ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

On Feature Removal for Explainability in Dynamic Environments

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

1- Bielefeld University, D-33615 Bielefeld, Germany

2- LMU Munich, D-80539 Munich, Germany

† equal contribution

Abstract. Removal-based explanations are a general framework to provide feature importance scores, where feature removal, i.e. restricting a model on a subset of features, is a central component. While many machine learning applications require dynamic modeling environments, where distributions and models change over time, removal-based explanations and feature removal have mainly been considered in a static batch learning environment. Recently, an interventional and observational perturbation method was presented that allows to remove features efficiently in dynamic learning environments with concept drift. In this paper, we compare these two algorithms on two synthetic data streams. We showcase how both yield substantially different explanations when features are correlated and provide guidance on the choice based on the application.

1 Introduction

Feature importance (FI) is a prominent technique to understand black-box machine learning (ML) models. FI scores are assigned to individual features to quantify their impact on the model's decision. Recently, many existing FI measures were summarized in the removal-based explanation framework [1] using three components: Feature removal, model behavior, and summary technique. Therein, the impact of individual features, referred to as FI scores, is quantified with respect to the *model behavior*, a specific property of the ML model. This model behavior is then evaluated by measuring the impact of *removing* a group of features and *summarizing* these evaluations in a single FI score for each feature. Among other insights, it was shown that the specific feature removal technique yields different views on the process and hence substantially different meaningful explanations [2, 3, 4, 5]. Chen et al. [2] provide first guidance based on the application. While FI has been mainly considered in a static environment, many contemporary real-world applications require dynamic models that quickly adapt over time. Providing explanations, such as FI, in non-stationary environments is a challenging task, as access to previous observations is limited, and FI scores may change abruptly due to adaptation of the model caused by concept drift. Online learning constitutes an important technology when models need to deal with possibly non-stationary environments in realistic scenarios [6].

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

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

This leads to the algorithmic challenge how to efficiently adapt models based on novel data, and the learning challenge how to reliably deal with a possibly changing underlying data distribution. Since the latter is ill-posed, explanations play a particularly relevant role in the context of potential concept drift.

In this work, we are interested in algorithmic approaches for efficient feature removal in dynamic environments, that can be applied to FI in online learning scenarios on data streams and are updated incrementally with new observations. We compare a realization of two canonical choices of measures for feature removal, referred to as the *interventional (int.)* and *observational (obs.)* approach, respectively. We apply both on synthetic data streams using global FI measures and demonstrate that there are substantial semantically meaningful differences of these approaches. We further provide supporting arguments for the claim in [2] that the obs. approach is "true to the data", i.e., reflects the causal structure of the features, and the int. approach is "true to the model", i.e., considers the model independent of the causal structure of the inputs.

Related Work

Feature removal has been introduced as part of the removal-based explanation framework [1] that summarizes many popular FI measures. For model behavior, common choices include the dataset loss (global explanation), such as permutation tests [7] and SAGE [8], or an individual prediction (local explanation), such as SHAP [9].The Shapley value [10], due to its axiomatic structure, often constitutes the preferred summary technique over pairwise subset comparisons. For feature removal, perturbing inputs is a widely applied approach, which does not require to fit a new model. Perturbation techniques are distinguished in the int. and obs. approach [2], which either break or maintain the feature dependencies, respectively. They are also referred to as marginal expectation and conditional expectation [3] or off- and on-manifold explanations [4]. The int. approach has been used in permutation tests [7]. The obs. approach is a more general concept, that reduces to marginal distributions in case of feature independence [9]. It has been approximated using unsupervised models [4], tree-based model structure [11] or assumptions on the structure of the conditional distribution [5]. The obs. approach is often approximated using the marginal distribution, i.e. assuming feature independence and using the int. approach [9, 8] and both methods have been discussed in the literature [2, 3, 4, 5]. It was argued in favor of the obs. approach [4, 5] and the int. approach [3], whereas in [2] the authors argue that this choice depends on the application scenario and neither is preferable in general.

While XAI has mainly considered static environments so far, recent advances have considered explanations in the context of dynamic environments, such as online learning on data streams. In the context of removal-based explanations, global FI measures have been introduced that generalize SAGE [12] and PFI [13], where incremental extensions of the int. and obs. approach were introduced. These recent developments enable us to dive deeper into the effects of int. and obs. approaches in the incremental setting.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

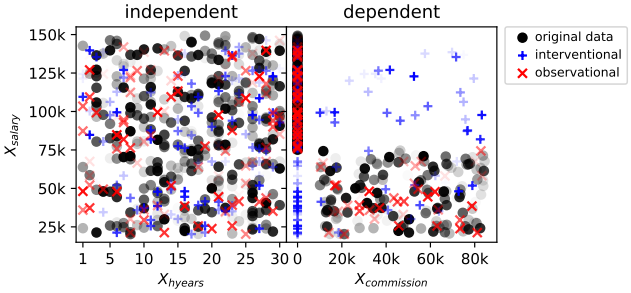

Fig. 1: The obs. (red) and int. (blue) approach lead to different perturbations on independent (left) and dependent (right) features. Int. removal creates synthetic data points that lie outside the data distribution.

2 Feature Removal in Dynamic Environments

In dynamic environments, we consider an unbound stream of data at time t as $(x_0, y_0), \ldots, (x_t, y_t)$ with a time-dependent model f_t, updated incrementally with each observation, and data-generating random variables (X_t, Y_t) at time t. Feature removal is a component of removal-based explanations [1], where the goal in a dynamic environment is to restrict the model $f_t : \mathcal{X} \to \mathcal{Y}$ on a d-dimensional input space $\mathcal{X}$ with features $D := \{1, \ldots, d\}$ to a subset $S \subset D$, while preserving accuracy. Formally, feature removal is executed by a (time-dependent) function $F_t : \mathcal{X} \times \mathcal{P}(D) \to \mathcal{Y}$, where $\mathcal{P}(D)$ refers to the power set of D, which enables the evaluation of the model for unknown features in $\bar{S} := D \setminus S$. While retraining the model on each subset of features is computationally prohibitive, we rely on a perturbation of the inputs of $\bar{S}$ for a given model f_t. One can distinguish between the *int.* and *obs.* approach [2]. The int. approach perturbs the features in $\bar{S}$ by using the marginal distribution, which breaks the feature dependency. In contrast, the obs. approach perturbs the features by using the conditional distribution given the present input values of features in S. Formally, these are defined [12] as

$$F_t^{\text{int}}(x, S) := \mathbb{E}\left[f_t(x^{(S)}, X_t^{(\bar{S})})\right] \text{ and } F_t^{\text{obs}}(x, S) := \mathbb{E}\left[f_t(X_t) \mid X_t^{(S)} = x^{(S)}\right],$$

where we write $f_t(x^{(S)}, x^{(\bar{S})})$ to distinguish between inputs of f_t in S, $x^{(S)}$, and inputs of f_t in $\bar{S}$, $x^{(\bar{S})}$. As the true data-generating distribution, in practice, is inaccessible, the expectation is approximated using Monte-Carlo integration, e.g. efficient implementations as proposed in the work [12, 13]. The int. approach relies on geometric sampling [13], where a reservoir of length L of observed data points is stored. Each new observation uniformly replaces an existing observation in the reservoir. To approximate $F_t^{\text{int}}(x, S)$, a random observation $\tilde{x}$ is drawn from the reservoir, where the feature values in $\bar{S}$ are considered, i.e. the model is evaluated with $f_t(x^{(S)}, \tilde{x}^{(\bar{S})})$. This yields to a geometric distribution

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

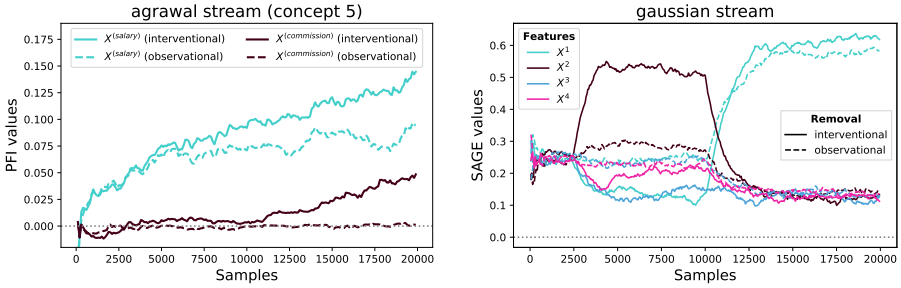

Fig. 2: Global FI for the Agrawal data stream using iPFI (left) and a synthetic multivariate Gaussian stream using iSAGE (right). Correlated features result in profoundly different FI scores.

for the probability of an observation to be chosen after $r > 0$ time steps. The obs. approach relies on a conditional subgroup approach [14], where a reservoir of length L is maintained using geometric sampling [13] for each conditional subgroup, which is represented as the leaf node of a decision tree. Similar to Molnar et al. [14], the subgroups are found individually for each feature by modeling each feature with respect to the rest. To maintain the subgroups dynamically, we use a Hoeffding Adaptive Tree (HAT) [12, 15]. For a subset $S \subset D$, the feature values for $\bar{S}$ are then inferred individually based on the reservoirs of the corresponding HAT. To find the subgroup, the HAT is traversed using the input x and the features in S, where the child is chosen based on random sampling, if a split node with features in $\bar{S}$ is encountered. This random sampling is based on the ratio of observed data points of each child node, which is an inherent statistic for HATs and parallels the TreeSHAP methodology [11]. The obs. approach thereby directly extends the conditional subgroup approach to arbitrary feature subsets and the incremental setting.

3 Experiments

We compare the int. and obs. approach on two synthetic data streams using the incremental variants of the global FI measures iPFI [13] and iSAGE [12][1].

Agrawal data stream. The Agrawal stream [16] is a well-established synthetic data stream for binary classification. In this experiment, we train an ARF [17] for concept 5 [16] and consider the *salary* $X^{\text{sal.}}$ and *commission* $X^{\text{com.}}$ feature. The commission feature depends on the uniformly distributed salary feature $X^{\text{sal.}} \sim \text{unif}(20k, 150k)$ with $X^{\text{com.}} = \mathbf{1}(X^{\text{sal.}} \leq 75k) \cdot Z$ with $Z \sim \text{unif}(10k, 75k)$, where $\mathbf{1}$ is the indicator function. The int. and obs. iPFI yield different perturbations for $X^{\text{sal.}}$ and $X^{\text{com.}}$, as illustrated in Fig. 1 (right), in

[1]For the implementation we refer to `https://github.com/mmschlk/On-Feature-Removal-for-Explainability-in-Dynamic-Environments`.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

contrast to the independent variables (left). Different perturbations then result in profoundly different explanations, as shown in Fig. 2 (left).

Obs. iPFI assigns a low FI score to $X^{\text{com.}}$, as the information is inferred from $X^{\text{sal.}}$. On the other hand, int. iPFI assigns a high FI score to $X^{\text{com.}}$, as the model uses this feature for prediction of the target variable in addition to $X^{\text{sal.}}$. Perturbing the inputs of the model with int. iPFI reveals this dependency, as the perturbed values for $X^{\text{com.}}$ do not depend on $X^{\text{sal.}}$.

Multivariate Gaussian distribution. In this experiment, we consider a multivariate Gaussian distribution $X_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t)$. The target variable does only depend on $X_t^{(1)}$ and is defined as $Y_t := \mathbf{1}(X_t^{(1)} > 0)$. At time $t_0 = 10k$, we induce a sudden concept drift: For $t < t_0$ we specify $\Sigma_t = 5 \cdot \mathbf{J}$, where $\mathbf{J}$ is a matrix of only ones, i.e. all features are highly correlated. For $t > t_0$, we specify $\Sigma_t = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, i.e. all features are uncorrelated. The resulting int. and obs. iSAGE values are shown in Fig. 2 (right).

When features are highly correlated or dependent, int. and obs. iSAGE yield profoundly different explanations. In the correlated setting, the HAT starts to split after around 2500 samples on $X^{(2)}$, as this variable is highly correlated with $X^{(1)}$, which is used in the classification function. This is reflected in a high FI score for $X^{(2)}$ of int. iSAGE, whereas all obs. iSAGE values remain relatively close, as the information of $X^{(2)}$ can be inferred from the remaining variables. After the concept drift, the features are independent and the HAT is forced to split on $X^{(1)}$ instead, which reflects the true classification function. Both, int. and obs. iSAGE then assign a high FI score to $X^{(1)}$, which confirms that both approaches yield similar results, if features are independent.

4 Conclusion

We compared incremental variants for int. and obs. feature removal in a dynamic environment. Our results confirm that this conceptional choice yields profoundly different explanations, in line with literature for the static setting [2, 3, 4, 5]. The difference in explanations arises from correlated features, which is expected as the obs. approach reduces to the int. approach in case of feature independence. We have shown on synthetic data streams that the int. approach reveals the model structure more reliably than the obs. approach, which aligns with the "true to the model" claim by Chen et al. [2]. In contrast, the obs. approach reveals the causal structure of the classification function more accurately, as it does not assign high FI scores to features that are replaceable by others, even though the model has learned to predict based on them. This confirms the notion of obs. explanations being "true to the data" [2]. Modeling perturbations for the obs. approach remains a challenging task, where we have discussed and compared one possible implementation using an extension of the conditional subgroup approach [14]. In future research other perturbation techniques, such as the unsupervised learning approach [4], can be extended to dynamic environments and analyzed rigorously to establish further guidance.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

References

[1] I. Covert, S. M. Lundberg, and S. Lee. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

[2] H. Chen, J. D. Janizek, S. M. Lundberg, and S. Lee. True to the model or true to the data? *CoRR*, abs/2006.16234, 2020.

[3] D. Janzing, L. Minorics, and P. Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 2020.

[4] C. Frye, D. Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

[5] K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[6] S. C. H. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[7] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[8] I. Covert, S. M. Lundberg, and S. Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems 33: (NeurIPS 2020)*, pages 17212–17223, 2020.

[9] S. M. Lundberg and S. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 4768–4777, 2017.

[10] L. S. Shapley. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, 1953.

[11] S. M. Lundberg, G. G. Erion, H. Chen, A. J. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[12] M. Muschalik, F. Fumagalli, B. Hammer, and E. Hüllermeier. iSAGE: An Incremental Version of SAGE for Online Explanation on Data Streams. *CoRR*, abs/2303.01181, 2023.

[13] F. Fumagalli, M. Muschalik, E. Hüllermeier, and B. Hammer. Incremental Permutation Feature Importance (iPFI): Towards Online Explanations on Data Streams. *CoRR*, abs/2209.01939, 2022.

[14] C. Molnar, G. König, B. Bischl, and G. Casalicchio. Model-agnostic feature importance and effects with dependent features - A conditional subgroup approach. *CoRR*, abs/2006.04628, 2020.

[15] A. Bifet and R. Gavaldà. Adaptive Learning from Evolving Data Streams. In *Proceedings of International Symposium on Intelligent Data Analysis (IDA 2009)*, pages 249–260, 2009.

[16] R. Agrawal, T. Imielinski, and A. Swami. Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):914–925, 1993.

[17] H. M Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, and T. Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9):1469–1495, 2017.