fixed size [5]. However, reducing the number of observations lowers the quality of the explanation, which increases the variance. In the following, we propose a framework for an incremental computation of ϕ_t, similar to the incremental update of the model f_t. Our goal is to leverage the previously calculated MAGE and update this explanation using the currently available datapoint, i.e.

$$\texttt{IncrementalUpdate}(\phi_t, f_t, x_t, y_t) \longrightarrow \phi_{t+1}$$

By introducing a smoothing parameter $0 < \alpha < 1$, we define the incremental MAGE.

Definition 3 (Incremental MAGE). *Let $0 < \alpha < 0$. We define an incremental MAGE as*

$$\phi_t(S) := (1 - \alpha) \cdot \phi_{t-1}(S) + \alpha \cdot \lambda_{f_t} \left(x_t, y_t, S, \mathcal{P}_{x_t,S}\right).$$

The incremental MAGE computes a single term of the sum in Eq. 1 at each time step and exploits the previously computed MAGE values. This drastically reduces the computational complexity, which is at time t equal to computing MAGE once with Eq. 1. However, the incremental MAGE allows to obtain ϕ_t for every time step t without sacrificing computational resources. Incremental variants of PFI [7] and

SAGE [8], as well as PDP [9] have been recently proposed. They can be viewed as an instantiation of incremental MAGEs.

A major challenge in computing incremental MAGEs is the maintenance of the perturbation dataset $\mathscr{P}_{x,S}$ over time, i.e. efficiently constructing perturbed data points that adhere to the data distribution. Reservoir sampling [34] has been adapted to efficiently store the data distribution with minimum resources [7]. Geometric sampling [7] proposes to store a reservoir of fixed lengths, where data points are replaced over time and more recent observations have a higher probability to be present in the reservoir compared to older observations. This mechanism allows to maintain a time-dependent marginal data distribution with limited resources. More advanced techniques maintain conditional distributions using online decision trees and allow for conditional sampling as required for instance in conditional SAGE [8]. It has been shown that both sampling techniques yield substantially different explanations [35]. Geometric sampling with marginal distributions highlights the structure of the model, whereas observational approaches via conditional sampling include the data distribution in the explanation [35].

4. Conclusion and Future Work

We summarized popular model-agnostic global explanation techniques, such as FI-based PFI and SAGE, as well as FE-based PDPs, into the MAGE framework for static learning environments. We then proposed the incremental MAGE framework to directly compute these explanations for online learning on data streams. Incremental MAGE allows to incrementally update previous estimates of MAGEs at each time step using minimal resources. We have shown that incremental variants, such as iPFP, iSAGE and iPDP can be summarized in the incremental MAGE framework. Incremental MAGE offers opportunities to expand the range of incremental variants of MAGE-techniques. For instance, recently proposed methods to estimate Shapley interactions [12, 25, 29] may be placed in the incremental MAGE framework to discover for complex interactions beyond isolated FE. Moreover, with increasing variety of explanations using different complexity levels, *human-centered presentations* and *visualizations* are important future work.

Acknowledgments

We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

[1] J. M. Clements, D. Xu, N. Yousefi, D. Efimov, Sequential Deep Learning for Credit Risk Monitoring with Tabular Financial Data, CoRR abs/2012.15330 (2020). arXiv:2012.15330.

[2] M. Bahri, A. Bifet, J. Gama, H. M. Gomes, S. Maniu, Data stream analysis: Foundations, major tasks and tools, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1405. doi:10.1002/widm.1405.

[3] N. Davari, B. Veloso, R. P. Ribeiro, P. M. Pereira, J. Gama, Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry, in: 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2021), IEEE, 2021, pp. 1–10. doi:10.1109/DSAA53316.2021.9564181.

[4] B. G. Atli, A. Jung, Online Feature Ranking for Intrusion Detection Systems, CoRR abs/1803.00530 (2018). arXiv:1803.00530.

[5] M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, Agnostic explanation of model change based on feature importance, Künstliche Intell. 36 (2022) 211–224. URL: https://doi.org/10.1007/s13218-022-00766-6. doi:10.1007/S13218-022-00766-6.

[6] V. Losing, B. Hammer, H. Wersing, Incremental On-line Learning: A Review and Comparison of State of the Art Algorithms, Neurocomputing 275 (2018) 1261–1274. doi:10.1016/j.neucom.2017.06.084.

[7] F. Fumagalli, M. Muschalik, E. Hüllermeier, B. Hammer, Incremental permutation feature importance (ipfi): towards online explanations on data streams, Mach. Learn. 112 (2023) 4863–4903. URL: https://doi.org/10.1007/s10994-023-06385-y. doi:10.1007/S10994-023-06385-Y.

[8] M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, isage: An incremental version of SAGE for online explanation on data streams, in: D. Koutra, C. Plant, M. G. Rodriguez, E. Baralis, F. Bonchi (Eds.), Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part III, volume 14171 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 428–445. URL: https://doi.org/10.1007/978-3-031-43418-1_26. doi:10.1007/978-3-031-43418-1_26.

[9] M. Muschalik, F. Fumagalli, R. Jagtani, B. Hammer, E. Hüllermeier, ipdp: On partial dependence plots in dynamic modeling scenarios, in: L. Longo (Ed.), Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part I, volume 1901 of *Communications in Computer and Information Science*, Springer, 2023, pp. 177–194. URL: https://doi.org/10.1007/978-3-031-44064-9_11. doi:10.1007/978-3-031-44064-9_11.

[10] A. P. Cassidy, F. A. Deviney, Calculating feature importance in data streams with concept drift using online random forest, in: 2014 IEEE International Conference on Big Data (Big Data 2014), 2014, pp. 23–28. doi:10.1109/BigData.2014.7004352.

[11] H. M. Gomes, R. F. d. Mello, B. Pfahringer, A. Bifet, Feature scoring using tree-based ensembles for evolving data streams, in: 2019 IEEE International Conference on Big Data (Big Data 2019), 2019, p. 761–769.

[12] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hüllermeier, B. E. Hammer, SHAP-IQ: Unified approximation of any-order shapley interactions, in: Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[13] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[14] I. Covert, S. M. Lundberg, S.-I. Lee, Understanding Global Feature Contributions With Additive Importance Measures, in: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS 2020), 2020, p. 17212–17223.

[15] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.

[16] A. Fisher, C. Rudin, F. Dominici, 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 (2019) 1–81.

[17] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the Feature Importance for Black Box Models, volume 11051 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2019, p. 655–670. doi:10.1007/978-3-030-10925-7_40.

[18] L. S. Shapley, A Value for n-Person Games, in: Contributions to the Theory of Games (AM-28), Volume II, Princeton University Press, New Jersey, USA, 1953, pp. 307–318.

[19] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics 29 (2001) 1189–1232. URL: http://www.jstor.org/stable/2699986.

[20] J. Herbinger, B. Bischl, G. Casalicchio, REPID: regional effect plots with implicit interaction detection, in: G. Camps-Valls, F. J. R. Ruiz, I. Valera (Eds.), International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event, volume 151 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 10209–10233. URL: https://proceedings.mlr.press/v151/herbinger22a.html.

[21] D. W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82 (2016). URL: https://api.semanticscholar.org/CorpusID:88522102.

[22] S. M. Lundberg, G. G. Erion, H. Chen, A. J. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S. Lee, From local explanations to global understanding with explainable AI for trees,

Nature Machine Intelligence 2 (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[23] J. Herbinger, B. Bischl, G. Casalicchio, Decomposing global feature effects based on feature interactions, CoRR abs/2306.00541 (2023). URL: https://doi.org/10.48550/arXiv.2306.00541. doi:10.48550/ARXIV.2306.00541. arXiv:2306.00541.

[24] M. Sundararajan, K. Dhamdhere, A. Agarwal, The Shapley Taylor Interaction Index, in: Proceedings of the 37th International Conference on Machine Learning, (ICML 2020), volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 9259–9268.

[25] C. Tsai, C. Yeh, P. Ravikumar, Faith-Shap: The Faithful Shapley Interaction Index, Journal of Machine Learning Research 24 (2023) 1–42.

[26] S. Bordt, U. von Luxburg, From Shapley Values to Generalized Additive Models and back, in: International Conference on Artificial Intelligence and Statistics (AISTATS 2023), volume 206 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 709–745.

[27] P. Kolpaczki, M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, SVARM-IQ: Efficient approximation of any-order Shapley interactions through stratification, in: Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, (AISTATS 2024), volume 238 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 3520–3528.

[28] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hüllermeier, B. Hammer, KernelSHAP-IQ: Weighted least square optimization for shapley interactions, in: Forty-first International Conference on Machine Learning, 2024. URL: https://openreview.net/forum?id=d5jXW2H4gg.

[29] M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles, in: Thirty-Eighth AAAI Conference on Artificial Intelligence, (AAAI 2024), AAAI Press, 2024, pp. 14388–14396. doi:10.1609/AAAI.V38I13.29352.

[30] I. Covert, S. Lundberg, S.-I. Lee, Explaining by Removing: A Unified Framework for Model Explanation, Journal of Machine Learning Research 22 (2021) 1–90.

[31] G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001), ACM Press, 2001, p. 97–106. doi:10.1145/502512.502529.

[32] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, T. Abdessalem, Adaptive random forests for evolving data stream classification, Machine Learning 106 (2017) 1469–1495. doi:10.1007/s10994-017-5642-8.

[33] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under Concept Drift: A Review, IEEE Transactions on Knowledge and Data Engineering (2018) 2346–2363. doi:10.1109/TKDE.2018.2876857.

[34] J. S. Vitter, Random Sampling with a Reservoir, ACM Transactions on Mathematical Software 11 (1985) 37–57. doi:10.1016/j.ipl.2005.11.003.

[35] F. Fumagalli, M. Muschalik, E. Hüllermeier, B. Hammer, On feature removal for explainability in dynamic environments, in: 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023, 2023. URL: https://doi.org/10.14428/esann/2023.ES2023-148. doi:10.14428/ESANN/2023.ES2023-148.