Assessing Explanation Quality by Venn Prediction

Amr Alkhatib AMAK2@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: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

Abstract

Rules output by explainable machine learning techniques naturally come with a degree of uncertainty, as the complex functionality of the underlying black-box model often can be difficult to approximate by a single, interpretable rule. However, the uncertainty of these approximations is not properly quantified by current explanatory techniques. The use of Venn prediction is here proposed and investigated as a means to quantify the uncertainty of the explanations and thereby also allow for competing explanation techniques to be evaluated with respect to their relative uncertainty. A number of metrics of rule explanation quality based on uncertainty are proposed and discussed, including metrics that capture the tendency of the explanations to predict the correct outcome of a black-box model on new instances, how informative (tight) the produced intervals are, and how certain a rule is when predicting one class. An empirical investigation is presented, in which explanations produced by the state-of-the-art technique Anchors are compared to explanatory rules obtained from association rule mining. The results suggest that the association rule mining approach may provide explanations with less uncertainty towards the correct label, as predicted by the black-box model, compared to Anchors. The results also show that the explanatory rules obtained through association rule mining result in tighter intervals and are closer to either one or zero compared to Anchors, i.e., they are more certain towards a specific class label.

Keywords: Venn prediction · Explainable machine learning · Rule mining

1. Introduction

Many machine learning algorithms achieve state-of-the-art performance in solving real-world problems in finance, medicine, biology, etc., but they usually produce black-box (non-interpretable) models (Linardatos et al., 2021). Understanding the logic used to produce the predictions is often a necessity for the user to place trust in such models. Using white box (interpretable) models is one possible solution, but they often come with a substantial loss of predictive performance (Loyola-González, 2019). Explainable machine learning is a practical solution to explain the predictions of a black-box model without losing the predictive performance.

Explanation techniques are classified into model-agnostic or model-specific. The first allows for explaining any black-box model, e.g., as in (Ribeiro et al., 2016a), while the latter exploits properties of the underlying model to produce the explanations, e.g., as in (Boström et al., 2018). The explanation techniques can also be classified according to the scope of the produced explanations into local and global techniques (Molnar, 2022). A local

explanation technique is instance-based, i.e., it explains a single prediction of a black-box model. On the other hand, global explanation techniques explain how a model behaves in general (Molnar, 2022).

Explanation techniques can produce explanations in various forms. Plots are one possible form that is conceivably intuitive and easy to understand, e.g., the Partial Dependence Plot (PDP) (Friedman, 2001) and the Accumulated Local Effects (ALE) Plot (Apley and Zhu, 2020). Other explanation techniques produce explanations in the form of (additive) feature importance scores, e.g., LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016b) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). Another popular approach to explaining predictions is to use rules; this format is claimed to be relatively easy to understand and is often preferred over alternative types of explanations. Anchors is a prominent example of such a technique (Ribeiro et al., 2018). However, these techniques have been observed to sometimes generate overly specific explanatory rules with low fidelity (faithfulness) to the underlying black-box model (Delaunay et al., 2020).

The low fidelity problem of some explanatory rules can partly be explained by the underlying black-box model being complex and not allowing for being approximated by a single explanatory rule. Since the fidelity can vary substantially between different explanations (rules), there is a need to properly quantify their uncertainty. Venn prediction (Vovk et al., 2005) is a multi-probabilistic method with proven validity guarantees that can be used to produce well-calibrated probability estimates (Johansson et al., 2019), and which we here propose and investigate as a means to quantify the uncertainty.

The main contributions of this study are:

- A novel method for quantifying the uncertainty of rule-based explanations
- A set of metrics designed to measure the uncertainty of the explanatory rules
- An empirical investigation comparing the uncertainty of explanations as produced by Anchors and an association rule mining explanatory technique

The following section briefly discusses related work on rule-based explanation techniques and Venn prediction. In Section 3, we describe how to quantify the uncertainty of the rules, then, a set of metrics to measure the uncertainty of the explanatory rules is proposed. In Section 4, we present and discuss the results of an empirical investigation in which explanations obtained through Anchors are compared to explanations generated by association rule mining. Finally, in Section 5, we summarize the main findings and outline directions for future work.

2. Related Work

In this section, we start out by discussing some model-agnostic rule-based explanation techniques. We then continue with a description of Venn predictors.

2.1. Rule-Based Explanations

Explainable machine learning is a research area that has recently gained some attention, and many explanation techniques provide its explanations in the form of rules. Anchors (Ribeiro et al., 2018) is an example of such a technique, which generates explanations such that non-included features should have no effect on the model's outcome. It is claimed that the rules are easier to understand and often preferred to alternative types of explanation (Ribeiro et al., 2018). However, it has been shown that Anchors may produce specific rules with low fidelity (Delaunay et al., 2020). Local Rule-based Explanations (LORE) (Guidotti et al., 2019) is another model-agnostic method, which trains a white-box model using synthetic data in the neighborhood of the instance of interest, while the neighborhood examples are generated using a genetic algorithm. The trained white-box model acts as a faithful explainer locally on a specific data instance but not as a general explainer on other data instances. The logic of the local white-box model is used to explain in the form of a decision rule. LORE also provides a set of counterfactual rules, pointing out some changes in the features that can result in another outcome. Since Anchors and LORE provide local explanations, GLocalX (Setzu et al., 2021) was proposed as a method to generate global explanatory rules. GLocalX measures the similarity of different rules and merges similar local rules into a more general one. The generated general rules can emulate the black-box model, and their faithfulness can be tested on a dataset. In order to provide guarantees on the fidelity of the extracted explanatory rules, Johansson et al. (2022) utilized the conformal prediction framework. However, this novel method is limited to regression models.

2.2. Venn Predictors

A probabilistic predictor aims to predict the probability distribution of the label on a test object, given the training data. A valid predictor has probability distributions that perform well against statistical tests based on subsequent observation of the labels. Such validity cannot be achieved in a general sense (Gammerman et al., 1998). However, Venn predictors (Vovk et al., 2003) can overcome this impossibility by two means. First, multiple probabilities for each label are produced, containing the valid one, and second, the statistical validity tests are restricted to calibration. Precisely, the observed frequencies must match the produced probabilities; for instance, if several predictions were made with a probability estimate 0.9, these predictions should be correct in about 90% of the cases.

Venn predictors were introduced in a transductive setting requiring one model to be trained for each class label for each instance, which incurs a high computational cost. However, Lambrou et al. (2015) developed the Inductive Venn Predictors (IVPs) to overcome the computational inefficiency problem. Since the inductive Venn prediction is a key concept used in our proposed method, next, we describe the IVPs.

Assume we have training data in the form of $\{z_1, \ldots, z_l\}$. Inductive Venn predictors require the data to be split into two subsets, a *proper training set* used to train the underlying model and *calibration set* to estimate label probabilities for each test example. Let the proper training set be $\{z_1, \ldots, z_q\}$, and similarly, the calibration set be $\{z_{q+1}, \ldots, z_l\}$. Each instance in the two sets is composed of two parts $z_i = (x_i, y_i)$, an *object* x_i and a *label* y_i.

Given a new test object x_{l+1}, Venn prediction estimates the probability that $y_{l+1} = Y_j$, for each label Y_j in the set of possible labels $Y_j \in \{Y_1, \ldots, Y_c\}$. Inductive Venn prediction is based on the fundamental idea of dividing the calibration data into several *categories* and using the frequency of label Y_j in each category to estimate label probabilities for new instances belong to the same category. A *Venn taxonomy* is used to specify the categories, and each taxonomy results in a different Venn predictor. Typically, the taxonomy is based on the underlying model, and a new object x_i is assigned to a category based on the model's output. One simple Venn taxonomy is to divide the instances into one category for each predicted label.

When estimating label probabilities for a test object x_{l+1}, first, the category of the object is determined by the model similar to the calibration objects. Afterwards, the label frequencies in the calibration data in the same category are used to calculate the label probabilities. All possible labels $Y_j \in \{Y_1, \ldots, Y_c\}$ are used since the true label y_{l+1} is unknown, resulting in a set of label probability distributions. However, a more compact representation can be used instead of the probability distributions. The compact representation uses the lower $L(Y_j)$ and upper $U(Y_j)$ probability estimates for each label Y_j. Assume that the Venn taxonomy assigns the category k to a test object x_{l+1} and Z_k is the of calibration instances that belong to category k. Consequently, the lower $L(Y_j)$ and upper $U(Y_j)$ probability estimates are defined by:

$$L(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}|}{|Z_k| + 1} \tag{1}$$

and:

$$U(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}| + 1}{|Z_k| + 1} \tag{2}$$

A prediction $\hat{y}_{l+1}$ for x_{l+1} can be made using the lower probability estimates as follows:

$$\hat{y}_{l+1} = \max_{Y_j \in \{Y_1, \ldots, Y_c\}} L(Y_j) \tag{3}$$

The output of a Venn predictor is the prediction $\hat{y}_{l+1}$ all together with the following probability interval:

$$[L(\hat{y}_{l+1}), U(\hat{y}_{l+1})] \tag{4}$$

Vovk et al. (2005) have shown that multi-probability predictions of Venn predictors are valid, no matter what taxonomy is used. However, it does not mean that the choice of taxonomy is unimportant, as it has an effect on the prediction accuracy and the size of the probability intervals of the Venn predictor. It is also essential to note that the smaller probability intervals are more informative, and the probability estimates should preferably be as close to one or zero as possible, which indicates high certainty in prediction.

3. Method

This section first describes a method to quantify the explanatory rules' uncertainty with respect to the underlying model predictions, allowing for selecting rules with better fidelity estimates. Afterwards, we propose a set of metrics to measure the quality of the explanatory rules based on the uncertainty.

3.1. Explanatory Rule Uncertainty Quantification

The proposed method for uncertainty quantification is agnostic to the explanation technique, as long as it outputs (explanatory) rules. The uncertainty quantification is done by using the upper and lower probability estimates of each class label (Y_j), where the correct classes are the predicted ones by the underlying model. Consequently, we will refer to the upper and lower probability estimates by the upper and lower fidelity estimates or the upper and lower fidelity bounds.

The uncertainty of each explanatory rule is quantified separately (as an independent classifier). A calibration set (X_c) is devoted to uncertainty quantification, from which each rule determines its calibration subset (X_r). A rule can be used to select its calibration subset by checking all the conditions in the rule's antecedent. If all the conditions are met by one instance, it will be included in the rule's calibration subset (X_r). The selection is made regardless of the rule's consequent or the true label (as predicted by the black-box model). After selecting the instances of the calibration subset, the black-box model is employed to obtain the labels (Y_r) of the instances in the calibration subset (X_r). X_r and Y_r form together the labeled calibration subset (Z_r) needed for computing the upper and lower fidelity estimates.

Finally, Equation (5) and Equation (6) are employed, using Z_r, to compute the lower and upper fidelity bounds of each class label, respectively.

$$L(Y_j) = \frac{|\{(x_i, y_i) \in Z_r | y_i = Y_j\}|}{|Z_r| + 1} \tag{5}$$

$$U(Y_j) = \frac{|\{(x_i, y_i) \in Z_r | y_i = Y_j\}| + 1}{|Z_r| + 1} \tag{6}$$

3.2. Evaluation Metrics

We propose three metrics (and three extensions of them) to measure the uncertainty of the explanatory rules and present the results as average values over all the rules. One of the proposed metrics assumes access to a test dataset. In contrast, the other two metrics can be determined using the upper and lower fidelity estimate values as computed on the calibration dataset.

The average lower bound (ALB) is the metric to be computed using a test dataset. The ALB values are reported on test instances with applicable rules, i.e., rules with all conditions in their antecedents are true with respect to the data instance. If two or more rule-based explanatory techniques are compared, the test instance must be covered by at least one rule

from each technique. Otherwise, the test instance is omitted from the evaluation, ensuring that the techniques are compared on precisely the same data. The ALB is computed for each data instance by checking the rules that cover that instance. For all the applicable rules, the average value of the lower bounds of each class label is computed. Only the average value of the correct class label, as determined by the black-box model, is returned (Equation (7)). Accordingly, a high ALB value indicates that the rules are predicting the correct label in accordance with the black-box model and with high certainty.

$$ALB = \frac{1}{n}\sum_{i=1}^{n} L(Y_{ji}), \text{ where } n \text{ is the number of rules} \tag{7}$$

The other two metrics are the difference between the upper and lower bound of each class label (the interval size, Equation (8)) and the absolute difference between the lower bounds of the two classes ($\triangle$LB, Equation (9)) since we consider only binary classification problems in this study.

$$\textit{Interval Size} = U(Y_j) - L(Y_j) \tag{8}$$

$$\triangle LB = |L(Y_1) - L(Y_0)| \tag{9}$$

The interval size and $\triangle$LB are computed for each rule without an essential need for a test dataset but using the upper and lower bound values computed on the calibration set. According to Johansson et al. (2019), a tighter probability interval is more informative, and the interval bounds should preferably be as close to one or zero as possible. For this specific motivation, we compute the $\triangle$LB, in addition to the interval size, to monitor how close the intervals are to one/zero. Since we have two classes, a certain rule should have one class's lower bound close to 1 and the other class's lower bound close to zero. Consequently, the absolute difference between the two classes' lower bounds should preferably be as close to one as possible. The $\triangle$LB reflects how confident a rule is toward predicting one class label. The uncertainty is high if a rule has a small difference between the lower bounds of the two classes and vice versa.

Since the interval size and $\triangle$LB are computed only on the calibration set and the value of both metrics is the average overall rules, which assumes that all explanatory rules are equally important. Therefore, we also provide weighted average values, where rules with higher coverage are rewarded with higher weights while computing the average value. In order to compute the weighted average, each interval size or $\triangle$LB value is multiplied by the number of instances that the rule covers, then all the multiplied values are summed and divided by the total count of multiplied values.

Finally, we provide ALB-P an extension to the ALB metric, which penalizes a method for explanations with low coverage. Thus, instead of computing the ALB scores on the data instances with coverage from all the compared methods, each method will be evaluated on all the data instances, and a zero ALB value is added for each data instance with no applicable rules.

4. Empirical Evaluation

In this section, the uncertainty of the rules obtained from two explanatory techniques, namely Anchors and the association rules, is compared. The quality of the generated explanatory rules is assessed using the proposed metrics in Subsection 3.2, where better explanatory rules are the rules with lower uncertainty. We conduct two sets of experiments. In the first set of experiments, Subsection 4.2, we evaluate the explanatory rules assuming that all the rules are equally important. While in the second set of experiments, Subsection 4.3, we assign higher weights to the rules with higher coverage.

4.1. Experimental Setup

The experiments are conducted using 12 public datasets[1]. Each dataset is split into training, development, calibration, and test sets. The ratios of the splits vary based on the size of the dataset, but the majority followed 40% training, 20% development, 20% calibration, and 20% test. The black-box model is trained on the first, the explanatory rules are generated using the black-box predictions on the second set, the uncertainty of the explanatory rules are quantified on the third, and finally, the quality of the produced rules is evaluated on the fourth set. All datasets are for binary classification problems except for Compas[2], which contains three classes (Low, Medium, and High) that have been reduced into two by combining Low and Medium in one class. The black-box models are generated by XGBoost (Chen and Guestrin, 2016). Some of the models' hyperparameters (e.g., learning rate, number of estimators, and the regularization parameter lambda) are tuned through grid search using 5-fold cross-validation on the training set.

In the following experiments, explanations obtained through Anchors and association rule mining are compared. Anchors is used with the default hyperparameters, and the confidence threshold has been set to 0.9. In subsection 4.1.1, we describe how the explanatory rules are obtained using Anchors and the association rule mining.

4.1.1. Obtaining Explanatory Rules

Anchors is used to generate explanations for the black-box model predictions on the development set. Then, all the conditions in the obtained rules are added to one set of conditions (*combined conditions*). Consequently, each data object is represented by an itemset of conditions (*explanation itemset*), where the contents of the *explanation itemset* are the true conditions in the *combined conditions* for that object all together with the predicted class label. The *explanation itemsets* are the input to the association rule mining algorithm.

An association rule mining algorithm is applied to the explanation itemsets using the specified confidence and support thresholds.

Finally, the set of obtained association rules is filtered by keeping only the rules for which some set of conditions in the antecedent and a single class label appear in the consequent.

In all experiments, the Apriori algorithm (Agrawal and Srikant, 1994) is employed for association rule mining. The association rule mining algorithm requires two hyperparameters; support and confidence. The former is primarily set to 4 but can also be raised if the

1. All the datasets were obtained from https://www.openml.org except Compas
2. https://github.com/propublica/compas-analysis

number of rules is overwhelming. The confidence threshold is selected based on a trade-off between the uncertainty (measured on the development set) and the number of rules; accordingly, high confidence may produce a very small number of rules with high certainty and vice versa.

4.2. Rule-Based Evaluation

In this experiment, the rules are evaluated using the metrics previously described in Subsection 3.2 with an assumption that all the rules are equally important. Consequently, the interval size and $\triangle$LB values are computed by taking the average value over all the rules. The results from comparing Anchors' rules to the association rules are summarized in Table 1. The association rules generally have higher certainty levels than the rules of Anchors.

To test the null hypothesis that there is no difference in the uncertainty, as measured by ALB, between Anchors and the association rules, and since we only compare two methods in this experiment, the Wilcoxon signed-rank test (Wilcoxon, 1945) is employed. It turns out that the null hypothesis may be rejected at the 0.05 level. Similar statistical significance tests have been carried out for the interval size and $\triangle$LB and showed that these differences are statistically significant as well.

We also illustrate the differences between explanations produced by Anchors and association rule mining, by presenting samples of both methods' rules in Table 2 and Table 3.

In summary, it can be concluded that with association rule mining, we can provide explanatory rules with reduced levels of uncertainty compared to the rules obtained directly through Anchors.

4.3. Instance-Based Evaluation

In this set of experiments, the same metrics that are used in Subsection 4.2 are employed again, but the scores of the rules with higher coverage are rewarded with higher weights while computing the average value. Only the values of the $\triangle$LB and the interval size are considered in the weighted averaging; consequently, the ALB values remain unaffected in this experiment. We replace ALB with ALB-P to reward higher coverage, as shown in Table 4. The results in Table 4 show no significant difference, using the Wilcoxon signed-rank test, between Anchors' rules and the association rules in terms of the sizes of the prediction intervals. However, the association rules result in significantly higher $\triangle$LB values than Anchors. On the other hand, Anchors provides significantly higher ALB-P values than association rules, reflecting the high coverage provided by Anchors' explanations.

Table 1: Average values of ALB, interval size, and △LB, in addition to the number of rules and coverage for Anchors and the association rules.

Dataset	**Association Rules**					**Anchors**				
	ALB	Interval Size	△LB	#Rules*	Cov.**	ALB	Interval Size	△LB	#Rules	Cov.
ada	0.926	0.01	0.94	209	0.60	0.917	0.09	0.81	311	1.00
Bank Marketing	0.953	0.01	0.97	149	0.84	0.918	0.38	0.54	118	1.00
BNG breast-w	0.988	0.003	0.99	18	0.86	0.981	0.002	0.98	121	1.00
Compas	0.931	0.025	0.93	135	0.80	0.906	0.282	0.54	313	1.00
Churn	0.873	0.03	0.88	96	0.90	0.863	0.189	0.71	398	1.00
Internet Advertisements	0.919	0.008	0.92	179	0.84	0.879	0.019	0.77	294	1.00
Jungle Chess 2pcs	0.99	0.01	0.99	102	1.00	0.99	0.03	0.97	22	1.00
mc1	0.999	0.0007	0.999	349	0.99	0.998	0.058	0.884	35	1.00
Mushroom	0.995	0.01	0.99	273	0.96	0.981	0.02	0.97	123	1.00
Phishing Websites	0.971	0.007	0.97	166	0.62	0.942	0.016	0.88	187	1.00
Spambase	0.937	0.01	0.95	210	0.91	0.928	0.032	0.94	336	0.997
Telco Customer Churn	0.92	0.06	0.88	217	0.80	0.907	0.08	0.74	196	1.00
Average rank***	1.08	1.08	1	–	–	1.92	1.92	2	–	–

* The number of rules.
** The coverage value here is the percentage of instances in the dataset that are covered by at least one rule.
*** The average rank shows which method is better on average, with 1 being the best and 2 the worst result.

Table 2: The top 4 explanatory rules of each class output by Anchors for the BNG breast dataset with the lower (L(Y)) and upper (U(Y)) fidelity estimates of the correct class label.

Conditions	Label	L(Y)	U(Y)
Cell Size Uniformity ≤ 1 & Cell Shape Uniformity ≤ 1	Benign	0.996894	0.997930
Cell Size Uniformity ≤ 1 & Clump Thickness ≤ 2.22	Benign	0.997630	1
Cell Shape Uniformity ≤ 1 & Clump Thickness ≤ 2.22	Benign	0.997354	1
Cell Shape Uniformity ≤ 1 & Clump Thickness ≤ 4	Benign	0.997465	0.998733
Clump Thickness > 5.66 & Cell Size Uniformity > 4.49	Malignant	0.996805	1
Clump Thickness > 5.66 & Normal Nucleoli > 3.55	Malignant	0.989209	0.992806
Cell Size Uniformity > 4.49 & Bare Nuclei > 6.61	Malignant	0.997135	1
Cell Size Uniformity > 4.49 & Single Epi Cell Size > 2	Malignant	0.960986	0.963039

Table 3: The top 4 explanatory rules of each class output by association rules for the BNG breast dataset with the lower (L(Y)) and upper (U(Y)) fidelity estimates of the correct class label.

Conditions	Label	L(Y)	U(Y)
Bare Nuclei $\leq$ 6.61 & Cell Shape Uniformity $\leq$ 1	Benign	0.998008	0.999004
Bland Chromatin $\leq$ 4.66 & Cell Shape Uniformity $\leq$ 1	Benign	0.996051	0.997038
Bland Chromatin $\leq$ 3 & Cell Shape Uniformity $\leq$ 1	Benign	0.997992	0.998996
Cell Shape Uniformity $\leq$ 1 & Bare Nuclei $\leq$ 1	Benign	0.997785	0.998893
Clump Thickness $>$ 5.66 & Cell Size Uniformity $>$ 4.49	Malignant	0.996805	1
Clump Thickness $>$ 5.66 & Cell Shape Uniformity $>$ 5.18	Malignant	0.99361	0.996805
Cell Shape Uniformity $>$ 5.18 & Mitoses $>$ 1	Malignant	0.995025	1
Cell Size Uniformity $>$ 4.49 & Mitoses $>$ 1	Malignant	0.995098	1

Table 4: Average values of ALB-P, weighted interval size, and weighted $\triangle$LB, in addition to the number of rules and coverage for Anchors and the association rules.

Dataset	**Association Rules**					**Anchors**				
	ALB-P	Interval Size	$\triangle$LB	#Rules	Cov.	ALB-P	Interval Size	$\triangle$LB	#Rules	Cov.
ada	0.56	0.008	0.94	209	0.60	0.82	0.011	0.90	311	1.00
Bank Marketing	0.798	0.006	0.99	149	0.84	0.896	0.002	0.89	118	1.00
BNG breast-w	0.85	0.002	0.995	18	0.86	0.97	0.002	0.986	121	1.00
Compas	0.74	0.005	0.96	135	0.80	0.85	0.003	0.89	313	1.00
Churn	0.78	0.01	0.91	96	0.90	0.85	0.006	0.89	398	1.00
Internet Advertisements	0.77	0.007	0.92	179	0.84	0.82	0.005	0.82	294	1.00
Jungle Chess 2pcs	0.99	0.012	0.988	102	1.00	0.99	0.013	0.987	22	1.00
mc1	0.989	0.0007	0.999	349	0.99	0.993	0.0014	0.997	35	1.00
Mushroom	0.95	0.0059	0.994	273	0.96	0.97	0.0067	0.976	123	1.00
Phishing Websites	0.60	0.0033	0.98	166	0.62	0.89	0.0047	0.91	187	1.00
Spambase	0.85	0.009	0.957	210	0.91	0.91	0.012	0.953	336	0.997
Telco Customer Churn	0.74	0.005	0.977	217	0.80	0.87	0.0095	0.899	196	1.00
Average rank	1.92	1.375	1.125	–	–	1.08	1.625	1.875	–	–

5. Concluding Remarks

We have proposed a method to quantify the uncertainty of the explanations expressed in the form of rules. Moreover, a set of metrics of rule explanation quality based on uncertainty is proposed. The metrics measure the tendency of the explanations to predict the correct label on new instances, how informative the produced intervals are, and the certainty of a rule to predict one class. We have presented results from an empirical evaluation, comparing explanatory rules obtained by the state-of-the-art technique Anchors to rules obtained from association rule mining. Two sets of experiments have been carried out. In the first set, the performance is averaged across the rules, where all rules are given the same weight, while in the second set of experiments, the performance is instead averaged over instances, where the rules are weighted according to their coverage. The explanations produced by association rule mining were observed to significantly outperform Anchors' explanations when averaging across rules. When instead averaging across instances, no significant difference between Anchors and the association rules was observed regarding the interval size; however, the association rules maintained significantly higher $\triangle$LB values, while in contrast, Anchors achieved significantly higher ALB-P values than association rules. When averaging over instances, the results indicate that Anchors may be more reliable, while when averaging over rules, Anchors may produce more uncertain rules compared to those produced by association rule mining.

One direction for future work is to investigate uncertainty quantification for other explanation types, in addition to rules, in particular additive feature importance scores, which are produced by several popular explanation techniques. The challenge there would be to find ways of verifying, or evaluating, the explanations.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8. URL http://dl.acm.org/citation.cfm?id=645920.672836.

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. doi: 10.1111/rssb.12377.

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. doi: 10.5445/KSP/1000087327/05.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016. doi: 10.1145/2939672.2939785.

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.

Jerome Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 10 2001. doi: 10.2307/2699986.

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.

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019. doi: 10.1109/MIS.2019.2957223.

Ulf Johansson, Tuve Löfström, Henrik Linusson, and Henrik Boström. Efficient venn predictors using random forests. *Machine Learning*, 108, 03 2019. doi: 10.1007/s10994-018-5753-x.

Ulf Johansson, Cecilia Sönströd, Tuwe Löfström, and Henrik Boström. Rule extraction with guarantees from regression models. *Pattern Recognition*, 126:108554, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.108554. URL https://www.sciencedirect.com/science/article/pii/S0031320322000358.

Antonis Lambrou, Ilia Nouretdinov, and Harris Papadopoulos. Inductive venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74(1):181–201, Jun 2015. ISSN 1573-7470. doi: 10.1007/s10472-014-9420-z. URL https://doi.org/10.1007/s10472-014-9420-z.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

Octavio Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 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. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Christoph Molnar. *Interpretable Machine Learning*. 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016a.

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, 2016b.

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

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 01 2021. doi: 10.1016/j.artint.2021.103457.

Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2003/file/10c66082c124f8afe3df4886f5e516e0-Paper.pdf.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Frank Wilcoxon. Individual comparisons by ranking methods. biometrics bulletin 1, 6 (1945), 80–83. *URL http://www. jstor. org/stable/3001968*, 1945.