DOI: 10.15439/2022F258
ISSN 2300-5963 ACSIS, Vol. 30

Rule-based approximation of black-box classifiers for tabular data to generate global and local explanations

Cezary Maszczyk*†, Michał Kozielski‡ and Marek Sikora†‡
*Doctoral School, Silesian Univeriesty of Technology,
ul. Akademicka 2A, 44-100 Gliwice, Poland
†Łukasiewicz Research Network – Institute of Innovative Technologies EMAG,
ul. Leopolda 31, 40-189 Katowice, Poland
‡Department of Computer Networks and Systems, Silesian Univeriesty of Technology,
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: marek.sikora@polsl.pl

***Abstract*—The need to understand the decision bases of artificial intelligence methods is becoming widespread. One method to obtain explanations of machine learning models and their decisions is the approximation of a complex model treated as a black box by an interpretable rule-based model. Such an approach allows detailed and understandable explanations to be generated from the elementary conditions contained in the rule premises. However, there is a lack of research on the evaluation of such an approximation and the influence of the parameters of the rule-based approximator. In this work, a rule-based approximation of complex classifier for tabular data is evaluated. Moreover, it was investigated how selected measures of rule quality affect the approximation. The obtained results show what quality of approximation can be expected and indicate which measure of rule quality is worth using in such application.**

I. Introduction

IN RECENT years, eXplainable Artificial Intelligence (XAI) [1], [2], [3], [4] has become an increasingly important field of artificial intelligence (AI). The explanations generated by XAI methods allow people to understand how artificial intelligence methods work and are useful for many types of users involved in different ways with AI applications.

This work focuses on XAI derived from rule induction [5]. Rule-based models represent knowledge embedded in data in the form of IF-THEN rules. The premise of a rule is a conjunction of elementary conditions $w_i \equiv a_i \odot x_i$, with x_i, being an element of the attribute a_i domain and $\odot$ representing a relation ($=$ for symbolic attributes; $<$, $\leq$, $>$, $\geq$ for ordinal and numerical ones). The conclusion of the rule contains a decision, which can be either symbolic or numeric.

Rules and decision trees (which can be easily transformed into mutually disjoint rules) are used to generate global explanations because of their interpretability [4]. Such global explanations are obtained by generating a rule-based interpretable model that approximates the non-interpretable (complex) base model. The use of rule-based approximation is justified for tabular data where attributes are interpretable and the generated rules can be understood by humans.

There are recent examples in the literature of the use of rule-based systems to obtain model agnostic explanations. Rule-based models approximating the complex base model locally were considered in [6]. The works [7], [8] considered rules as interpretable expressions used to present explanations within a proposed local-to-global approach for black-box explanations.

The parameters of the rule-based model induction method affect the quality of the base model approximation. Furthermore, they affect the explanations generated from the set of input rules. To the best of the authors' knowledge, there is no published analysis showing how well a rule-based model can approximate the complex base model being explained. Therefore, the aim of this paper is to present and evaluate the rule-based, model-agnostic approach to explaining machine learning models. The rule-based approximation of a black-box model may be evaluated in terms of its accuracy and number of the generated rules. Therefore, three measures of rule quality used in the induction process are verified in this work: Correlation, C2 and Precision [9]. These measures, in the above order, generate more and more specific rules and thus build probably more and more accurate approximators but composed of an increasing number of rules. Verification of these parameters should provide clues that are useful in the development of rule-based approximators used to create explanations.

The contribution of this paper includes: (i) analysis and evaluation of the approach that generates global and local explanations from a rule-based approximation of a black-box model, (ii) verification of how the application of different measures of rule quality assessment (Correlation, C2 and Precision), which enable variation in the approximation accuracy of a machine learning model, affects the explanations obtained.

This work was partially supported by: Computer Networks and Systems Department at Silesian University of Technology within the statutory research project and the Łukasiewicz Research Network (ROLAP-ML, R&D grant)

II. The idea of rule-based explanations

The proposed approach to generating machine learning model explanations involves approximating the base model treated as a black-box using a rule-based model. Base model approximation means that the rule-based model learns the decisions of the base model. Such a learning process requires transformed training data in which the existing decision variable is replaced by the decision of the base model. In the extreme case (which may mean over-fitting the base model), the transformed training data may be identical to the original.

The generated rule-based model, although interpretable in theory, is not always clear, e.g. due to the number of rules. However, once the rule-based model is generated, the elementary conditions of the rules and their importance can be extracted. On their basis, it is possible to generate the explanations showing which attributes, and which attribute value ranges, are most important for decision making by the base model. Importance based rankings of elementary conditions are generated using the Shapley index [10] and details on the assessment of the importance of rule elementary conditions are presented in the work [11].

Using the above method to explain the black-box classifier, global and local explanations can be obtained. The global explanation of the classifier is in the form of a ranking of elementary rule conditions indicating which ones have the highest importance in the decision of the base model about the selected class. The local explanation, on the other hand, consists of a set of rules covering the data instance and an analogous ranking indicating the importance of the elementary conditions in the classifier's decision.

III. Evaluation of rule-based approximation

The first step in the presented concept of generating rule-based explanations of the base machine learning model is to approximate this model with a rule-based model. The base model approximation should be of high accuracy to generate reliable explanations. The rule induction algorithm can be controlled by a number of parameters. This study focuses on the rule quality measure used and verifies impact of three such measures: Correlation, C2 and Precision.

The approach presented in this paper was implemented in Python. The scikit-learn package [12] was used to generate base models and the coverage algorithm implementation available in the RuleKit package [13] via a Python wrapper[1] was used to generate the rule-based models. Four different algorithms were selected to generate complex, non-interpretable classifiers used as base models. The selected algorithms are: Artificial Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM) and Extreem Gradient Boosting (XGBoost). The base models and the approximating rule-based models were evaluated in 10-fold cross validation process. Experiments were conducted on a set of 29 data sets available in the OpenML[2] and UCI[3] repositories. These data sets consist of tabular data and relate to the task of classification.

In the experiments, the quality of the implemented approximation with the rule-based model was evaluated in two ways. Firstly, the classification quality measures, such as Accuracy (Acc) and Balanced Accuracy (BAcc), were calculated. The quality of the approximator was determined for the transformed data in which the decision attribute represents the decision of the base model. Table I presents mean values calculated for all the data sets that were divided into train and test data. In the experiment, four base models were approximated and three rule quality measures were verified as values of the approximator parameter. Additionally, Table I presents the average number of rules that were generated by the induction algorithm.

TABLE I
Quality of rule-based approximations on transformed training and test data, for which the decision attribute is the decision of the base model

Base model	Approx. param.	#Rules	Train data		Test data	
			Acc	BAcc	Acc	BAcc
NN	Correlation	19.9	0.938	0.932	0.894	0.878
	C2	34.2	0.968	0.965	0.916	0.902
	Precision	75.3	0.988	0.986	0.912	0.892
RF	Correlation	21.1	0.924	0.92	0.873	0.858
	C2	38.3	0.958	0.953	0.899	0.884
	Precision	82.5	0.979	0.976	0.894	0.877
SVM	Correlation	18.7	0.952	0.949	0.915	0.905
	C2	30.4	0.978	0.976	0.935	0.922
	Precision	67.1	0.989	0.986	0.938	0.922
XGB	Correlation	21.1	0.924	0.921	0.877	0.863
	C2	38.4	0.958	0.953	0.9	0.886
	Precision	81.1	0.98	0.977	0.893	0.877

Based on the results in Table I, it can be concluded that the rule-based models approximate complex models well. Furthermore, the results show that the accuracy of this approximation decreases significantly on the test data compared to the training data. This is an expected result, as rule-based models were generated to obtain the best possible approximation on the training data.

Besides, the results in Table I show that the selected measures - in order: Correlation, C2 and Precision - enable to obtain increasingly fit rule-based models. The increase in quality results from the increase in the number of rules. In other words, the more general the rules are the less accurate the approximation becomes. On test data, however, the best quality results are obtained by models generated using the C2 measure.

The generated base models and their approximators can be further evaluated on the original data sets. Fig. 1 presents the difference in Balanced Accuracy (BAcc) between the generated approximations and base models. The BAcc values were calculated on test data for each of the data sets. XGBoost was used as a base model and the C2 measure was used in rule-based approximation generation.

[1] https://github.com/adaa-polsl/RuleKit-python

[2] https://openml.org/

[3] https://archive.ics.uci.edu/ml/datasets.php

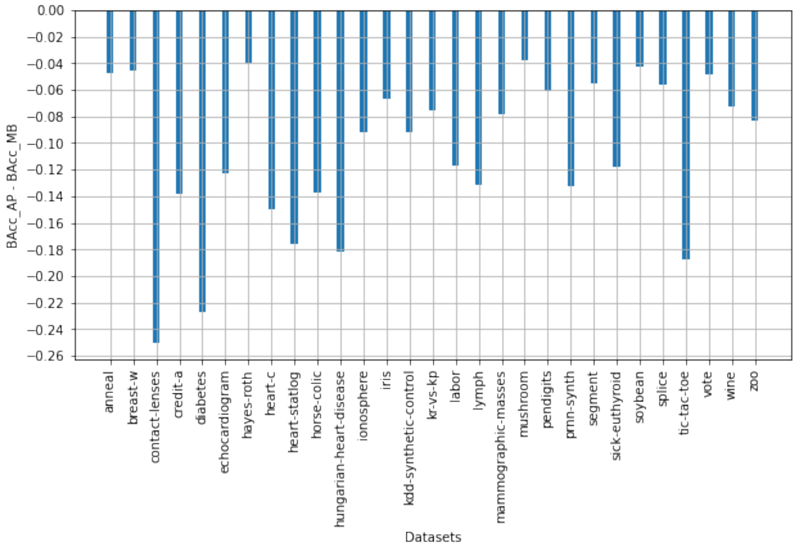

Fig. 1. Approximation quality illustrated as the difference of the Balanced Accuracy values for rule-based approximation (BAcc_AP) and base model (BAcc_MB) - comparison on test data

The results illustrated in Figure 1 show that an approximator operating on the original data is never better than the base model. These results provide a negative answer to the question of whether creating a rule-based model that cannot distinguish between examples that even complex models cannot correctly classify will produce rule sets with good classification quality. Moreover, it is again apparent that over-fitting to the train data results in lower accuracy on test data.

Within the second approach to rule-based approximation evaluation the generated explanations were compared. Therefore, using the SHAP method, attribute rankings were generated for the base models and their approximations. If the size of the data set exceeded 100 instances a subset of that size was selected to determine the Shapley values. Next, the top three features from both rankings were compared to verify if these three most important attributes were the same for both models. The results of the comparison are illustrated in Fig. 2. It presents for each data set the average number of identical features selected as the top three. This average is calculated as four base models (NN, Random Forest, SVM and XGBoost) were generated again for each data set. The rule-based approximation was performed using the C2 measure, which proved to be the most reasonable choice in earlier experiments. The comparison was performed on 31 data sets.

The results presented in Fig. 2 show that in most cases at least two out of three most important features are the same for the base model and its approximator.

IV. Exemplary use cases

Having the rule-based approximator of sufficient accuracy induced it is possible to generate explanations. The example global and local explanations presented below were generated from the rule-based model using RuleXAI package[4]. This package generates explanations by analysing the importance of the rules' elementary conditions [11]. The C2 measure was used for rule quality evaluation. The rule-based model approximated the XGBoost classifier.

The base model was trained on the wine[5] data set representing a three-class problem and consisting of the following attributes: Alcohol, Malic_acid, Ash, Alcalinity_of_ash, Magnesium, Total_phenols, Flavanoids, Nonflavanoid_phenols, Proanthocyanins, Color_intensity, Hue, OD280_OD315_of_diluted_wines, Proline, Class.

The generated global explanations are presented in Table II. They consist of rankings built separately for each class. The rankings were built on the basis of feature importance, or more precisely on the basis of the importance of the elementary conditions of the rules.

Local explanations are generated for a specific data instance. The analysed data instance takes the following values for the above list of attributes {0.213157895, 0.029644269, 0.652406417, 0.381443299, 0.260869565, 0.420689655, 0.394514768, 0.169811321, 0.611987382, 0.151023891, 0.25203252, 0.663003663, 0.172610556, 1}. This data instance was covered by the following two rules from the approximating model:

```
IF Color_intensity = (-inf, 0.19) THEN
ref_prediction = 1.0
IF Alcohol = (-inf, 0.39) AND Flavanoids =
<0.13, inf) THEN ref_prediction = 1.0
```

In addition to the explanation in the form of rules, RuleXAI generated a ranking presented in Fig. 3. It shows which ranges of attribute values were most important for a given decision.

[4] https://github.com/adaa-polsl/RuleXAI

[5] https://archive.ics.uci.edu/ml/datasets/wine

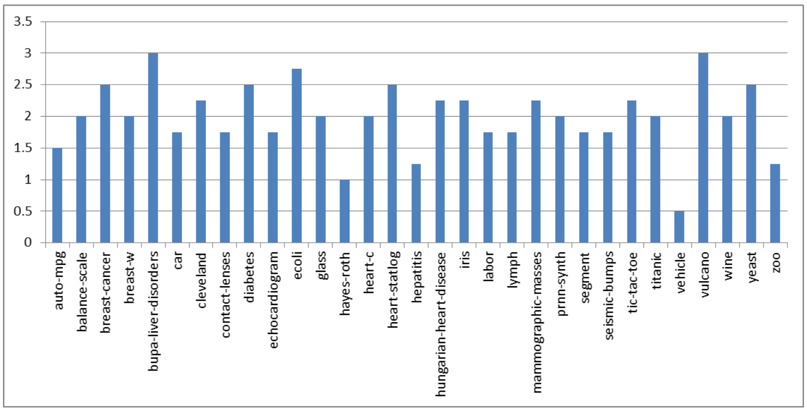

Fig. 2. Number of the same features in the first three positions of the rankings generated by the SHAP method for the base model and its approximator

TABLE II
RULE-BASED GLOBAL EXPLANATION GENERATED BY RULEXAI - THE EXPLANATION OF THE XGBOOST MODEL GENERATED ON WINE DATA SET

Class 0		Class 1		Class 2	
Condition	**Importance**	**Condition**	**Importance**	**Condition**	**Importance**
Proline $\geq$ 0.51	0.872	Color_intensity < 0.19	0.906	Flavanoids < 0.13	0.841
Proline $\geq$ 0.4	0.609	Alcohol < 0.39	0.728	OD280_OD315_of_diluted_wines < 0.2	0.825
Alcohol $\geq$ 0.52	0.352	Flavanoids $\geq$ 0.13	0.147	Ash $\geq$ 0.37	0.05
				Malic_acid $\geq$ 0.074	0.034

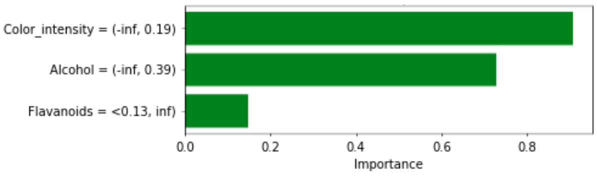

Fig. 3. Rule-based local explanation showing which attributes and which ranges of attribute values were most important for a given decision

V. CONCLUSIONS

The method generating global and local explanations using rule-based approximation of the black-box model was analysed within the presented research. In addition, global and local explanations generated from the rule-based approximation were presented. The explanations generated from the rules provide a detailed understanding of which features in which value ranges are important to the classifier's decisions.

Application of the rule-based model to approximation of the complex classifier generated on tabular data was positively evaluated. The approximation obtained with the rule-based models was consistent with the approximated base models both from the point of view of classification quality measures and from the perspective of consistency in the most important features indicated by the generated explanations.

REFERENCES

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. [Online]. Available: https://doi.org/10.1145/3236009

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[3] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[4] P. Biecek and T. Burzykowski, *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman and Hall/CRC, 2021.

[5] J. W. Grzymala-Busse, *Rule Induction*. Boston, MA: Springer US, 2005, pp. 277–294. [Online]. Available: https://doi.org/10.1007/0-387-25465-X_13

[6] E. Pastor and E. Baralis, "Explaining black box models by means of local rules," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 510–517. [Online]. Available: https://doi.org/10.1145/3297280.3297328

[7] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, "Meaningful explanations of black box ai decision systems," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9780–9784.

[8] M. Setzu, R. Guidotti, A. Monreale, and F. Turini, "Global explanations with local scoring," in *Machine Learning and Knowledge Discovery in Databases*, P. Cellier and K. Driessens, Eds. Cham: Springer International Publishing, 2020, pp. 159–171.

[9] M. Sikora and Ł. Wróbel, "Data-driven adaptive selection of rule quality measures for improving rule induction and filtration algorithms," *International Journal of General Systems*, vol. 42, no. 6, pp. 594–613, 2013.

[10] L. S. Shapley, "A value for n-person games," *Classics in game theory*, vol. 69, 1997.

[11] M. Sikora, "Redefinition of decision rules based on the importance of elementary conditions evaluation," *Fundamenta Informaticae*, vol. 123, no. 2, pp. 171–197, 2013.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[13] A. Gudyś, M. Sikora, and Łukasz Wróbel, "Rulekit: A comprehensive suite for rule-based learning," *Knowledge-Based Systems*, vol. 194, p. 105480, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120300046