MODEL AGNOSTIC GLOBALLY INTERPRETABLE EXPLANATIONS

Piyush Gupta, Nikaash Puri, Sukriti Verma, Pratiksha Agarwal & Balaji Krishnamurthy*
Adobe Systems
Noida, Uttar Pradesh - 201304, India
{piygupta, nikpuri, sukrverm, pratagar, kbalaji}@adobe.com

ABSTRACT

Explaining the behavior of black box machine learning through human interpretable rules is an important research area. Several recent works have focused on explaining model behavior locally i.e. for specific predictions. However, it is also important to understand the behavior of models globally. In this work, we present a novel approach that captures model behavior globally in an accurate, succinct, and human understandable manner. It uses local model explanation methods to extract conditions important for specific instances followed by an evolutionary algorithm that optimizes an information theory based fitness measure to construct global rules. We show how our approach can be used in different domains to extract patterns in a dataset through a trained model. Our approach outperforms existing approaches across several publicly available data sets.

1 INTRODUCTION

Black box models are increasingly used to assist in crucial decisions (Louzada et al., 2016; Corbett-Davies et al., 2017). It is therefore important to interpret the decisions taken by them (Lipton, 2018; Ribeiro et al., 2016a; Doshi-Velez, 2017). We describe a model-agnostic approach called **MAGIX** or **M**odel **A**gnostic **G**lobally **I**nterpretable E**x**planations to interpret classification models as if-then rules that explain model behavior globally. These rules provide useful insights into both the data and the model.

2 RELATED WORK

There are several approaches that explain model behavior at a local level, i.e. in a limited region of the input space (Lundberg & Lee, 2016; 2017; Shrikumar et al., 2017). LIME (Ribeiro et al., 2016b) explains the classification of a particular instance by a trained model. It has been extended to Anchors (Ribeiro et al., 2018) that outputs rules that explain only the reasons for a specific decision. Local approaches can be used to derive rules R_k that correctly explain a small region of the input space. However, since each R_k covers a small fraction of the input space, the number of rules that explain model behavior globally is large. We show this in Section 4.

Bastani et al. (2017) propose a surrogate model approach where a decision tree is trained on the predictions made by the model. However, rules extracted from a decision tree are in the form of decision lists. Lakkaraju et al. (2016) show that decision sets with independent rules are more interpretable than decision lists. They optimize an objective function that balances accuracy and interpretability of the ruleset. The candidate set of rules is derived using association rule mining. In contrast, our approach builds upon local model interpretation algorithms using them to mine conditions and employing an evolutionary algorithm to build the global ruleset out of these local conditions. We show in Section 4 that our approach outperforms existing approaches.

*All authors contributed equally.

Table 1: Contingency table used to compute *Fitness* of rule R_i

Rule/Class	y_{R_i}	NOT y_{R_i}
R_i	n_{11} = Number of instances in cover(R_i) with class y_{R_i}	n_{12} = Number of instances in cover(R_i) with class different from y_{R_i}
NOT R_i	n_{21} = Number of instances with class y_{R_i} but not covered by R_i	n_{22} = Number of instances not covered by R_i and with class different from y_{R_i}

3 APPROACH

Definitions and notations frequently used hereafter are as follows:

1. Dataset D is a set of x_i and class y_{x_i}. The trained model is a function $M: X \rightarrow Y$ that maps input x_i to class y_j. F_1 to F_m are the features used by the model.
2. Condition C_i is defined by a feature F_{C_i} and a value v_{C_i}. It represents the predicate $F_{C_i} = v_{C_i}$. A condition C_i is said to hold for an instance x_j if $x_j[F_{C_i}]$ is equal to v_{C_i}.
3. Rule R_k is a conjunction of conditions from C_1 to C_l with a class y_{R_k}. R_k covers x_j if each condition in R_k holds for x_j. R_k correctly covers x_j if R_k covers x_j and $y_{R_k} = M(x_j)$.
4. *coverage(R_k)* is the fraction of instances that are covered by R_k.
5. *precision(R_k)* is the fraction of instances in *coverage(R_k)* that are correctly covered by R_k.
6. Interpretation I of a machine learning model M is a set of rules R_1, R_2, ..., R_n. I should have the characteristics of **Interpretability**, **Accuracy** and **Fidelity** (Guidotti et al., 2018).

Our algorithm works in two phases. In the first, we use LIME (Ribeiro et al., 2016b) to find conditions that are important to explain the classification of a specific instance. This is repeated until we find at least one condition for each instance in the training set. At the end of this phase, we have a set of locally important conditions.

In the second phase, the locally important conditions are input to a genetic algorithm (Whitley, 1994) to evolve rules at the global level. The algorithm explores combinations of conditions, guided by a fitness function, to build a global, accurate and interpretable ruleset. It is run independently for each class to ensure that model behavior is explained for all classes. A candidate rule is encoded as a bit string with each bit marking the presence or absence of one condition. The desiderata for a rule are **precision**, **coverage** and **rule length**, as defined at the start of this section. In order to capture these, we introduce a fitness measure based on mutual information.

The Mutual Information between two variables quantifies their dependence. Within the context of learning a rule, the two variables are the model and the rule. We want to maximize the information that each rule provides us about the model. For a rule R_i that has class label y_{R_i}, we construct a contingency table as shown in Table 1. Mutual Information (MI) for a rule R_i is:

$$MI = \frac{1}{N} \sum_{a,b=1}^{2} n_{ab} log(\frac{n_{ab} \times N}{r_a \times c_b}) \tag{1}$$

where, N is sum of all values, r_a is summation of values in row a, c_b is summation of values in column b. The fitness measure for a rule is as follows:

$$Fitness(R_i) = \begin{cases} MI & when\ n_{11} \geq \frac{r_1 \times c_1}{N} \\ -1 \times MI & otherwise \end{cases}$$

A high value of MI means that the rule R_i accurately captures the model behavior for class y_{R_i}. However, a high MI could also mean a high negative correlation between the rule and predicted class. To penalize this, we negate the value of MI when the value in cell n_{11} is less than the expected value. The genetic algorithm is run for 600 generations. All the individuals of the final generation having positive fitness are selected as rules that explain model behavior.

4 RESULTS

We demonstrate our approach on publicly available datasets that have been summarized in Appendix A. Each dataset was split into a training, validation and scoring set. A Random Forest Model was trained on the training set. Rules were learned on the predictions made by this model using MAGIX. Figure 1 shows some of these rules. They give useful insights into these models.

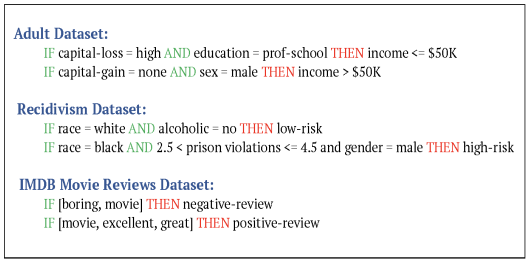


Figure 1: Illustrative Rules learned using MAGIX

Dataset	Approach	SC	SP
Adult	MAGIX	100.00	**88.95**
	Anchors	62.53	82.36
	SLS	99.81	76.08
Recidivism	MAGIX	99.40	**83.49**
	Anchors	28.23	69.86
	SLS	99.98	62.66

Table 2: Set-Coverage and Set-Precision with Simulated Users on rulesets of 20 rules

For instance, we find that the model trained on the **Recidivism** dataset had a racial bias. To predict whether a convict will re-offend, race was a distinguishing factor. The condition *"Race = White"* was a part of a large number of rules that predict a low recidivism score, whereas the condition *"Race = Black"* was present in most rules that predicted a high score.

On the **IMDB Movie Reviews** Dataset, we interpreted a pre-trained LSTM Network. Both the code and the model is open-source (Deshpande, Adit). We learn rules of the form, *If [W_1, W_2, ..., W_n] Then Class K* where *[W_1, W_2, ..., W_n]* signifies a presence of all W_i in the review. A few rules learned on the LSTM Model are listed in Figure 1.

4.1 EXPERIMENTAL SETUP

We compare the performance of our approach with Anchors (Ribeiro et al., 2018) and Smooth Local Search (SLS) (Lakkaraju et al., 2016) on the **Adult** and **Recidivism** datasets. The training set was used to train a random forest model. Rules explaining model behavior were learned using different approaches. For generating rules using Anchors, we repeatedly generate anchors on instances such that all the instances are correctly covered by at least one anchor. Apriori algorithm (Agrawal et al., 1996) is used for generating candidate rules on which a selection is performed using SLS (Lakkaraju et al., 2016).

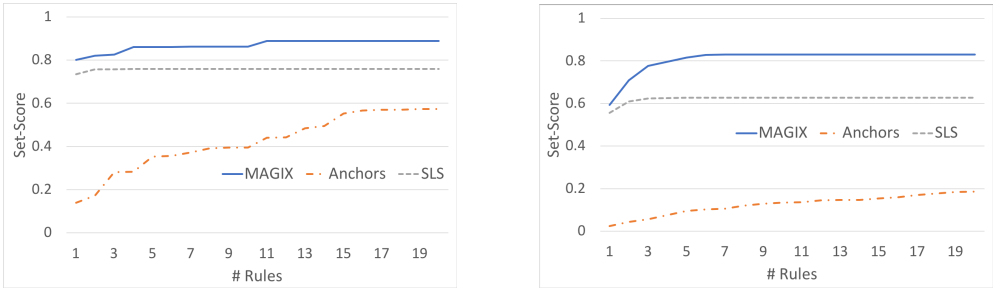
To build a global ruleset we use a variant of the submodular pick technique described in LIME (Ribeiro et al., 2016b). The rules obtained are sorted in the order of decreasing marginal coverage gain. Global rulesets limited by two specifications are formed for each of the approaches being compared. The size of the rule sets is limited by either the number of rules or the fraction of instances covered by the rule set. We call these Cognitive Budget and Global Coverage respectively.

4.2 SIMULATED USER STUDY

This study was performed on validation sets from **Adult** and **Recidivism** datasets (Experiments section, Ribeiro et al. (2018)). An Interpretation of 20 rules is learned using each approach. *Set-Coverage* and *Set-Precision* are computed for each rule set. The *Set-Coverage* (SC) of an interpretation I, is the fraction of instances that are covered by at least one rule in I. The *Set-Precision* (SP) of an interpretation I, is the fraction of instances from the set of covered instances for which the model and the interpretation agree on the class label. For instances covered by multiple rules, the highest precision rule is selected and associated class assigned.

Table 2 shows the results. It shows that while both MAGIX and SLS generate interpretations that cover the entire set of instances, the rules generated by MAGIX correctly explain model behavior on a larger fraction of the data. The rules generated by Anchors cover a smaller fraction of instances.

To evaluate the fidelity of different approaches, we define the *Set-Score* metric. The *Set-Score* of an interpretation I, is the fraction of all instances that are labelled with the correct class using I. The



(a) Adult Dataset

(b) Recidivism Dataset

Figure 2: Set-Score v/s Number of Rules for various approaches

Table 3: Comparing Global Rulesets

		Ruleset with 20 Rules				Ruleset with 90% Coverage			
Dataset	Approach	AMB	UNC	LEN	NUM	AMB	UNC	LEN	NUM
Adult	MAGIX	0.27	0.00	1.95	20	0.25	0.00	1.00	2
	Anchors	0.08	0.38	3.25	20	0.33	0.04	5.40	337
	SLS	0.97	0.00	1.91	20	0.85	0.01	1.00	2
Recidiv.	MAGIX	0.10	0.00	1.95	20	0.00	0.05	1.00	2
	Anchors	0.03	0.72	4.75	20	0.17	0.11	8.7	1101
	SLS	0.99	0.00	1.59	20	0.76	0.01	1.00	2

correct class is the class that is predicted by the trained model. If an instance is covered by several conflicting rules, the highest precision rule is retained for that instance.

Figure 2 plots the value of $Set\text{-}Score$ vs the size of the rule set for both data sets. It shows that the rules learned using MAGIX are better at explaining model behavior than those from SLS or Anchors.

4.3 Comparing Global Rule sets

In this section, we show comparisons along 4 metrics introduced in work by Lakkaraju et al. (2016). Table 3 records these comparisons.

1. Fraction Ambiguous (AMB): Fraction of instances in the data set that are covered by multiple conflicting rules.

2. Fraction Uncovered (UNC): Fraction of instances in the data set that are not covered by any of the rules in the ruleset.

3. Rule Length (LEN): Average number of conditions in a rule from the ruleset.

4. Number of Rules (NUM): Number of rules in the ruleset.

As we can see from the table, Anchors requires a very high number of rules to cover 90% of the data set. Further, the average rule learned by Anchors is also longer. MAGIX instead explains model behavior on 90% of the instances with just 2 rules. For SLS, the value of ambiguity is quite high. Such rules would convey confusing model behavior and compromise interpretability. The interpretations produced by MAGIX are concise, unambiguous and explain model behavior on a large part of the dataset.

5 CONCLUSION

We have presented a novel approach to learn rules that explain the behavior of a black box model globally. Our approach optimizes the rules on multiple dimensions of accuracy, coverage and human interpretability. This claim is supported with experiments. By interpreting models trained on large data sets, we can extract patterns inherent in the original data. This gives us a useful way to understand large and complex data sets. Moreover, rules obtained using MAGIX provide insights into model behavior and can be used to identify biases. Our approach is scalable and has been implemented and made available to marketers as part of a leading digital marketing suite of products.

REFERENCES

Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1): 307–328, 1996.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Deshpande, Adit. Perform sentiment analysis with lstms, using tensorflow. https://www.oreilly.com/learning/perform-sentiment-analysis-with-lstms-using-tensorflow.

Been Doshi-Velez, Finale; Kim. Towards a rigorous science of interpretable machine learning. In *eprint arXiv:1702.08608*, 2017.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684. ACM, 2016.

Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43, 2018.

Francisco Louzada, Anderson Ara, and Guilherme B Fernandes. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134, 2016.

Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*, 2016.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 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*, pp. 1135–1144. ACM, 2016b.

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

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153. JMLR. org, 2017.

Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

A DATASETS USED

Dataset	Rows	Features	Classes
Adult Dataset	32,561	Age, Workclass, Education, Marital Status, Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Weekly Hours, Country	<=$50K, >$50K
Recidivism Dataset	9,549	No. of priors, Age, Gender, Race, Marital Status, Severity of crime, Years of schooling, Alcoholic, Junky, Prison Violations, Months Served	Low-risk, High-risk
IMDB Movie Reviews	25,000	Movie Review in English	Positive, Negative

Table 4: Dataset Description