Explaining Decisions of Black-box Models using BARBE

Mohammad Motallebi, Md Tanvir Alam Anik, and
Osmar R. Zaïane[0000−−0002−0060−5988]

[1] University of Alberta, Edmonton Alberta, Canada
[2] Alberta Machine Intelligence Institute
{zaiane@cs.ualberta.ca}

Abstract. Machine learning models are ubiquitous today in most application domains and are often taken for granted. While integrated into many systems, oftentimes even unnoticed by the user, these powerful models frequently remain as black-boxes. They are black-boxes because while they are powerful predictive models, it is commonly the case that one cannot understand the decision-making process behind their predictions. Even if we understand the inner workings of a learning algorithm building a predictive model, the mechanism during inference is more often than not obscure. How can we trust that a certain prediction from a model is correct? How can we trust that the model is making reasonable predictions in general? Debugging a predictive model is unworkable in the absence of explanations.
We propose herein a new framework, called BARBE, a model-independent explainer, that learns a surrogate rule-based model on data labeled by the black-box. BARBE makes use of an interpretable associative classifier to create a rule-based model that provides various explanations, including salient features, associations between features, and rule-based representations. Our experimental analysis illustrates the effectiveness of BARBE in generating rule-based explanations for both numerical and text data, when compared to state-of-the-art explainers. Our study demonstrates the faithfulness of BARBE to black-box models. The text-based explanations generated by BARBE are more meaningful to show the fidelity and trustworthiness of the explanation.

Keywords: Machine Learning, Explainable AI, Associative Classification, Model Independent Explanation

1 Introduction

Explainable Artificial Intelligence (XAI) has attracted the attention of many researchers in recent years. This surge in interest is prompted by the need to obtain explainability in different AI domains. Providing an explanation is indeed a requirement in many jurisdictions when an AI system is used to make critical decisions for humans [1], [23]. The objective of augmenting systems with explainability is to provide supplementary information on top of the main output

created by them (such as the class label in a classifier). This new information allows or empowers the user to know why the AI system provided the aforementioned output. One example is a model trained to detect colon cancer in people based on their medical records. Researchers later noticed the data included the name of medical clinics in which patients were admitted later. The presence of this feature was erroneously picked up by the trained model and had caused a significant fictitious boost in the performance of the system [26]. In all such cases, some sort of explainability providing transparency could have helped them avoid the consequences, consequences that are often far-reaching like contributing to the distrust of machine learning.

To tackle this issue, various XAI methods have been developed in recent years that attempt to provide explainability in one way or another. Most effort has been on attempting to justify or elucidate neural networks since the spotlight is currently on deep learning. However, there is also an effort for more generic approaches. For the classification task, in particular, methods have been developed in which the explanation framework is either independent of the classifier or is integrated into it. The former approach is called model-independent explanation (or sometimes called model-agnostic), while the latter is called model-dependent. One significant advantage of model-independent approach frameworks is that these frameworks can be added to different classifiers. This addition allows machine learning enthusiasts to inject some sort of explainability into any existing classifier and leverage them readily.

One problem with some of the current model-independent approaches is that they do not provide highly precise explanations. In other words, regardless of what the "real" explanation is, the user can ask for a fixed number of important features (e.g., give me top k features relevant to the decision), and the system then provides k important features accordingly. Additionally, another notable aspect is to take into account the correlations among input features. Some approaches, as we discuss in the next section, pay no attention to this aspect. In our view, this is a critical part in which the end-user should be able to depend on to better understand the underlying "reasoning" of the black-box model.

In this work, we introduce BARBE, for **B**lack-box **A**ssociation **R**ule-**B**ased **E**xplanation, a model-independent method that explains the decisions of any black-box classifier for tabular and text datasets with high precision. Moreover, the black-box classifier is not required to provide any probability score to take advantage of BARBE. Furthermore, BARBE presents explanations in three alternative forms: 1) the importance score for salient features, which many methods also benefit from; 2) significant associations between pertinent features; and 3) the construction of classification rules, which distinguishes BARBE from other methods. BARBE exploits association rules, a particular kind of rules that take into account the associations between features, helping users grasp different underlying potential causes of a decision.

The rest of this paper is organized as follows. We discuss a few of the main XAI approaches in the next section. Section 3 contains some preliminaries on associative classification. This section is needed as we take advantage of an as-

sociative classifier as the core of our work. We introduce BARBE, the main contribution of this paper, in Section 4. Later, we report the experiments we conducted in Section 5 and Section 6 to show correctness and fidelity to the black-box predictions. In this section, we show how BARBE performs and compare it against other methods. We conclude this work in Section 7 and provide some thoughts about future work.

2 Related Work

Attempts to explain classification decisions are not new. ExplainD is a tool introduced by Poulin et al. [19] that visualizes the decisions of well-known classifiers, which helps users understand their behavior during inference. Most researchers focus on explaining Deep Neural Networks (DNNs), particularly for image classification, by exposing the internals of the model using methods such as computing gradients and propagating them back to input to capture important pixels, which can be presented as the explanation [7], [25] (e.g., Grad CAM [24]).

LIME [21] is a popular model-agnostic method that uses perturbed samples to train a linear regression model for explaining black-box models. It relies on the input and output of the model to generate explanations, without knowledge of the internal structure of the black-box model, and its explanation is a ranked list of important features for the prediction of each data point.Anchor [22] is an approach for explaining black-box models that provides a set of salient features in the form of a single "if-then" rule to overcome the limitation of the linear model associated with LIME. A weakness of this approach is that it does not reveal the associations among features. Also, in contrast to LIME, it cannot provide any relative feature importance scores anymore.

Guidotti et al. introduce LORE [6] that takes advantage of rules for providing explanations. In their method, they create a neighbourhood around the instance using a Genetic Algorithm. Moreover, they enforce the data point selection algorithm to choose at most half of the data points from the class of the original data point. Note that while data points are created by the genetic algorithm, the class labels are obtained by querying the black-box model. With the labeled synthetic data points they train a decision tree. They take advantage of the decision tree to produce two types of rules; a single decision rule, and a set of counter-factual rules.

Pattern Aided Local Explanation (PALEX) is another method proposed to provide explanations for black-box models. In their method, Jia et al. [10] suggest a set of patterns as the explanation using FP-Growth algorithm [9]. Their method, however, requires defining a few hyper-parameters such as minimum support, and minimum growth ratio thresholds as well as the probability score provided by the black-box model. Alternatively, LACE [17] directly learns an associative classifier by exploiting the nearest data points in training data. Their method, however, requires the training data to be available, and this may not always be realistic. Additionally, the sparsity of the training data in that neighbourhood, can have a substantial impact on the performance of their system.

CoSP (Co-Selection Pick) is a recent framework proposed by Meddahi et al. [15], which aims to provide global explainability for black-box machine learning models. It does not explain a specific prediction but piggybacks on an existing explainer to co-select the most important test instances and features of the model as a whole. The framework selects individual explanations based on a similarity preserving approach, achieving a co-selection of instances and features. Unlike submodular optimization methods, CoSP considers the problem as a co-selection task and can be applied in both supervised and unsupervised scenarios with supposedly any local explainer. In their paper they used LIME.

3 Associative Classification

Rule-based classifiers, such as Ripper [4] or SigDirect [11], generate easily interpretable models by learning classification rules of the form "**If** *condition* **Then** *class*". Being known as transparent classifiers, they use attribute-value pairs as the antecedent and a class label as the consequent. During inference, applicable rules are selected based on whether their antecedent matches the instance's features, and a heuristic is used to assign the consequent as the prediction.

Associative classifiers learn their classification rules by applying association rule mining, a canonical task in data mining on the data after modeling the training data into transactions, each transaction being a set of attribute-value pairs and the class label. The rules are conjunctions of feature-values implying a class label: f_1, and f_2, and f_3, and f_4, and ..., $f_n \rightarrow class1$.

Associative classifiers have, for the most part, after the rule generation, a rule pruning phase to weed out redundant and noisy rules, and this is where the various approaches differ, in addition to the heuristics used to select rules at inference time. The most recent associative classifier approach that outperforms all preceding algorithms is SigDirect [11], which we take advantage of in our framework BARBE. The authors of SigDirect showed that not only did their algorithm outperform the other associative classifier contenders in accuracy on various datasets, it also generates a classification model with significantly less rules. Having fewer and more accurate rules is particularly pertinent for providing explanations in BARBE. Another advantage is the lack of cumbersome parameters. With other associative classifiers like CBA [13], CMAR [12] or ARC [3] hyper-parameter tuning is required. They heavily rely on support and confidence thresholds which are notoriously difficult to assess. SigDirect uses instead statistical significance to appraise rules.

SigDirect uses an Apriori-like strategy to first generate the rules and then leverage a new instance-based approach for the pruning step to only keep rules with the highest quality [11]. Similar to Apriori [2], it expands one level at a time but uses the Kingfisher algorithm [8] to find globally optimum, non-redundant dependencies with a scalable branch and bound approach with supplementary pruning by means of Fisher exact test and a P-value for statistical significance. Therefore we use SigDirect, a strong interpretable rule-based classifier that gen-

erates a minimal number of statistically significant classification rules, as the core of our model-independent explanation framework.

4 BARBE: Black-box Association Rule-Based Explanation

4.1 Shortcoming of other methods

Take what LIME generates for the text "A movie where tensions build and conflicts arise" as shown in Figure 6A. The number below each feature is simply the "importance scores" used for ranking. For example, 0.10 for *movie* and 0.06 for $conflicts$ highlight that *movie* has slightly higher importance than $conflicts$ in making the sentence negative. This leads us to the conclusion that only the order among features matters to the users and not the numbers generated in the explanations. Sine LIME uses a weighted loss function for its linear model that also benefits from regularisation, it is likely that the instances which are not in the very close proximity of the original instance would be misclassified by this linear model, thus providing wrong explanations to the user. Ribeiro et al. [22], the same authors of LIME, also point out the fact that features are taken independently (see example in Figure 6B). They introduce Anchor to overcome this issue. In their new method, an explanation is a set of features that whenever they co-occur, the class label is determined with a 95% confidence. This Anchor essentially resembles a rule (with a high confidence threshold of 95%).

The authors of LORE [6] benefit from the idea of using a set of counterfactual rules as the explanation in their method as well. Despite the fact that these methods, to some degree, overcome the problem mentioned above, one issue remains: is there always only one set of correlative features (and hence one reason) behind the final outcome of the model? What if there were multiple sets of correlative features that independently derive the final conclusion of the system [16]. Therefore an explanation should not solely focus on independent features or one unique set of associated features but on possibly a set of causes. Hence the interest in an associative classifier that can provide a set of rules as an explanation.

To overcome the above shortcomings, we introduce Black-box Association Rule-Based Explanations or BARBE. Our method, unlike LIME, provides a set of rules as the explanation, where not only do rules provide users with important features (what LIME does), but also takes care of the associations among them (what LORE and Anchor do). In addition, since we provide multiple rules as an explanation, we can hint at multiple causes that have led to that decision by the system, something that the aforementioned methods are unable to provide. Note that using a decision tree (in systems like LORE [6]) the path in the tree leading to the predicted label results in a single applicable rule which constitutes only one unique cause.

4.2 Explanations by BARBE

BARBE generates a descriptive model learned on data labeled by the black-box and provides as the explanation a subset of rules from the model that apply to the instance for which the explanation is expected. From this set of rules and their individual measure of confidence and significance, BARBE can provide an ordered set of important features as an alternative way of providing explanations. This allows the users to have the choice to look at these two types and get a better understanding of the underlying causes. Moreover, as mentioned earlier, each rule in addition to the items in its antecedent and the class label, comes with added information such as its confidence, support value, and p-value. Not only does this provide an alternative means for users to comprehend black-box models, but it also opens the door for researchers to conduct comparisons with other techniques such as LIME.

Figure 1 shows an example of what BARBE outputs for an instance of the Glass dataset [5]. In this example, BARBE produces three rules in which they not only provide important features to the users but also hint at the associations among the features. For compactness, the feature number in both the table and histogram is displayed instead of the full name. The first rule could be written as "Magnesium = [0.38, 2.13], Aluminum = [1.64, 1.76], Calcium = [7.80, 8.23] → Vehicle Window". The second rule could be expressed as "Magnesium = [0.38, 2.13], Potassium = [0.00, 0.61] → Vehicle Window" and the third rule could be written as "Potassium = [0.00, 0.61] → Vehicle Window". Here, the rules are inferring class label 3.

RULES	*Support*	*Confidence*	*ln pF*
3="1" AND 4="3" AND 7="3" → 3	0.0900	0.833	-17.095
3="1" AND 6="1" → 3	0.2230	0.791	-33.641
6="1" → 3	0.3730	0.696	-24.174

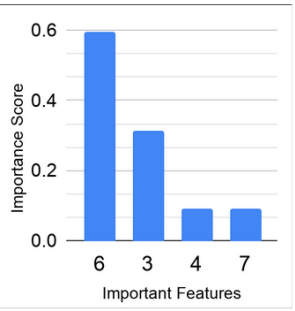


Fig. 1. The explanation provided by BARBE for an instance of the Glass dataset [5]. Here, feature tokens are shown for conciseness. The right side contains the important features ranked based on their importance. The left side contains important rules with their support, confidence, and the logarithm of the p-value reported by SigDirect.

4.3 How does BARBE work?

A high-level representation of BARBE's activity diagram is shown in Figure 2. BARBE creates a neighbourhood around the instance to explain with synthetic data points produced by perturbing the features of the instance. The synthetic data points are labeled by the black-box which produces a training set for the SigDirect classifier. The outcome of the training is a set of rules. Rules from the trained model relevant to the original instance are extracted. Lastly, BARBE derives important features from these rules. BARBE needs to have access to the set of possible values for each attribute. Moreover, SigDirect, the heart of BARBE, relies on associations between discrete features. Indeed numerical values of continuous attributes need to be discretized into intervals. Associative classification rules demand ordinal features. Therefore, if buckets are not predefined, and in order to define buckets for continuous data, BARBE needs to access a sample of data from which the instance to be explained was drawn.

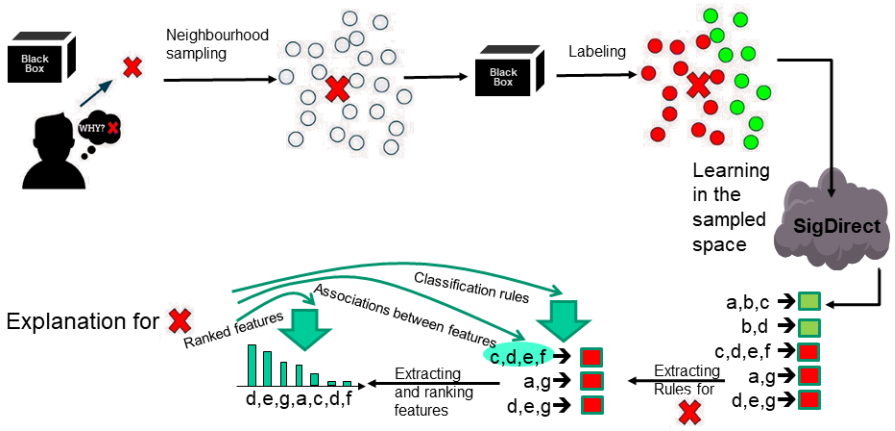


Fig. 2. Coarse representation of BARBE's framework.

BARBE receives an instance and a label to explain. If there are continuous features, and buckets for value intervals are not defined, a dataset of data points is used to discretize numerical attributes and define buckets. These buckets are then used to perturb the original instance and generate a neighbourhood of synthetic data points around the original instance. All these points are then labeled by the black-box after being converted to the input format of the black-box. For instance, before the labeling by the black-box, a reverse quantization may be required for continuous values that were mapped to the set of discrete finite buckets. This reverse transformation creates a normal distribution for each bucket of a feature and then randomly samples from the distribution. For text data, BARBE uses a simple strategy to generate a synthetic dataset around the original instance. BARBE takes advantage of random word removal from the

sentence. The algorithm takes the input sentence and a number which tells the algorithm to iterate the process for n times to generate n number of synthetic text data. It selects a set of random positions within the input sentence and deletes the words from those positions. The resulting sentence after deleting random words is returned as the synthetic sentence. This process is repeated n times to generate n new sentences which form the neighbourhood dataset around the original text.

The resulting set of classification rules is relevant to the original instance since the training data is made of instances from its vicinity. The set is further reduced by selecting the most relevant rules to the instance to explain. The most important features are thereafter selected from these selected rules. We have experimented with different metrics to rank the features and found that summing the supports of all applied rules in which a feature f appears provides the best accuracy when compared with the features used by the black-box.

5 Experiments

5.1 Experiments Setup

To evaluate our method, we compare the explanation produced by BARBE against the true explanation. We replace the black-box model with a fully-transparent model and conduct experiments on this "open box". The interpretable model we leverage in our experiments is a Decision Tree (DT)[3]. Not every DT is interpretable. Its depth should be limited to a reasonable level so humans can track different paths in this data structure [7]. We limit the depth of the DT to a specific level k at train time, with k set at 5 in our experiment.

5.2 Experiments' Metrics

We use *Precision*, *Recall*, F_β*-score*, and *Rank-Biased Overlap (RBO)* as our comparative metrics in the experiments. We make use of $F_{0.5}$*-score* in our experiments to put more importance on *Precision* than *Recall*. If the explanation includes only a few of the features which are mostly tagged correctly as important (i.e., a case where *Precision* is high but *Recall* is low), then the end-user can still trust the system as this case indicates the black-box model is focusing on some of the right features. Moreover, since BARBE is presenting a ranked list of features as explanation, we take advantage of Rank-Biased Overlap (RBO) [27] to evaluate the order of important features compared to the ground truth explanation[4].

In our experiments, we compare the explanation provided by BARBE with the ground truth explanation obtained from the DT for each instance in the dataset. We then calculated the metrics for each one. These metrics are later averaged over all instances used in the experiment and finally reported in Table 1

[3] We use scikit-learn [18] for implementing the DT.
[4] We used the implementation in `https://github.com/changyaochen/rbo`.

and Figure 3 in the next section. All results are averages for 100 experiments of explanation evaluations.

	$F_{0.5}$-*score*			*RBO*		
dataset	**LIME**	**Anchor**	**BARBE**	**LIME**	**Anchor**	**BARBE**
Glass	0.736	0.534	**0.796**	0.777	0.287	**0.796**
Wine	0.554	0.656	**0.680**	0.314	**0.484**	0.416
Hungarian	0.483	0.508	**0.570**	**0.776**	0.264	0.427
Poker	**0.666**	0.331	0.637	0.353	**0.713**	0.368
Breast	0.633	0.497	**0.715**	0.300	0.246	**0.332**
Image	0.566	0.570	**0.852**	0.483	0.321	**0.500**
Magic	0.596	**0.729**	0.712	0.684	**0.835**	0.515
Vowel	0.526	0.683	**0.840**	0.671	0.444	**0.675**
Hepatitis	0.432	**0.492**	0.340	0.106	**0.218**	0.055
WPBC	0.489	0.536	**0.642**	0.543	0.594	**0.606**
WDBC	0.468	0.131	**0.494**	0.228	0.056	**0.320**
Average	0.559	0.515	**0.662**	**0.476**	0.406	0.455
Nb. Wins	1/11	2/11	**8/11**	1/11	4/11	**6/11**

Table 1. $F_{0.5}$ and *RBO* for different methods and datasets with sample size at $5,000$.

5.3 Comparison with Other Explainers

We choose LIME and Anchor to compare with BARBE through different experiments with an interpretable decision tree disguised as a black-box. We exploit 11 different UCI datasets [5] (Table 1) to conduct these experiments where the results we report are averages over five runs with different random seeds. In Figure 3, we provide *Precision*, *Recall*, and $F_{0.5}$-*score* for all data points that the method has predicted the class label correctly when the generated sample size around the instance increases from 1,000 to 5,000. For lack of space, we report here the results of the average for all 11 datasets.

To estimate the faithfulness of LIME, we examine its prediction score. Because the regression model is trained on values originating from a probability space, we can expect the regression model to generate a number in the same domain. Additionally, for the original point, if the interpretable model is trained properly, the predicted value should be close to 1. If the score, however, is below $\frac{1}{n}$, where n is the number of classes, that explanation is not faithful, and thus we do not include the incorrect case. Our criterion generously considers the outcome of LIME's regression model faithful, since most instances would belong to the target class and thus their probability score should be close to 1, yet, we observe BARBE providing competing results throughout our experiments. Anchor, however, does not rely on any interpretable model, and therefore, we are not able to determine its fidelity with the ground truth. Consequently, we assume in our experiments that all its explanations are faithful, and as a result, we include all instances of Anchor in the evaluation.

We also report the $F_{0.5}$ and *RBO* for the three methods and for all the datasets in Table 1 where the sample size is $5,000$.

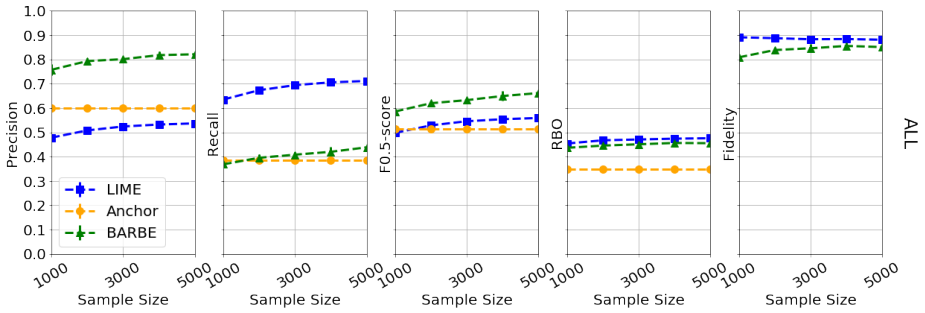


Fig. 3. Performance of LIME, Anchor, and BARBE.

In terms of $F_{0.5}$-*score*, BARBE displays the best results for all datasets but two, for which it was still a good contender. This good performance of BARBE is mainly due to the very high *Precision*. It was able to pinpoint correct features as per the ground truth. Moreover, LIME had typically higher *Recall* as LIME ranked all available features and therefore would rarely miss relevant ones. Interestingly, BARBE outperforms Anchor in not only for the $F_{0.5}$-*score*, but also *Precision* in most datasets, even though Anchor depends on a high precision rule to explain an instance. Moreover, the results demonstrate that BARBE has a better capacity to order the importance of features. In more than half the datasets, BARBE gives a better*RBO* score, followed by Anchor which beats LIME in terms of arranging the discovered salient features by importance.

5.4 Faithfulness to the Black-Box

To show the real importance of the features claimed important by BARBE on the decision of the black-box model, we make changes to the values of those features and request the black-box to do another classification. The more those changes are significant, the higher the chance that the black-box flips its decision. This is illustrated in Figure 4. We changed the value of the most important features by 1 standard deviation up to 2 standard deviations for 100 randomly selected data points from the Pen Digits dataset of UCI repository and classified by a neural network with 2 hidden layers as a black-box. The accuracy clearly and steadily drops as we increase the extent of the change which indicates that the features highlighted by BARBE are indeed the influencers for the black-box.

6 Experiments on BARBE for Text

In this section, we discuss the settings under which we conduct experiments to evaluate BARBE for text. Our goal is to develop a framework that can be

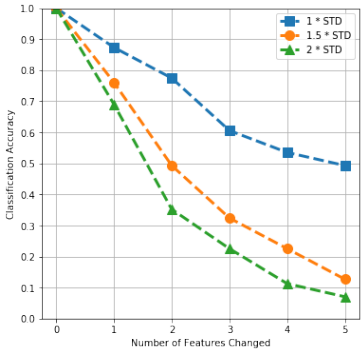


Fig. 4. Impact on Accuracy of changing values of important features.

employed not only on tabular datasets but also on text datasets. For text data, we demonstrate the efficacy of BARBE for binary classification tasks in this paper. We choose the IMDB movie review dataset [14] for the binary classification task since it is widely used in the literature for text classification.

6.1 Results

BARBE uses the data labeled by the black-box model to train a descriptive model that generates a set of rules as the explanation. Each rule has a support, confidence, and statistical significance value associated with it. The rules correspond to the features in the data which constitute the set of important features as the form of explanation. We have used support vector machine (SVM) as the underlying black-box and trained it with the IMDB movie review dataset. TF-IDF [20] has been used to convert the text into features in BARBE. When the black-box is ready, we use the neighbourhood generation process as discussed in Section 4.3 to create the synthetic dataset. This dataset once labeled by the black-box is used by BARBE to generate the rules in the form of explanation. Figure 5 shows the result generated by BARBE for a sentence with negative sentiment as labeled by the black-box. The sentence depicted in Figure 5A contains words that have been highlighted with a color gradient of red for negative words and green for positive words. Figure 5B depicts the rules generated by BARBE. Here, BARBE generates a total of 5 rules for this sentence to identify the most important features. The important features are highlighted from strong red to light red for negative and strong green to light green for positive ones.

It is evident from the figure that the words “conflicts” and “tensions” are the two negative words that are responsible for labeling the sentence as negative by the black-box model. It is noteworthy that BARBE generates a set of rules, not a single rule, and each rule contains a set of words. For the sentence of Figure 5A, BARBE generates “conflicts arise” as the association rule that essentially has semantic context in the sentence. “conflicts arise” has more semantic significance than the word “conflicts” only. The three other rules generated by BARBE are:

"tensions", "movie", and "build" as shown in Figure 5B. It is important to note that the word "build" has been detected as a feature having a positive label here. With 31.01% confidence, this rule has a positive label meaning this rule has little impact on making the sentence positive. Thus, BARBE not only identifies the rules that explain the negative sentiment of the sentence but also highlights the rules that may contribute slightly to the sentence being positive. Each rule has its support, confidence, and the logarithm of statistical significance values shown in the table in Figure 5B inside brackets. BARBE represents the black-box prediction probabilities as depicted in Figure 5C. In Figure 5D, the vertical bar chart showcases the most important features based on their frequency within the rules, along with the corresponding weighted confidence value. It is evident from this figure that the word "conflicts" and "tensions" are the most important feature whereas "movie", and "arise" are the least important ones.

We also demonstrate the explanation obtained by BARBE for a sentence having positive sentiment. Figure 5E presents the sentence. BARBE generates 5 rules to explain why the sentence has been labeled as positive by the black-box model as shown in Figure 5F. The words "wonderful" and "success" are enough to justify the sentence as positive. Figure 5G is the prediction probabilities of the black-box model and Figure 5H highlights the most important features in terms of their frequency weighted by the confidence values within the rules.

6.2 Comparison with Other Explainers

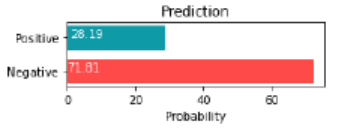
We compare the explanation generated by BARBE with LIME and Anchor for the text dataset. LIME for text differs from LIME for tabular data in terms of the neighborhood data generation methodology. Starting from the original instance, new instances are created by randomly removing words from the original instance. There is a major drawback here. While generating such neighborhood instances, LIME creates a large number of empty sentences as we explore the neighborhood generation algorithm of LIME. An empty sentence does not make any sense when it is used to be labeled by the black-box model. On the other hand, Anchor deploys a perturbation-based strategy to generate local explanations for predictions of black-box model.

Figure 6 illustrates the explanation generated by LIME and Anchor for the sentence "A movie where tensions build and conflicts arise". LIME highlights the feature "movie" as the most important word for making the sentence negative. It provides fewer weights to "tensions" and "conflicts". The word "movie" cannot justify the decision of the black-box alone with such a significant weight of 0.10. Besides, there is no association within the features generated by LIME. LIME provides less significance to "tensions" and "conflicts" whereas BARBE provides higher significance to them (see Figure 5). Moreover, BARBE discovers the rules containing a set of features e.g. a conjunction of features that makes more sense when explaining the decision of the black-box. All the rules have their support, confidence, and statistical significance values which are not available in LIME. On the other hand, Anchor generates only a single feature "conflicts" and treats this word as solely responsible for making the sentence negative. It misses the

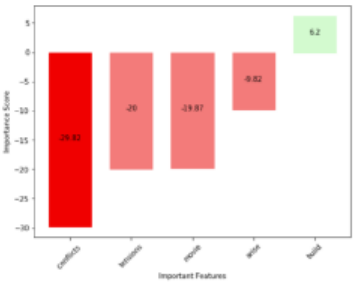
A. Sentence: A movie where tensions build and conflicts arise

B. Rules with Support, Confidence & Logarithm of Statistical Significance Values:

Sentence	Rules (Support, Confidence, and SS)
A movie where tensions build and conflicts arise	*tensions -> Negative* (48.91, 100%, -53.82)
A movie where tensions build and conflicts arise	*conflicts -> Negative* (43.56, 100%, -56.92)
A movie where tensions build and conflicts arise	*conflicts arise -> Negative* (25.36, 98.19%, -56.45)
A movie where tensions build and conflicts arise	*movie -> Negative* (50.81, 99.35%, -52.87)
A movie where tensions build and conflicts arise	*build -> Positive* (2, 31.01%, -55.92)



C. Blackbox Prediction Probabilities

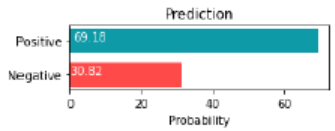


D. Feature Importance

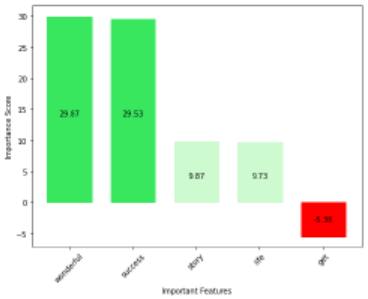
E. Sentence: Get this wonderful story about success in life

F. Rules with Support, Confidence & Statistical Significance Values:

Sentence	Rules (Support, Confidence, Statistical Significance)
Get this wonderful story about success in life	*wonderful -> Positive* (46.18, 100%, -382.93)
Get this wonderful story about success in life	*wonderful story -> Positive* (25.5, 98.72%, -382.02)
Get this wonderful story about success in life	*success -> Positive* (45.84, 99%, -269.74)
Get this wonderful story about success in life	*success life -> Positive* (26.86, 97.29%, -269.09)
Get this wonderful story about success in life	*get -> Negative* (13.75, 26.92%, -7.80)



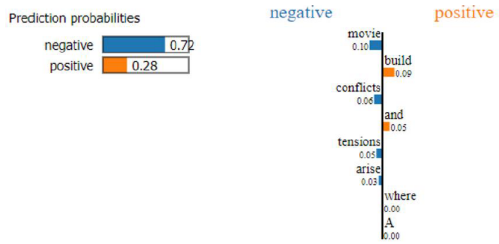
G. Blackbox Prediction Probabilities



H. Feature Importance

Fig. 5. The explanation provided by BARBE for two instances of the IMDB movie review dataset labeled by the black-box model. (A) shows a negative sentence with features highlighted. Red presents the negative words and green presents the positive ones. (B) presents the set of important rules with their support, confidence, and logarithm of statistical significance values Found by BARBE. (C) presents the prediction probability of the black-box, and (D) presents the histogram of important features ranked by BARBE. Figures E-F present the case for a positive sentence.

other significant words present in the sentence as BARBE and LIME figure out in their explanation.



A. Explanation generated by LIME

B. Explanation generated by Anchor

Fig. 6. Comparing the explanation generated by Lime and Anchor for the sentence: "A movie where tensions build and conflicts arise". (A) presents the explanation of LIME. The probabilities on the left of Figure A are the prediction probabilities of the underlying black-box model. On the right of Figure A, the features and their corresponding importance scores generated by LIME are shown in order of their importance. (B) presents the explanation of Anchor which only depends on the single word "conflicts" .

7 Conclusion and perspectives

We have presented a model-independent explanation framework based on associative classifiers, BARBE, that provides explanations for any black-box classifier in three forms: a set of ranked salient features that are relevant in the prediction of an instance; significant associations between features; and a set of interpretable classification rules that could explain the attribution of a class label to an instance. Unlike other methods, BARBE does not require from the black-box anything more than the predicted label for a given input. For a given input and its imputed label to explain, BARBE creates a perturbed sample around the input, requests labels from the black-box, and learns from the sample classification rules using an effective associative classifier. Taking advantage of the interpretability of the model generated by the surrogate learner, BARBE can then provide useful explanations. Compared to other prevalent methods that provide salient features, we have shown that BARBE has a better *Precision*, presents a better balance between *Precision* and *Recall* via the $F_{0.5}$ score,

and ranks better the features as indicated by a respectable *RBO* score. In addition, BARBE provides classification rules with associations between the features. Demonstrating the performance of BARBE on text makes the association rule more feasible to explain. By providing a conjunction of rules e.g. conjunction of features, the semantic fidelity is preserved by BARBE.

Associative classifiers are highly accurate but can generate noisy rules, which can be misleading when used in explanations. Pruning techniques can help address this issue, but finding more effective techniques could be beneficial. High dimensionality is also an issue, and an ensemble of associative classifiers can be used to partition the feature space. However, a better pruning of the search space would be an improvement. Associative classifiers are suitable for text categorization[3], and using BARBE for text classification explanations is straightforward, allowing the discovery of n-gram causes like with Anchor.

References

1. General data protection regulation (2 2020), `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation`
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Int. Conf. VLDB. vol. 1215, pp. 487–499 (1994)
3. Antonie, M.L., Zaiane, O.R.: Text document categorization by term association. In: IEEE International Conference on Data Mining. pp. 19–26. IEEE (2002)
4. Cohen, W.W.: Fast effective rule induction. In: Twelfth International Conference on Machine Learning (1995)
5. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`
6. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 (2018)
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 93 (2018)
8. Hamalainen, W.: Efficient discovery of the top-k optimal dependency rules with fisher's exact test of significance. In: 2010 IEEE International Conference on Data Mining. pp. 196–205. IEEE (2010)
9. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM sigmod record. vol. 29, pp. 1–12 (2000)
10. Jia, Y., Bailey, J., Ramamohanarao, K., Leckie, C., Ma, X.: Exploiting patterns to explain individual predictions. Knowledge and Information Systems pp. 1–24 (2019)
11. Li, J., Zaiane, O.R.: Exploiting statistically significant dependent rules for associative classification. Intelligent Data Analysis **21**(5), 1155–1172 (2017)
12. Li, W., Han, J., Pei, J.: Cmar: Accurate and efficient classification based on multiple class-association rules. In: Proceedings 2001 IEEE international conference on data mining. pp. 369–376. IEEE (2001)
13. Liu, B., Hsu, W., Ma, Y., et al.: Integrating classification and association rule mining. In: KDD. vol. 98, pp. 80–86 (1998)

14. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. pp. 142–150 (2011)
15. Meddahi, K., Benkabou, S.E., Hadjali, A., Mesmoudi, A., El Kefel Mansouri, D., Benabdeslem, K., Chaib, S.: Towards a co-selection approach for a global explainability of black box machine learning models. In: Chbeir, R., Huang, H., Silvestri, F., Manolopoulos, Y., Zhang, Y. (eds.) Web Information Systems Engineering – WISE 2022. pp. 97–109. Springer International Publishing, Cham (2022)
16. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)
17. Pastor, E., Baralis, E.: Explaining black box models by means of local rules. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 510–517 (2019)
18. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., MacDonell, C., Anvik, J.: Visual explanation of evidence with additive classifiers. In: Proceedings of the National Conference on Artificial Intelligence. vol. 21, p. 1822. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2006)
20. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)
21. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: 22nd ACM SIGKDD international conference on Knowledge Discovery and Data mining. pp. 1135–1144 (2016)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
23. Secretariat, T.B.o.C.: Directive on automated decision-making (Aug 2017), `https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592`
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
25. Shahroudnejad, A.: A survey on understanding, visualizations, and explanation of deep neural networks. preprint arXiv:2102.01792 (2021)
26. Sukel, K.: Artificial intelligence ushers in the era of superhuman doctors (7 2017), `https://www.newscientist.com/article/mg23531340-800-artificial-intelligence-ushers-in-the-era-of-superhuman-doctors/`
27. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS) **28**(4), 1–38 (2010)