Discussion Paper
Deriving Local Explanations for Black Box Models

Eliana Pastor and Elena Baralis

Politecnico di Torino, Italy
{eliana.pastor, elena.baralis}@polito.it

Abstract. Many high performance machine learning methods produce black box models, which do not disclose their internal working that yields the prediction. We propose a novel explanation method that explains the predictions of any classifier by analyzing the prediction change obtained by omitting relevant subsets of attribute values. Our method overcomes the exponential time complexity of previous works by learning a local model in the neighborhood of the prediction to explain. Preliminary experiments show that, despite the approximation introduced by the local model, the explanations provided by our method are effective in detecting also correlation among attributes. Our method is model-agnostic. Hence, experts can compare explanations and local behaviors of the predictions for the same instance made by different classifiers.

Keywords: Interpretability · Prediction Explanation · Local model.

1 Introduction

The application of machine learning algorithms is becoming pervasive in every aspect of our society. Since classification models could greatly affect people lives, understanding how a classification model works or why a decision is made is gaining more and more importance. Accuracy and interpretability of a machine learning model are frequently considered as a trade-off: the greater is the accuracy of a model, the lower is its understandability. The experts often favor accuracy over interpretability. However, an accurate model does not imply a trustworthy one.

Given the importance of interpretability, we propose a novel explanation method that explains the predictions made on single instances by any classifier. This methodology is model-agnostic. Hence, it is applicable to any classification method without making any assumption on its internal logic. The explanation highlights the feature values of a particular instance that are relevant for the prediction made by a specific classifier. The explanation is based on the knowledge of the local behavior of the model, captured by an interpretable local model.

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy.

The paper is organized as follows. Section 2 describes related work. Section 3 introduces the proposed technique, while Section 4 describes some preliminary experiments that show the effectiveness of our method. Finally, Section 5 draws conclusions and outlines future works.

2 Related Work

Many algorithms have been proposed for improving the interpretability of already existing classification models. We can identify two main approaches: model-dependent solutions and model-agnostic ones.

Model-dependent solutions are applicable only for specific classification models. Ad hoc solutions have been proposed for improving the understandability of neural networks [2, 20], Naive Bayes models [16], Support Vector Machines (SVM) [3, 6, 13], and random forest models [5, 12, 15]. These methods only address some specific classification algorithms. The explanation of how the models work is presented by means of different techniques, e.g. rules [2, 3, 12, 15], visualizations [6, 20], nomograms [13, 16], global feature importance [5]. Thus, no comparison among different techniques in terms of model interpretability is possible.

Model-agnostic or model-independent solutions treat the machine learning model as a black box. Some solutions try to explain the original model globally. As an example, the algorithm TREPAN approximates a generic model f learning a classification tree on the predictions of f [7]. It can be argued that the interpretable but simple decision tree could not be able to mimic the complexity of the whole model.

Other approaches propose a general method for explaining individual predictions, i.e. why particular decisions are made. Ribeiro et al. [18] introduce a model-agnostic method for explaining individual prediction by learning an interpretable and linear model in the locality of the prediction to be explained. However, the linear approximation may not be faithful if the model is highly non-linear even in the locality of the prediction [18].

Several works study how a prediction changes if parts of the input components are omitted. Fong and Vedaldi apply this approach for images' classification, approximating the elimination of parts of an image with meaningful perturbations [11]. Lemaire et al. [14] and Robnik-Šikonja and Kononenko [19] consider how each attribute value is relevant for the prediction for tabular data, by omitting one attribute value at a time. Štrumbelj et al. study also the omission of more attribute values together, thus also addressing the attribute interaction [22]. However, they compute the omission effect for the power set of the attributes. Hence, the method is affected by an exponential time complexity. A first solution for overcoming this problem is based on a sampling-based approximation [21]. The sampling is quasi-random and adaptive and it is based on a greedy approach, considering data characteristics, as the feature variance. We overcome the problem of the exponential time complexity by exploiting local properties of the original model to be explained.

3 Deriving Local Explanations

We propose a novel method to explain individual predictions of any classification model. Our method highlights the relevance of each attribute value in an instance for the prediction of its class label. The classification model is treated as a black box. Given the prediction that we want to explain, we "remove" one or more attribute values at a time and we measure how the prediction changes. If the prediction output changes, it means that the considered attributes are relevant for the prediction of the instance. The relevance of the attributes can be estimated as a difference of prediction probabilities with respect to a particular target class. The larger the difference, the more the omitted attribute values are relevant for the prediction.
With respect to previous approaches [14, 19, 22], we propose a novel methodology based on a local and interpretable model learned in the neighborhood of the prediction that we want to explain. The local model is provided by an associative classifier and highlights the subsets of feature values that are relevant for the prediction. Our explanation method is characterized by the following features.

- Only the relevant attribute subsets provided by the local model are considered, instead of the complete power set of all attribute combinations. Hence, our approach overcomes the exponential time complexity. Furthermore, the local approximation is based on the behavior of the original model.
- The local model is rule-based and it returns the association between subsets of feature values and class. The rules, being understandable, provide preliminary insight of why a particular decision is made by the considered model. This allows a qualitative understanding of the original model behavior.

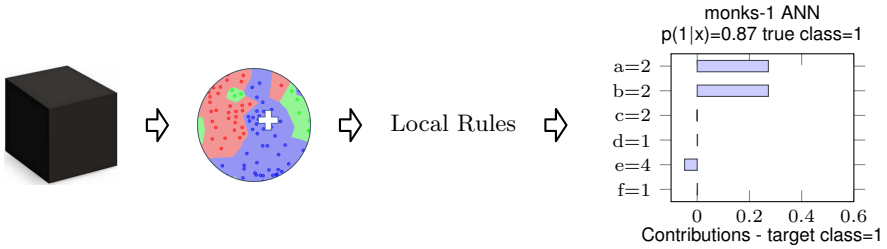


Fig. 1: Steps of the explanation method.

In Figure 1, the steps of the explanation method are outlined. Considering the original model as a black box, we learn a local model in the locality of the prediction to be explained, the big white cross. We obtain from the local model a set of relevant rules that indicate the relevant subsets of feature values. Only these subsets are then used for estimating the contribution of each attribute value.

Let be f a generic trained classification model whose predictions we want to explain. f can be produced by any classifier because the proposed explanation method is model-agnostic. Given an instance x, our technique requires that, for each class c, the model f returns the class probability, $p(y{=}c \mid x)$, i.e. the probability that x belongs to class c. Many classification algorithms, the probabilistic classifiers (e.g. Naive Bayes, artificial neural networks), naturally provide class probabilities. When class probabilities are not available, we firstly apply post-modeling methods, the probability calibration methods, to obtain posterior probabilities [17].

3.1 Capturing the Locality by means of K Neighbors

Our goal is to understand the subsets of feature values that are relevant for the prediction of a particular instance x. We train a local interpretable model able to directly produce the relevant subsets. The interpretable model is trained only in the locality of x. The local training data is computed considering the K instances in the training set that are nearest to the instance x that we want to explain.

The choice of parameter K is important because it affects the generated model. To estimate K, we exploit techniques proposed for estimating the parameter K of the K-Nearest Neighbors classifier [10]. As future work, we will study a heuristic algorithm that automatically selects an optimal K for the particular prediction to be explained.

Once the K neighbors of the instance x are selected, they are labeled by the model f, whose prediction we want to explain.

3.2 Extracting Local Rules

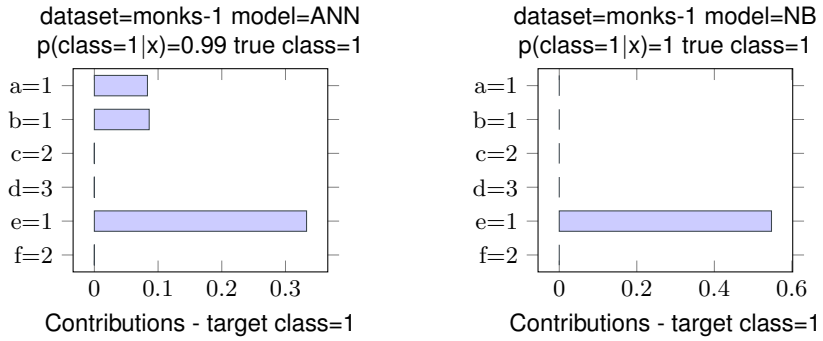
The K labelled neighbors of instance x are used for training the local model. The local model provides the feature values that are relevant for the prediction. We use a rule-based classifier, more specifically an associative classifier. This kind of classifiers extracts classification rules from the training data. Association rules are in the form $A \rightarrow B$, where A is a set of items, and B, if the rules are used for classification purposes, is a class label [1]. These rules are denoted as CARs. An item is a pair (*attribute, value*). CARs highlight the subsets of feature values that are associated with the class label. In our implementation, as rule-based system we use the classifier L^3, *Live and Let Live*, an associative classifier that is based on a lazy pruning approach [4]. L^3 is trained with the K neighbors of the instance x that we want to explain. This local model is used for obtaining the *local rules*. From the local rules we retrieve the set of (*attribute=value*) pairs that are considered relevant for the classification task in the locality of the instance x to be explained.

3.3 Computation and Visualization of Attribute Contributions

The relevant feature subsets provided by the selected CARs drive the estimate of the feature contributions. We estimate the contributions of each attribute

value by means of the definitions of *prediction difference* and *interaction contribution* proposed in [22]. We modify the original definitions to consider the relevant subset instead of the complete power set. These contributions represent the influence of each feature value in the determination of the class, for the prediction on a single instance x of the model f. The larger the value of its contribution, the more the corresponding attribute value determines the class. A positive contribution means that the attribute value has a positive influence in determining the class. A negative one, instead, means that the attribute value speaks against the prediction for that particular target class. The estimation of the contributions considers the subsets of attributes highlighted by the local model. The contributions of each attribute value can be visualized through a bar plot representation, following the visualization method proposed by Kononenko et al. [19]. The visualization allows the final users to understand in a simple and uniform way the motivation driving the prediction for instance x made by model f. An example of visualization is presented in the last step of Figure 1.

4 Preliminary Results



(a) (b)

Fig. 2: Comparison of explanations of a particular instance of the *monks-1* data set. Explanation of (a) the neural network prediction and (b) the Naive Bayes prediction.

In this section, preliminary outcomes of our explanation method are presented and discussed. The considered data set is the artificial data set *Monk1* [9]. The data set is composed by 6 discrete attributes *a,b,c,d,e,f* and the class label can take value 1 or 0. Being artificial, the relationship between the attributes and the class value is known. The class is 1 if *a=b* or if *e=1*, 0 otherwise. Thus, it is possible to verify the results of our explanation methods, comparing explanations with the true relations among attributes.

To build the models and perform classification, we exploit the Python-based Orange 2.7 Data Mining Library [8].

As a first example, we train a multilayered feed-forward artificial neural network (ANN) using the *Monk1* data set. Let x = (a=1, b=1, c=2, d=3, e=1, f=2) be the instance that we want to explain. We know that the "true class" is 1 because *e=1* and *a=b*. Thus, e and both a and b in this case are important for the prediction. The ANN correctly predicts the class label as 1 with probability *p(class=1|x)* equal to 0.99. To estimate what are the relevant subsets of feature values for the ANN for this particular instance, we train the associative classifier L^3 in the locality of instance x. In these experiments, the parameter K is set to the square root of the number of instances of the training data set [10]. For the support and confidence thresholds, we use L^3 default values, in particular 1% and 50% respectively. Preliminary experiments of sensitivity for these thresholds show that the results of the local model are stable if the values of support and confidence are changed in the neighborhood of the default values. The local model returns the following CARs:

$$\{e = 1\} \rightarrow class = 1$$
$$\{a = 1,\ b = 1\} \longrightarrow class = 1$$

Hence, if *e=1* the instance is assigned to the class 1 or also if a and b are both equal to 1. These relations, based on our knowledge of the *Monk1* data set, should indeed determine the class. The local behavior captures the true explanation. Once this relevant subset is defined, we can compute the contribution of each attribute value to the prediction [22]. The result is shown in Figure 2a. The largest contribution is given by term *e=1*, followed by the terms a and b.

If we explain the prediction for the same instance and still with respect to class 1, made by another model, we may obtain a different result. The explanation should capture how the model behaves in the locality of the instance. Different models work differently. This difference may be on the predicted class label, but also on the feature values that drive the prediction.

Consider the explanation of the same instance x and still built with respect to class 1, but classified by the Naive Bayes classifier. The local model returns a single relevant rule:

$$\{e = 1\} \rightarrow class = 1$$

Only attribute e equal to 1 is considered relevant for class 1. The resulting contributions are shown in Figure 2b. The Naive Bayes classifier assigns correctly the instance x to class 1, but only because *e=1*. The local model and the explanation highlight that the Naive Bayes classifier has not learned the association that if *a=b* then *class=1*. The Naive Bayes classifier, because of its assumption of independence between features, is not able to learn the importance that a and b have together. Hence, the local model and the explanation in this case successfully reflect the model behavior.

Despite the approximation introduced by considering only relevant subsets, in our preliminary tests we obtain explanations comparable to the results in

previous works, in which the complete power set of attribute values is taken into consideration [22].

We can compare the predictions of the same instance made by different classifiers by simply comparing the attribute value contributions and local rules. The comparison allows experts to inspect the local behavior of different classifiers. Based on their prior knowledge, they can decide which prediction to trust. It is important to notice that these considerations can be made by domain experts regardless of the intrinsic interpretability of the model. Being our method model-agnostic, the explanations are presented in the same way for all classifiers, ranging from the naturally interpretable, as decision trees, to the black box ones, as the neural networks.

5 Conclusions and Future Work

We propose a novel model-agnostic explanation method that explains the individual predictions of any classifier. The original classification model is treated as a black box. We omit subsets of attribute values and we measure how the prediction changes. We overcome the exponential time complexity that derives from the computation of the power set of the feature values by learning a local model. The local model is an associative classifier that is learned in the locality of the instance whose prediction is to be explained. It returns the subsets of feature values that are relevant for that particular prediction: only these subsets are omitted.

Preliminary tests show that our technique is able to capture the diverse internal logic of different classification techniques. Accuracy and interpretability in this way can no longer be considered as a trade-off. The experts can choose accurate models but also verify if these models can be trusted. Hence, our explanation method helps users in the selection of the best classification approach.

As future work we plan to (i) evaluate the approximation error introduced by considering only the relevant subsets highlighted by the local model and (ii) define a technique to automatically detect an appropriate value of K, the number of neighbors to be considered in the local model, to adapt its value to different data distributions.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec. **22**(2), 207–216 (Jun 1993)
2. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems **8**(6), 373–389 (1995)
3. Barakat, N., Bradley, A.P.: Rule extraction from support vector machines: a review. Neurocomputing **74**(1-3), 178–190 (2010)
4. Baralis, E., Chiusano, S., Garza, P.: A lazy approach to associative classification. IEEE Transactions on Knowledge and Data Engineering **20**(2), 156–171 (2008)

5. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
6. Caragea, D., Cook, D., Honavar, V.G.: Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 251–256. KDD '01, ACM, New York, NY, USA (2001)
7. Craven, M., Shavlik, J.W.: Extracting tree-structured representations of trained networks. In: Advances in neural information processing systems. pp. 24–30 (1996)
8. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., et al.: Orange: data mining toolbox in python. The Journal of Machine Learning Research **14**(1), 2349–2353 (2013)
9. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
11. Fong, R., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. arXiv preprint arXiv:1704.03296 (2017)
12. Hara, S., Hayashi, K.: Making tree ensembles interpretable. arXiv preprint arXiv:1606.05390 (2016)
13. Jakulin, A., Možina, M., Demšar, J., Bratko, I., Zupan, B.: Nomograms for visualizing support vector machines. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 108–117. ACM (2005)
14. Lemaire, V., Feraud, R., Voisine, N.: Contact personalization using a score understanding method. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 649–654 (June 2008)
15. Mashayekhi, M., Gras, R.: Rule extraction from random forest: the rf+ hc methods. In: Canadian Conference on Artificial Intelligence. pp. 223–237. Springer (2015)
16. Možina, M., Demšar, J., Kattan, M., Zupan, B.: Nomograms for visualization of naive bayesian classifier. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 337–348. Springer (2004)
17. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632. ACM (2005)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. pp. 1135–1144. KDD '16, ACM, New York, NY, USA (2016)
19. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering **20**(5), 589–600 (May 2008)
20. Tzeng, F.Y., Ma, K.L.: Opening the black box-data driven visualization of neural networks. In: Visualization, 2005. VIS 05. IEEE. pp. 383–390. IEEE (2005)
21. Štrumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. **11**, 1–18 (Mar 2010)
22. Štrumbelj, E., Kononenko, I., Robnik Šikonja, M.: Explaining instance classifications with interactions of subsets of feature values. Data Knowl. Eng. **68**(10), 886–904 (Oct 2009)