Proceedings of Machine Learning Research 230:1–17, 2024 Conformal and Probabilistic Prediction with Applications

Estimating Quality of Approximated Shapley Values Using Conformal Prediction

Amr Alkhatib ALKHAT@KTH.SE
Henrik Boström BOSTROMH@KTH.SE
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden
Ulf Johansson ULF.JOHANSSON@JU.SE
Dept. of Computing, Jönköping University, Sweden

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Thanks to their theoretically proven properties, Shapley values have received a lot of attention as a means to explain predictions within the area of explainable machine learning. However, the computation of Shapley values is time-consuming and computationally expensive, in particular for datasets with high dimensionality, often rendering them impractical for generating timely explanations. Methods to approximate Shapley values, e.g., FastSHAP, offer a solution with adequate computational cost. However, such approximations come with a degree of uncertainty. Therefore, we propose a method to measure the fidelity of Shapley value approximations and use the conformal prediction framework to provide validity guarantees for the whole explanation in contrast to an earlier approach that offered validity guarantees on a per-feature importance basis, disregarding the relative importance of the remaining feature scores within the same explanation. We propose a set of difficulty estimation functions devised to consider the difficulty of explanation approximations. We provide a large-scale empirical investigation where the proposed difficulty estimators are evaluated with respect to their efficiency (interval size) in measuring the similarity to the ground truth Shapley values. The results suggest that the proposed approach can provide predictions coupled with informative validity guarantees (tight intervals), allowing the user to trust/reject the provided explanations based on their similarity to the ground truth values.

Keywords: Conformal prediction · Explainable machine learning

1. Introduction

Advanced machine learning algorithms tend to generate black-box models, which limits their application in real-world problems across various domains when explainability is a central prerequisite, e.g., in domains such as medicine and law (Lakkaraju et al., 2017). Using interpretable models may address the explainability issue; however, employing white-box models, e.g., linear models and decision trees, often fails to achieve the predictive accuracy of state-of-the-art black-box models (Loyola-González, 2019). Consequently, explainable machine learning provides a solution to the explainability problem without sacrificing performance. Explanations can be provided through model-agnostic methods that explain any model regardless of the used algorithm or alternatively using a model-specific method that exploits the architecture and characteristics of the black-box model (Bénard et al., 2021; Boström et al., 2018; Wang et al., 2021; Ying et al., 2019). LIME (Ribeiro et al., 2016) and

SHAP (Lundberg and Lee, 2017) are prominent examples of model-agnostic methods that provide explanations in the form of feature importance scores that reflect the relative importance of the features with respect to the prediction. However, explanations generated using model-agnostic methods such as LIME and SHAP are computationally expensive, as LIME creates local white-box surrogate models to explain predictions, and SHAP approximates Shapley values.

Methods, e.g., INVASE (Yoon et al., 2019), REAL-X (Jethani et al., 2021), and L2X (Chen et al., 2018), have been proposed to reduce the computational cost of model-agnostic explanations. FastSHAP (Jethani et al., 2022) is one such approach that approximates Shapley values. Shapley values have been shown to provide a unique solution in the class of additive feature attribution methods that satisfies the following properties (Lundberg and Lee, 2017):

- local accuracy: the explanation matches the model
- missingness: a missing feature is attributed a value of zero
- consistency: if the contribution of a feature increases or remains unchanged, the Shapley value increases or remains unchanged

However, since FastSHAP provides mere approximations of the true Shapley values using a pre-trained neural network (the explainer) the provided approximations may be of close or far proximity to the true Shapley values. In previous work (Alkhatib et al., 2023), we proposed using the conformal prediction framework to provide validity guarantees on the approximated importance scores, i.e., instead of providing a point value for each approximated score, a prediction interval containing the true Shapley value at a specified confidence level was produced. However, the validity is, by this approach, granted per feature score; the Shapley value for each feature is handled as a separate regression problem. Hence, this approach only provides a local guarantee for the approximation without taking into account the order of the features or the general orientation of the approximate explanation vector in relation to the ground truth. Therefore, in this work, we propose and investigate an alternative approach, by which the entire explanation (feature attribution) is taken into account, while still providing a validity guarantee through the inductive conformal prediction framework (Vovk et al., 2005).

The main contributions of the study are:

- an approach for measuring the similarity of approximated Shapley explanations to the ground truth Shapley values accompanied with validity guarantees obtained using the inductive conformal prediction framework
- a set of difficulty estimation functions designed to measure the difficulty of approximating an explanation
- a large-scale empirical investigation using 20 publicly available datasets comparing the performance efficiency of the proposed difficulty estimators

In the following section, we provide a background on explainable machine learning as well as conformal prediction. Section 3 provides an overview of related work. Section 4 describes the proposed methodology for quantifying the uncertainty of the approximated explanations. Section 5 presents and discusses the results of the empirical investigation. Finally, Section 6 outlines the main findings and points out directions for future work.

2. Background

In the first part of this subsection, we provide a short summary of the different explainable machine learning methods and some of their limitations. In the second part, We briefly overview inductive conformal prediction for regression models and provide pointers to possible non-conformity functions and difficulty estimates.

2.1. Machine Learning Explanation Methods

An explanation, in the context of machine learning, refers to a clear understanding or rationale provided for a model's predictions or decisions. Explanations based on Shapley values and other feature attribution methods typically involve determining the contribution of the input variable to the model's output (Molnar, 2022). For instance, Shapley values quantify each feature's impact on a particular prediction, offering insights into the model's decision-making process and helping to clarify why a model made a specific prediction or classification.

Explanation methods provide explanations in various forms. Plots, such as Partial Dependence Plot (PDP) (Friedman, 2001) and Accumulated Local Effects (ALE) Plot (Apley and Zhu, 2020), offer intuitive visual explanations. Rule-based approaches, like Anchors (Ribeiro et al., 2018), offer explanations in a structured format, often favored for their clarity. Alternatively, methods like LIME and SHAP provide additive feature importance scores that can be summed to get the prediction of the black-box model. We illustrate the explanations generated using LIME and SHAP for the same prediction made by an XGBoost model in Figure 1 and Figure 2, respectively.

Despite their advantages, employing explanation methods poses some challenges. One issue is fidelity, which concerns how accurately an explanation reflects the model's decision-making process. Many explanation methods that produce explanations in the form of feature importance scores lack an objective means of verification. For instance, the explanations provided in Figure 1 and Figure 2 are essentially for the same exact prediction, yet they differ. Therefore, the user can expect one explanation to be more precise (has higher fidelity) than the other. However, such fidelity to the underlying black-box model is undetermined, and we aim to provide a solution for this problem in the following sections.

Rule-based methods, e.g., Anchors (Ribeiro et al., 2018), offer explanations in a user-friendly format that can be easily interpreted as a simple rule with some conditions in the antecedent that lead to a consequent prediction. The rules can also be applied to predictions, allowing for measurement of their alignment with the underlying black-box model. Nonetheless, the fidelity of certain explanatory rules may be low (Delaunay et al., 2020), partly due to the complexity of the underlying model and the challenge of encapsulating it within a single explanatory rule.

Computational cost presents another challenge. For instance, KernelSHAP (Lundberg and Lee, 2017) and LIME employ a permutation-based approach, which involves computing the model's predictions on many subsets of features, imposing computational overhead.

Such costs may limit the practical application of explanation methods, particularly with large datasets or time constraints. Therefore, approaches to provide fast approximations, e.g., FastSHAP (Jethani et al., 2022), have been introduced. However, the fidelity of such approximations may raise doubts.

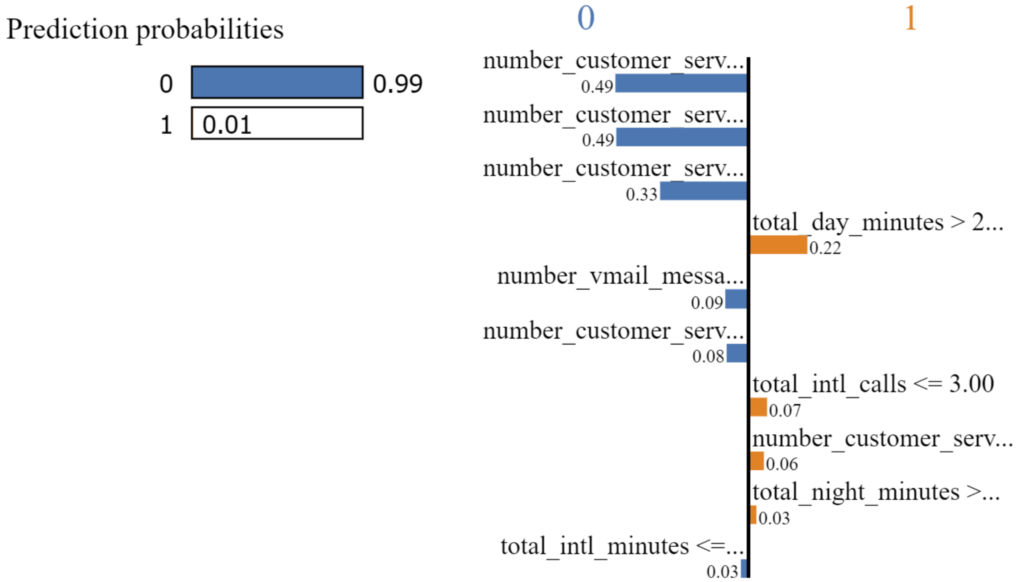

Figure 1: Example generated through LIME for a single data object from the Churn dataset

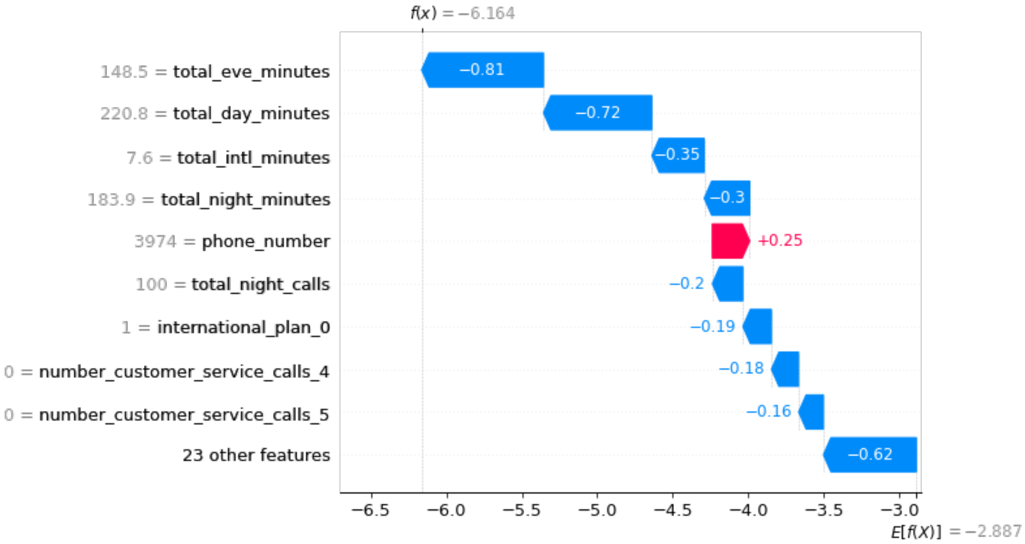

Figure 2: Example generated through SHAP for a single data object from the Churn dataset

2.2. Conformal Prediction

Conformal Prediction (CP) has been introduced to provide guarantees on prediction errors using a user-specified confidence level to restrict the probability of incorrect predictions (Johansson et al., 2014). CP was proposed as a transductive method that requires constructing a model for each data instance (Gammerman et al., 1998; Saunders et al., 1999), which is computationally demanding. Vovk et al. (2005) introduced the Inductive Conformal Prediction (ICP) to overcome the computational cost problem, where a single model is generated from the available data and used to predict new instances.

2.2.1. Inductive Conformal Prediction for Regression

Let X represent the feature space and $y \in \mathbb{R}$ denote the target variable. Consider a dataset $Z = \{z_1, z_2, ..., z_n\}$, where each $z_i = (x_i, y_i)$, with $x_i \in X$ and $y_i \in \mathbb{R}$. Assuming the data examples are independent and identically distributed (i.i.d), the inductive conformal regression involves the following steps:

1. Divide the dataset Z into a training subset $Z_t = \{z_1, z_2, ..., z_m\}$ and a calibration subset $Z_c = \{z_{m+1}, z_{m+2}, ..., z_n\}$.

2. The underlying model $\mathcal{M}$ is trained using Z_t.

3. Compute a non-conformity score α_i using a non-conformity function for all data example $z_i \in Z_c$, which produces a sequence $S = \{\alpha_{m+1}, \alpha_{m+2}, ..., \alpha_n\}$. The absolute error can be a simple non-conformity function (Papadopoulos et al., 2002):

$$\alpha_i = |y_i - \tilde{y}_i| \tag{1}$$

   where $\tilde{y}_i$ is the prediction made by the underlying model $\mathcal{M}$ for data example z_i.

4. Using the sequence of non-conformity scores S and a predefined significance level ϵ, find the smallest $\alpha_\epsilon \in S$ such that:

$$\frac{|\{z_i \in Z_c | \alpha_i < \alpha_\epsilon\}| + 1}{|Z_c| + 1} \geq 1 - \epsilon \tag{2}$$

   α_ϵ provides a probabilistic bound for the non-conformity scores at the significance level ϵ. Thus, the non-conformity score of a new data instance x_{n+1} will be less than or equal to α_ϵ with a probability of 1-ϵ. The interval covering the true prediction with a probability of 1-ϵ is given by:

$$\tilde{\mathcal{Y}}_{n+1} = [\tilde{y}_{n+1} - \alpha_\epsilon, \tilde{y}_{n+1} + \alpha_\epsilon] \tag{3}$$

2.2.2. Normalized Non-Conformity Measures

Since some predictions are more accurate than others, smaller intervals can be generated for the more accurate cases (easier cases). Therefore, the non-conformity score α_i for instance x_i can be adjusted by a difficulty estimate σ_i:

$$\alpha_i = \frac{|y_i - \tilde{y}_i|}{\sigma_i} \tag{4}$$

Consequently, the predicted interval for a new data instance x_{n+1} is given by:

$$\tilde{\mathcal{Y}}_{n+1} = [\tilde{y}_{n+1} - \alpha_\epsilon \sigma_{n+1}, \tilde{y}_{n+1} + \alpha_\epsilon \sigma_{n+1}] \tag{5}$$

The difficulty can be estimated using different possible functions, Papadopoulos et al. (2011) for instance, suggested estimating the difficulty using the k-nearest neighbors (KNN), which involves calculating the sum of the distances between x_i and its k-nearest neighbors as follows:

$$d_i^k = \sum_{\eta \in \mathcal{N}} distance(x_i, x_\eta) \tag{6}$$

where $\mathcal{N}$ represents the set of the k-nearest neighbours. Afterwards, d_i^k is normalized using the median distance value from the training data:

$$\lambda_i^k = \frac{d_i^k}{median(d_j : z_j \in Z_t)} \tag{7}$$

The non-conformity score then is given by:

$$\alpha_i = \left| \frac{y_i - \tilde{y}_i}{\gamma + \lambda_i^k} \right|, \tag{8}$$

and

$$\alpha_i = \left| \frac{y_i - \tilde{y}_i}{exp(\gamma \lambda_i^k)} \right|. \tag{9}$$

where γ is a parameter that controls the sensitivity of the measure to any changes in λ_i^k and $\gamma \geq 0$.

Papadopoulos et al. (2011) provided another difficulty estimate that relies on the variation in the labels of the k-nearest neighbors as measured by their standard deviation. A high agreement between the k-nearest neighbors indicates a more accurate prediction (an easy case). The standard deviation of the labels of the k-nearest neighbors s_i^k is computed as follows:

$$s_i^k = \sqrt{\frac{1}{k} \sum_{j=1}^{k} (y_{i_j} - \overline{y_{i_{1,2,\ldots,k}}})^2}, \tag{10}$$

where

$$\overline{y_{i_{1,2,\ldots,k}}} = \frac{1}{k}\sum_{j=1}^{k} y_{i_j} \tag{11}$$

The s_i^k can also be normalized:

$$\delta_i^k = \frac{s_i^k}{median(s_j : z_j \in Z_t)} \tag{12}$$

Accordingly, the non-conformity measure can be calculated as follows:

$$\alpha_i = \left|\frac{y_i - \tilde{y}_i}{\gamma + \delta_i^k}\right|, \tag{13}$$

and

$$\alpha_i = \left|\frac{y_i - \tilde{y}_i}{exp(\gamma\delta_i^k)}\right|. \tag{14}$$

One more non-conformity measure can be obtained by combining λ_i^k and δ_i^k as follows:

$$\alpha_i = \left|\frac{y_i - \tilde{y}_i}{\gamma + \lambda_i^k + \delta_i^k}\right|, \tag{15}$$

and

$$\alpha_i = \left|\frac{y_i - \tilde{y}_i}{exp(\gamma\lambda_i^k) + exp(\rho\delta_i^k)}\right|, \tag{16}$$

Where ρ is a parameter that controls the sensitivity of the measure to changes in δ_i^k.

3. Related Work

In this section, we briefly review previous approaches to quantify the uncertainty of explanations.

In (Schulz et al., 2022), tools were provided for evaluating the reliability of LIME explanations by estimating the uncertainty associated with a particular explanation, through evaluating the ordinal consensus among various bootstrapped surrogate explainers. In (Zhang et al., 2022), a framework was proposed to improve the explainability and quantify uncertainty in deep learning models for image classification. The proposed framework includes employing a Bayesian neural network, using entropy to estimate the uncertainty, and combining prediction difference analysis and the Bayesian neural network to quantify the uncertainty in feature importance and enhance model explainability.

Importance scores can be complex for human comprehension when explaining high-dimensional data and multimodal machine-learning models. Therefore, Folgado et al. (2023) proposed quantifying the explained models' uncertainty and rejecting the ones with high uncertainty, which reduces the number of modalities required to explain the predictions, resulting in a lower complexity explanation. Yang and Li (2023) have developed an explainable method for a domain application that can capture both aleatoric and epistemic uncertainties. The method targets molecular property prediction using deep learning models and can attribute the uncertainties to specific atoms in the input molecule. Cohen et al. (2023) proposed the quantification of uncertainty in stochastically estimated Shapley value explanations using the variance of the additive Shapley values.

Venn prediction has been employed to quantify the uncertainty of explanations (Alkhatib et al., 2022). However, the method considers explanatory rules only and does not apply to Shapley values. A set of metrics has been proposed that allows different explanations to be compared as solitary or groups of explanatory rules. The explanatory rules generated by Anchors are compared to those obtained from applying association rule mining to the Anchors' explanations. The findings indicate that association rule mining yields explanations with lower uncertainty compared to Anchors; additionally, rules obtained through association rule mining result in more informative predictions with tighter intervals.

Alkhatib et al. (2023) proposed the use of conformal regression to quantify the uncertainty of the approximated score-based explanations. Their method involves approximating any score-based explanation method using a much faster trained multi-target regression model in order to save computational cost and then provide validity guarantees using the inductive conformal prediction framework. The validity guarantees are provided at the level of individual features, and not for the whole explanation, by generating a prediction interval per approximated feature importance score, where each approximated score is considered a regression target. The prediction interval is computed using a non-conformity function and a calibration set. The findings of the empirical investigation in (Alkhatib et al., 2023) demonstrate that the approach can generate informative intervals covering the true importance scores of the underlying explanation method at the desired level of confidence.

4. Method

In this section, we first introduce the proposed approach to estimate the similarity of an approximate explanation to the ground truth explanation, while providing a validity guarantee following the use of the inductive conformal prediction framework. We then propose three difficulty estimation functions that are suitable for approximating explanations using conformal regressors.

4.1. Fidelity Estimation

The proposed method can, in principle, be applied to any fast explanation algorithm that produces approximations of additive feature importance scores in order to save the high cost of the exact computations. Here, we will focus on the case of FastSHAP (Jethani et al., 2022), a computationally efficient algorithm to approximate Shapley values contrasted with a computationally intensive yet precise method, e.g., unbiased KernelSHAP (Covert and Lee, 2021). We aim to assess the fidelity of approximate explanations compared to precise ones and provide validity guarantees on their similarity using the conformal prediction framework.

Given the impracticality of computing exact Shapley values during inference time to evaluate the fidelity of FastSHAP's approximate explanations, we assume that the similarity of an approximate explanation $\hat{\phi}$ to the exact (ground truth) explanation ϕ can be computed using a function f that takes the feature vector $(\mathbf{x})$, $\hat{\phi}$, and the predicted outcome by the black-box model (p) as inputs and produces (y) a similarity score to the ground truth explanation as an output. A similarity score (y) in this context represents a quantitative measure, chosen by the user, of how similar or close the approximate explanation ($\hat{\phi}$) is to the ground truth explanation (ϕ), e.g., cosine similarity or Euclidean distance. In this study, we will consider cosine similarity for this purpose, as it conveniently provides values between 1 and -1, which allows for comparing the sizes of the intervals generated by the conformal prediction framework using different models and comparing intervals across different datasets. Consequently, we propose that a mapping function ($f(\mathbf{x}, \hat{\phi}, p; \theta) = y$) can be learned using a machine learning model ($\mathcal{A} : f(\mathbf{x}, \hat{\phi}, p; \theta) = y$), where θ represents the model parameters. Given that the predicted outcome y represents a similarity score, model $\mathcal{A}$ can be trained as a regression model, and the conformal prediction framework can be employed to provide validity guarantees on the predictions of $\mathcal{A}$.

A proper training set $\boldsymbol{X}^{(train)}$ with the ground truth similarity scores ($\mathbf{y}^{(train)}$) is provided. The black-box model generates predictions $\mathbf{p}^{(train)}$ for the data objects in $\boldsymbol{X}^{(train)}$. The explanation method generates explanations ($\hat{\Phi}^{(train)}$) for the predicted outcomes. The input features $\boldsymbol{X}^{(train)}$, explanations $\hat{\Phi}^{(train)}$, the predictions by the black-box model $\mathbf{p}^{(train)}$, and targets $\mathbf{y}^{(train)}$ collectively constitute set $\boldsymbol{Z}^{(train)} = \{((\mathbf{x}_1, \hat{\phi}_1, p_1), y_1)$, $((\mathbf{x}_2, \hat{\phi}_2, p_2), y_2)$, ..., $((\mathbf{x}_n, \hat{\phi}_n, p_n), y_n)\}$, upon which the regression model $\mathcal{A}$ is trained.

The regression model $\mathcal{A}$ predicts the similarity between the approximate and ground truth explanations, while the conformal prediction framework offers a means to restrain the error level associated with each prediction. Hence, a calibration set $\boldsymbol{Z}^{(cal)}$ is devised, similar to the training set, where $\boldsymbol{Z}^{(cal)} = \{((\mathbf{x}_{n+1}, \hat{\phi}_{n+1}, p_{n+1}), y_{n+1})$, $((\mathbf{x}_{n+2}, \hat{\phi}_{n+2}, p_{n+2}), y_{n+2})$, ..., $((\mathbf{x}_c, \hat{\phi}_c, p_c), y_c)\}$. The estimated similarity scores $\hat{\mathbf{y}}^{(cal)}$ produced by the regression model $\mathcal{A}$, altogether with the ground truth targets $\mathbf{y}^{(cal)}$ are used to compute a non-conformity score α_i for each data objects $\mathbf{x}_i$ in $\boldsymbol{Z}^{(cal)}$. Consequently, if α_ϵ is the non-conformity score at a significance level ϵ, then during inference time, all values with a distance exceeding α_ϵ are disregarded. Finally, the interval that guarantees the probability of covering the true target is at least 1-ϵ is constructed as follows:

$$\tilde{\mathcal{Y}}_i = [\tilde{y}_i - \alpha_\epsilon, \tilde{y}_i + \alpha_\epsilon] \tag{17}$$

The idea of the proposed approach is outlined in Algorithm 1.

4.2. Difficulty Estimators

The non-conformity score can be normalized using a difficulty estimate φ_i, where some predictions are expected to be more precise than others. Subsequently, smaller intervals are anticipated for more accurate (easier) cases. The difficulty estimates can be computed using different approaches, e.g., the distance to the k-nearest neighbors (KNN) (Papadopoulos et al., 2011) or employing a trained model. Accordingly, we propose the following three functions to estimate the difficulty of the prediction designed to align with the primary objective of generating explanations and evaluating their fidelity:

- **Probability of the explanation:** we assume that if the prediction inferred from the importance scores of the explanation shows a high probability of one class, then it is likely an easy example to explain. Consequently, we propose the following difficulty estimate:

Algorithm 1 Fidelity Assessment with Conformal Prediction

```
Data: training set $\boldsymbol{Z}^{(train)}$, calibration set $\boldsymbol{Z}^{(cal)}$, significance level $\epsilon$, non-conformity measure $\Delta$
Result: Fidelity prediction model $\mathcal{A}$, non-conformity threshold $\alpha_\epsilon$
Initialize parameters $\theta$ of $\mathcal{A}$
$\mathcal{L} \leftarrow 0$
for $((\boldsymbol{x}_i, \hat{\phi}_i, p_i), y_i)$ in $\boldsymbol{Z}^{(train)}$ do
    $\tilde{y}_i \leftarrow \mathcal{A}(\mathbf{x}_i, \hat{\phi}_i, p_i)$
    $\mathcal{L} \leftarrow \mathcal{L}$+loss$(\tilde{y}_i, y_i)$
end
Compute gradients $\nabla_\theta \mathcal{L}$
Update $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}$
$C \leftarrow []$
for $((\boldsymbol{x}_j, \hat{\phi}_j, p_j), y_j)$ in $\boldsymbol{Z}^{(cal)}$ do
    $\tilde{y}_j \leftarrow \mathcal{A}(\mathbf{x}_j, \hat{\phi}_j, p_j)$
    $\alpha_j \leftarrow \Delta(\tilde{y}_j, y_j)$
    append $C \overset{+}{\leftarrow} \alpha_j$
end
$\alpha_\epsilon \leftarrow$ value in $C$ at significance level $\epsilon$
```

$$\varphi_i = 0.5 - \left| \frac{1}{1 + e^{-(\sum \hat{\phi})}} - 0.5 \right| \tag{18}$$

- **Probability difference:** the assumption is that if the prediction formulated using the importance scores deviates from the prediction of the black box, then the example is harder and vice versa

$$\varphi_i = \left| \frac{1}{1 + e^{-(\sum \hat{\phi})}} - \mathcal{B}(x_i; \theta) \right| \tag{19}$$

  where $\mathcal{B}(x_i; \theta)$ is the predicted probability by the black-box model.

- **Similarity to null:** an explanation is generated for a feature vector with all features masked ($\phi^{(null)}$), e.g., all zero features, when standard normalization is employed for preprocessing, which serves as a baseline explanation or explanation of $p(y = 1)$, which is compared to the explanation of each instance or $p(y = 1|x_i)$. Consequently, we measure the cosine similarity between the two explanations, and if they are different from each other, then the example is assumed to be easier to explain:

$$\varphi_i = 1 - \left| \frac{\phi^{(null)} \hat{\phi}_i}{\|\phi^{(null)}\| \|\hat{\phi}_i\|} \right| \tag{20}$$

5. Empirical Evaluation

This section evaluates the proposed fidelity estimation method and compares the three difficulty estimation functions on different datasets. The quality of the approximated explanations is assessed based on their similarity to ground truth explanations, and the difficulty estimation functions are evaluated based on the informativeness of the generated intervals using the inductive conformal prediction.

5.1. Experimental Setup

In the following experiments, 20 publicly available datasets[1] are used in the evaluation. Each dataset is split into a training set, test set, and calibration set. The ratios of the splits may differ according to the size of the dataset; however, the majority were split into 60% training, 20% calibration, and 20% test. The training set is used to train the black-box model, the explainer to approximate explanations for the black box, and the regression model to predict the similarity of the approximate explanations to the ground truth values. The calibration set is employed to compute the non-conformity and confidence percentile scores. The conformal regressors are obtained using the `crepes`[2] package (Boström, 2022). Finally, the test set is used to evaluate the quality of the generated explanations and compare the proposed difficulty estimation functions. The categorical features are binarized using one-hot encoding.

The black-box models ($\mathcal{B}$) are obtained using the XGBoost algorithm. The models are trained using the default set of hyperparameters of XGBoost. The explainers are learned using FastSHAP. The ground truth Shapley values (Φ) for the explanations of the calibration and the test sets are obtained using the unbiased KernelSHAP[3], shown by Covert and Lee (2021), to converge to the true Shapley values if a sufficiently large number of samples is provided. The similarity between explanations is measured using cosine similarity, which measures the similarity in the orientations of the compared vectors. The regression models ($\mathcal{A}$) to estimate the similarity of the approximate explanations to the ground truth are gradient-boosting regressors with 600 estimators. All the experiments are done at 0.95 confidence level.

5.2. Experimental Results

The required validity guarantees are ensured by the inductive conformal regression, and the generated intervals cover the true similarity values between the approximate and the ground truth explanations at the predefined confidence level $\epsilon = 0.95$. Therefore, the proposed difficulty estimates are compared based on the efficiency of the generated intervals, i.e., how tight the generated intervals are. In the following experiment, we compare the efficiency of the following difficulty estimation functions proposed in Subsection 4.2, which are: the probability of computed using the explanation scores (Exp. Prob.), the probability difference between the explanation and the black-box model prediction (Prob. Diff.), and the similarity to the default explanation when all the input features are masked (Sim. to

[1] The datasets were obtained from https://www.openml.org

[2] https://github.com/henrikbostrom/crepes

[3] The following efficient online implementation of KernelSHAP is used: https://github.com/iancovert/shapley-regression

null). We also compare proposed functions to the intervals without any difficulty estimation (baseline) and the distance to the k-nearest neighbors (KNN). The similarity between the ground truth and the approximations of FastSHAP is computed using the cosine similarity. Therefore, the valid range of the predicted similarities spans between 1 and -1, where 1 indicates a complete agreement, and -1 indicates a complete dissimilarity. The difficulty estimation functions are evaluated based on the average size of the predicted intervals across the test set. The results of using the 5 mentioned difficulty estimation functions are shown in Table 1. The results show that the explanation probability of the approximated scores provides the most efficient intervals followed by the baseline. In order to test the null hypothesis that the differences between the compared difficulty estimation functions are not significant, the Friedman (Friedman, 1939) test is applied, and the result of the Friedman test allows the rejection of the null hypothesis at the 0.05 level. Consequently, the post-hoc Nemenyi test (Nemenyi, 1963) is applied to detect which pairwise differences are significant. The results of the post-hoc test are summarized in Figure 3. The result of the post-hoc Nemenyi test shows that the differences between the efficiency of the explanation probability and both KNN and the baseline are insignificant, while it significantly outperforms the similarity to the default explanation and the probability difference between the black box and the explanation.

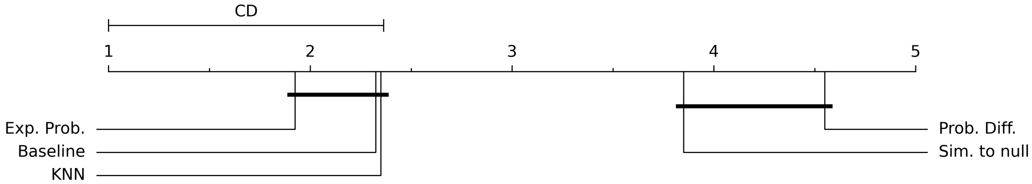

Figure 3: The average rank of the compared difficulty estimation functions on the 20 datasets. The ranking is conducted with respect to the interval size, where a lower rank is better (tighter intervals). The critical difference (CD) represents the largest difference that is not statistically significant.

Table 1: The average interval size of the similarity of the approximated explanation to the ground truth. The smallest average interval size is highlighted in light blue.

Dataset	Baseline	KNN	Prob. Diff.	Exp. Prob.	Sim. to Null
Abalone	0.119	0.078	0.313	0.115	0.115
Ada Prior	0.398	0.433	0.783	0.404	0.567
Bank32nh	0.226	0.247	0.262	0.228	0.276
Breast Cancer	0.102	0.064	0.129	0.078	0.106
Churn	0.495	0.577	0.637	0.495	0.513
Delta Ailerons	0.146	0.138	0.187	0.143	0.218
Electricity	0.136	0.126	0.186	0.129	0.194
Elevators	0.023	0.021	0.032	0.023	0.03
JM1	0.23	0.188	0.596	0.225	0.24
Heloc	0.314	0.322	0.98	0.309	0.346
MagicTelescope	0.077	0.069	0.107	0.072	0.079
MC1	0.15	0.091	0.155	0.152	0.187
Mozilla4	0.415	0.398	0.461	0.379	0.657
Numerai28.6	0.588	0.641	0.998	0.585	0.625
PC2	0.115	0.08	0.116	0.115	0.121
Phonemes	0.132	0.118	0.365	0.129	0.153
Satellite	0.371	0.4	0.386	0.378	0.392
Sick	0.432	0.529	0.451	0.429	0.429
Telco Customer Churn	0.345	0.402	1.118	0.35	0.349
Waveform-5000	0.215	0.224	0.273	0.206	0.218

6. Concluding Remarks

We proposed a method to estimate the similarity of the approximate explanations generated by FastSHAP to the ground truth explanations without a need to compute the exact Shapley values at the test time and provide validity guarantees using the inductive conformal prediction framework. The method estimates the similarity to the ground truth and provides a validity guarantee for the entire explanation rather than per individual feature score as proposed in a previous approach that dismisses the relative importance of the remaining feature scores in one instance explanation. We also proposed a set of difficulty estimation functions to generate adaptive intervals, which are devised for explanations. The proposed difficulty estimation functions are evaluated based on their efficiency, i.e., the tightness of the generated intervals that entail the actual similarity of the approximate explanation to the ground truth. The results show that using the probability predicted by the explanation scores provided the most efficient intervals compared to the remainder of the proposed functions; however, it is not statistically different from the KNN difficulty estimate or the baseline without difficulty estimation.

A possible direction for future work is to compare different explanation similarity functions. Additionally, more difficulty estimation functions can be investigated in order to produce more efficient intervals. Another direction could be to study how the generated intervals can support decision-making. Finally, exploring uncertainty quantification methods for different explanation types, e.g., explanations for text and image classification models, is a possible direction.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. UJ is partly funded by the Swedish Knowledge Foundation through the AFAIR project with grant number 20200223. HB was partly funded by Vinnova (RAPIDS, grant no. 2021-02522).

References

Amr Alkhatib, Henrik Boström, and Ulf Johansson. Assessing explanation quality by venn prediction. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 42–54. PMLR, 24–26 Aug 2022.

Amr Alkhatib, Henrik Boström, Sofiane Ennadir, and Ulf Johansson. Approximating score-based explanation techniques using conformal regression. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 450–469. PMLR, 13–15 Sep 2023. URL https://proceedings.mlr.press/v204/alkhatib23a.html.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(4):1059–1086, September 2020. ISSN 1369-7412.

Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. Interpretable random forests via rule extraction. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 937–945. PMLR, 13–15 Apr 2021.

Henrik Boström. crepes: a python package for generating conformal regressors and predictive systems. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction and Applications*, volume 179 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Henrik Boström, Ram B. Gurung, Tony Lindgren, and Ulf Johansson. Explaining random forest predictions with association rules. *Archives of Data Science, Series A (Online First)*, 5(1):A05, 20 S. online, 2018. ISSN 2363-9881.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 883–892. PMLR, 10–15 Jul 2018.

Joseph Cohen, Eunshin Byon, and Xun Huan. To trust or not: Towards efficient uncertainty quantification for stochastic shapley explanations. In *PHM Society Asia-Pacific Conference*, volume 4, 2023.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3457–3465. PMLR, 13–15 Apr 2021.

Julien Delaunay, Luis Galárraga, and Christine Largouët. Improving Anchor-based Explanations. In *CIKM 2020 - 29th ACM International Conference on Information and Knowledge Management*, pages 3269–3272, Galway / Virtual, Ireland, October 2020. ACM. doi: 10.1145/3340531.3417461. URL https://hal.inria.fr/hal-03133223.

Duarte Folgado, Marília Barandas, Lorenzo Famiglini, Ricardo Santos, Federico Cabitza, and Hugo Gamboa. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. *Information Fusion*, 100:101955, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101955.

Jerome Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 10 2001.

Milton Friedman. A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 34 (205):109–109, 1939.

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1459–1467. PMLR, 13–15 Apr 2021.

Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Mach. Learn.*, 97(1–2):155–176, oct 2014.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *CoRR*, abs/1707.01154, 2017.

Octavio Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 10 2019. doi: 10.1109/ACCESS.2019.2949286.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

Christoph Molnar. *Interpretable Machine Learning.* 2022.

Peter Björn Nemenyi. *Distribution-free multiple comparisons.* PhD thesis, Princeton University, 1963.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, page 345–356, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540440364.

Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. *J. Artif. Int. Res.*, 40(1):815–840, jan 2011. ISSN 1076-9757.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, page 722–726, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606130.

Jonas Schulz, Raul Santos-Rodriguez, and Rafael Poyiadzi. Uncertainty quantification of surrogate explanations: an ordinal consensus approach. *Proceedings of the Northern Lights Deep Learning Workshop*, 3, 03 2022. doi: 10.7557/18.6294.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2021. doi: 10.1109/TVCG.2020.3030418.

Chu-I Yang and Yi-Pei Li. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15(1):13, Feb 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00682-3. URL https://doi.org/10.1186/s13321-023-00682-3.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. *GNNExplainer: Generating Explanations for Graph Neural Networks.* Curran Associates Inc., Red Hook, NY, USA, 2019.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019.

Xiaoge Zhang, Felix T.S. Chan, and Sankaran Mahadevan. Explainable machine learning in image classification models: An uncertainty quantification perspective. *Knowledge-Based Systems*, 243:108418, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.108418. URL https://www.sciencedirect.com/science/article/pii/S095070512200168X.