ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Explaining Outliers using Isolation Forest and Shapley Interactions

Roel Visser[1], Fabian Fumagalli[1], Maximilian Muschalik[2],
Eyke Hüllermeier[2], Barbara Hammer[1] *

1- Bielefeld University, CITEC, D-33615 Bielefeld, Germany

2- LMU Munich, MCML, D-80539 Munich, Germany

Abstract. In unsupervised machine learning, Isolation Forest (IsoForest) is a widely used algorithm for the efficient detection of outliers. Identifying the features responsible for observed anomalies is crucial for practitioners, yet the ensemble nature of IsoForest complicates interpretation and comparison. As a remedy, SHAP is a prevalent method to interpret outlier scoring models by assigning contributions to individual features based on the Shapley Value (SV). However, complex anomalies typically involve *interaction* of features, and it is paramount for practitioners to distinguish such complex anomalies from simple cases. In this work, we propose Shapley Interactions (SIs) to enrich explanations of outliers with feature interactions. SIs, as an extension of the SV, decompose the outlier score into contributions of individual features *and* interactions of features up to a specified explanation order. We modify IsoForest to compute SI using TreeSHAP-IQ, an extension of TreeSHAP for tree-based models, using the `shapiq` package. Using a qualitative and quantitative analysis on synthetic and real-world datasets, we demonstrate the benefit of SI and feature interactions for outlier explanations over feature contributions alone.

1 Introduction

Detecting outliers in data and identifying their cause is crucial for many machine learning applications, such as critical security systems [10]. In this context, Isolation Forest (IsoForest) [3] is an efficient method that detects outliers based on an ensemble of trees in an unsupervised setting. Yet, understanding the impact of features in IsoForest's detection algorithm is impossible due to the large number of splits and trees. As a remedy, the Shapley Value (SV) [9] is used to quantify feature attributions in the field of eXplainable Artificial Intelligence (XAI) for outlier metrics [6] or black-box predictions with SHAP [5].

While feature attribution methods indicate contributions of individual features, they do not quantify *feature interactions*, which provide valuable insights into the interplay between multiple features. To understand feature interactions, Shapley Interactions (SIs) [7], as an extension of the SV, have emerged recently as an important tool to enrich SHAP scores with interactions.

In this work, we propose SIs to quantify feature interactions in outlier detections. Specifically, we modify IsoForest and apply TreeSHAP-IQ [8] to compute

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

exact SIs of IsoForest, which are otherwise intractable. Using SIs, we demonstrate and address the limitations of existing outlier explanation methods, such as SHAP [5] or DIFFI [2], which fail to identify feature interactions in complex outliers. We conduct experiments in different synthetic and real-world scenarios, and show the added information that interactions provide allowing us to detect behavior, which feature attributions do not capture appropriately.

2 Background

IsoForest [3] is an efficient method for detecting outliers using ensembles of decision trees. By randomly generating splits on randomly sampled features, IsoForest scores outliers using the path lengths within each tree, where outliers are those points which are more susceptible to isolation. More formally, at each leaf of the decision tree the outlier score is defined as

$$\texttt{IsoTreeScore}(\text{tree, leaf, n}) := \texttt{depth}(\text{tree}, \text{leaf}) + c(n), \tag{1}$$

where $c(n)$ adjusts for the number of samples n observed in the leaf [3] and small values indicate outlier. While IsoForest is an efficient method for outlier detection, it is a black box model. Yet, understanding which feature value was responsible for the outlier detection is crucial in applications since outlier detection is essentially ill-posed [10].

SHAP [5] is a prevalent method to explain outputs of black box ML models. Given the output of a black box model $f : \mathbb{R}^d \to \mathbb{R}$ in a d-dimensional feature space $D = \{1, \ldots, d\}$, and given an instance $x_0 \in \mathbb{R}^d$, SHAP assigns real-valued attributions $\phi(i) \in \mathbb{R}$ to every feature $i \in D$, such that the model output at x_0 is decomposed as $f(x_0) = \phi_0 + \sum_{i \in D} \phi(i)$, where ϕ_0 is a baseline term. SHAP assigns these attributions by defining a cooperative game $\nu(T) := \mathbb{E}[f(X) \mid X_T = x_T]$, which captures the model's expected output restricted to any subset of features $S \subseteq D$, and computing the SV [9] as a weighted average over marginal contributions $\Delta_i(T)$,

$$\phi(i) := \sum_{T \subseteq D} \frac{1}{d \cdot \binom{d-1}{|T|}} \Delta_i(T) \text{ with } \Delta_i(T) := \nu(T \cup \{i\}) - \nu(T).$$

The SV can be motivated by reasonable axioms and uniquely allocates a *fair* contribution to each feature $i \in D$ [9]. In the context of outlier detection, we formally introduce the outlier explanation game.

Definition 1. *Given an outlier scoring $\eta : \mathbb{R}^d \to \mathbb{R}$, let $\eta_S(x) = \mathbb{E}[\eta(X) \mid X_S = x_S]$ be the expected scores restricted to any subset of features $S \subseteq D$. We then define the **outlier explanation game** of an instance x_0 as $\nu_{x_0}(T) := \eta_T(X)$.*

Due to the exponential complexity of the SV requiring 2^d game evaluations, SVs typically have to be approximated. However, for tree-based models TreeSHAP [4] reduces the complexity of computing exact SHAP scores to polynomial

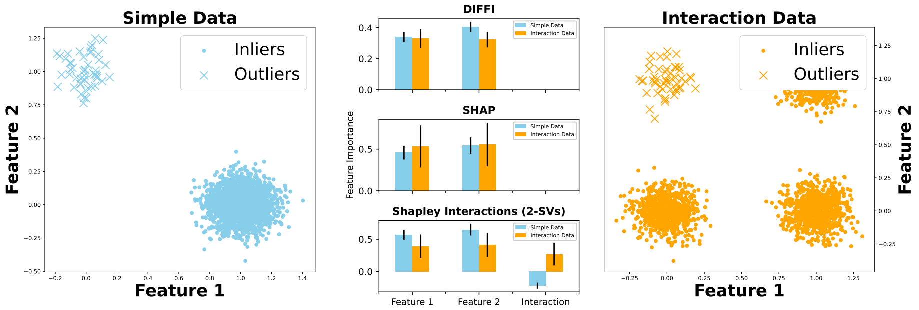

Fig. 1: Visualization of outlier dimensions (Feature 1, Feature 2) for synthetic datasets. Outliers in *simple data* (left) can be detected using each feature separately, while *interaction data* (right) requires both feature values. Individual feature attributions (middle, first and second row) do not display any differences. Yet, 2-SVs (middle, third row) showcase a negative interaction in simple data, indicating redundancy of both features. In contrast, the positive values for the interaction data indicates a synergy of these feature values for detecting outlier.

time by exploiting the tree structure to accelerate SV computation. Given a decision tree and a set of features $S \subseteq D$, TreeSHAP models ν by introducing weighted splits at each decision node of features in $D \setminus S$. The weights are determined by the conditional probabilities observed at each nodes, mimicking the conditional expectation, and allow us to efficiently compute the SV. Due to the linearity of SVs, SHAP scores of individual trees can be computed and aggregated into SHAP scores of the ensemble. In the context of outlier explanations, IsoForest exhibits a similar structure as any tree-based prediction model, and thus the outlier score. Hence, by introducing outlier scores at each decision and leaf node, TreeSHAP can be applied, as implemented in the `shap` package[1]. While TreeSHAP efficiently computes SHAP scores for IsoForest, it does not give any insights about interactions of features. However, in practice, both individual features and the interplay between multiple features may cause outliers, which we disentangle in the following by introducing SIs for outlier explanations.

3 Explaining Outliers with Shapley Interactions

Explanations of outliers that assign contributions to individual features highlight which feature value contributed to the detected anomaly. In practice, complex anomalies often involve a combination of features, indicating feature *interaction*. In such cases, it is crucial for practitioners to be able to distinguish complex anomalies with interactions from cases where a single individual feature value is responsible for the anomaly.

As an illustration, Figure 1 shows two synthetic settings. In both settings,

[1] `https://github.com/shap/shap`

the features x_1 and x_2 are responsible for the outliers, which is identified by SHAP. However, while it is possible to detect these outliers based on each individual feature in *simple data* (left), *interaction data* (right) requires knowing the values of both features, i.e. there exists a pairwise interaction. In the following, we introduce SIs to enrich SHAP explanations with feature interactions, which allows us to distinguish between such interactions and individual causes of outliers. Using TreeSHAP-IQ [8], we compute exact SIs for IsoForest—targeting the following question: Did the involved features individually or jointly cause the detected anomaly? This allows us to distinguish simple anomalies, where individual features are responsible, from complex interaction-based anomalies.

SIs [7], as an extension of the SV, decompose the model output additively into contributions for individual features *and* groups of features up to a given *explanation order* $k = 1, \ldots, d$. In this context, $k = 1$ is the SV, as the least complex explanation. In contrast, $k = d$ is the most complex explanation, which is the most faithful but entails 2^d components [7]. SIs thus yield an adjustable complexity-accuracy trade-off. Similar to the SV being a weighted average over *marginal contributions* $\Delta_i(T)$, SIs are a weighted average over discrete derivatives $\Delta_S(T)$, which directly extend marginal contributions to higher order. The discrete derivative of two features i, j is given for $T \subseteq N \setminus \{i, j\}$ by $\Delta_{\{i,j\}}(T) := \nu(T \cup \{i, j\}) - \nu(T) - \Delta_i(T) - \Delta_j(T)$, i.e. the effect of adding both features jointly, minus their individual contributions. In other words, a positive value indicates *synergy*, i.e. jointly knowing these features yields additional information about the cause of outlier, whereas a negative value indicates *redundancy*. SIs of order k can be formally defined with weights $w_k \geq 0$ as

$$\Phi_k(S) := \sum_{T \subseteq D \setminus S} w_k(|S|, |T|) \cdot \Delta_S(T) \text{ with } \Delta_S(T) := \sum_{L \subseteq S} (-1)^{|S|-|L|} \cdot \nu(T \cup L),$$

where $\nu(D) = \sum_{S \subseteq D} \Phi_k(S)$ additively decomposes the model output. The weights w_k depend on the sizes of S and T, and are SI-specific, where multiple choices have been proposed that differently extend the axioms of the SV [7], where we rely on k-SVs [4, 1]. Similar to the SV, SIs require an exponential number of evaluations. For tree-based models, TreeSHAP-IQ [8] extends TreeSHAP to SIs. Using Definition 1 with the modified IsoForest, we use TreeSHAP-IQ to efficiently compute any-order SIs.

4 Experiments

In this section, we empirically validate the capabilities of SIs to address limitations of existing feature attribution methods. We compare SIs (2-SVs) using the `shapiq`[2] package with DIFFI and SHAP as baselines. For visualization, we normalize all scores (higher is more important). All experiments are available at `https://github.com/r-visser/isolationforest-treeshapiq-paper`.

[2] `https://github.com/mmschlk/shapiq`

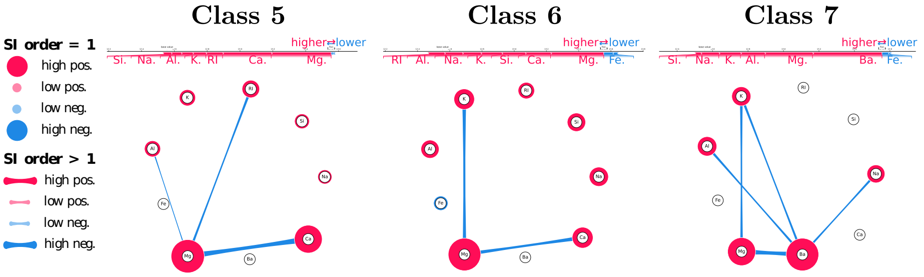

Fig. 2: Average positive (red) and negative (blue) SVs (force plot) and 2-SVs (graph plot) of all instances in the outlier classes (5,6,7) of the *glass* dataset.

Shapley Interactions on Synthetic Data. In this experiment, we generate synthetic outlier datasets, where we generate inliers and outliers based on two features X_1, X_2, and add two non-informative features $X_3, X_4 \sim \mathcal{N}(0, \sigma^2)$. For X_1, X_2, we generate *simple data*, illustrated in Figure 1 (left) with inliers centered at $[1, 0]$ and outliers centered at $[0, 1]$, perturbed by Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. For *interaction data*, illustrated in Figure 1 (right), we introduce additional inliers centered at $[0, 0]$ and $[1, 1]$, which are again perturbed by Gaussian noise. Clearly, outliers in *simple data* can be detected using either X_1 or X_2, whereas outliers in *interaction data* can only be detected with both values of X_1 and X_2. We generate 2500 instances for each inlier, and 100 for each outlier across 10 runs. We train IsoForest on a 20/80 train-test split of the outliers and using all generated inliers, which achieves an F1-score of 100%. We then compute explanations for all outliers using SHAP and DIFFI, as well as SIs (2-SVs). Figure 1 (middle) shows the average scores of X_1, X_2 and the interaction between both features. Individual feature attributions (first and second row) exhibit similar positive importance scores for features x_1 and x_2 but fail to identify differences in the type of outlier. In contrast, SIs (third row) show a negative interaction value for *simple data*, which indicates the redundancy of both features for the outlier detection task. While for the *interaction data*, this interaction is positive, indicating the synergy between both features, which is crucial for detecting outliers.

Shapley Interactions on Real-World Data. In this experiment, we investigate SIs on the real-world dataset *glass* [2]. It consist of 213 glass samples containing the glass' refractive index and the concentration of 8 chemical compounds. The dataset contains 7 classes. We replicate the experimental setup from [2], in which classes 1-4 are used as inliers. We train IsoForest on the inliers and outlier classes 5 and 6, yielding an F1-score of 98%. For class 7 we evaluate quantitatively whether the explanations of each method is able to detect the most relevant features for this class given the ground-truth labels *Barium (Ba.)* and *Aluminium (Al.)*, proposed by [2]. Table 1 shows Receiver Operating Characteristic (ROC) - Area Under the Curve (AUC) score [10] for

DIFFI, SHAP and SIs (2-SVs). Clearly, SIs outperform DIFFI and SHAP with smaller standard deviation. We also qualitatively evaluate the explanations for each outlier, comparing the difference in explanation between classes 5, 6, and 7. Figure 2 shows the average explanations produced by both SHAP, showing the SV (upper row), and TreeSHAP-IQ, giving the SIs as 2-SVs (lower row), across all instances in each class. We observe that there are clear differences between each class and many redundancies (negative interactions) between the most important features. Contrary to [2] we find that besides *Ba.* and *Al.*, *Magnesium (Mg.)* is also highly important for class 7.

DIFFI	**SHAP**	**2-SVs**
.81 (.23)	.84 (.16)	.96 (.04)

Table 1: *Glass* data ROC-AUC

5 Conclusion

We introduced SIs to interpret outliers detected by IsoForest. In contrast to feature attributions, like DIFFI or SHAP, SIs are capable of distinguishing between outliers that are caused by single individual features as well as combinations of multiple features, known as interactions. To compute SIs on IsoForest, we introduced the outlier explanation game and modified IsoForest's scoring to apply TreeSHAP-IQ [8]. In synthetic and real-world settings, we addressed limitations of feature attribution methods, such as SHAP and DIFFI, where SIs enable us to detect outliers caused by interactions of multiple features.

References

[1] S. Bordt and U. von Luxburg. From Shapley Values to Generalized Additive Models and back. In *AISTATS*, 2023.

[2] M. Carletti, M. Terzi, and G. A. Susto. Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. *Engineering Applications of Artificial Intelligence*, 119:105730, 2023.

[3] F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *ICDM*, pages 413–422. IEEE Computer Society, 2008.

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

[5] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, pages 4768–4777, 2017.

[6] M. Mayrhofer and P. Filzmoser. Multivariate outlier explanations using shapley values and mahalanobis distances. *Econometrics and Statistics*, 2023.

[7] M. Muschalik, H. Baniecki, F. Fumagalli, P. Kolpaczki, B. Hammer, and E. Hüllermeier. shapiq: Shapley interactions for machine learning. In *NeurIPS Datasets and Benchmarks*, 2024.

[8] M. Muschalik, F. Fumagalli, B. Hammer, and E. Hüllermeier. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *AAAI*, 2024.

[9] 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.

[10] J. Tritscher, A. Krause, and A. Hotho. Feature relevance xai in anomaly detection: Reviewing approaches and challenges. *Frontiers in Artificial Intelligence*, 6:1099521, 2023.