Guided-LIME: Structured Sampling based Hybrid Approach towards Explaining Blackbox Machine Learning Models

Amit Sangroya, Mouli Rastogi, C. Anantaram and Lovekesh Vig

TCS Innovation Labs, Tata Consultancy Services Ltd., Delhi, India

Abstract

Many approaches to explain machine learning models and interpret its results have been proposed. These include shadow model approaches, like LIME and SHAP; model inspection approaches like Grad-CAM and data-based approaches like Formal Concept Analysis (FCA). Explanations of the decisions of blackbox ML models using any one of these approaches has their limitations as the underlying model is rather complex. Running explanation model for each sample is not cost-efficient. This motivates to design a hybrid approach for evaluating interpretability of blackbox ML models. One of the major limitations of widely-used LIME explanation framework is the sampling criteria that is employed in SP-LIME algorithm for generating a global explanation of the model. In this work, we investigate a hybrid approach based on LIME using FCA for structured sampling of instances. The approach combines the benefits of using a data-based approach (FCA) and proxy model-based approach (LIME). We evaluate these models on three real-world datasets: IRIS, Heart Disease and Adult Earning dataset. We evaluate our approach based on two parameters: 1) by measuring the prominent features in the explanations, and 2) proximity of the proxy model to the original blackbox ML model. We use calibration error metric in order to measure the closeness between blackbox ML model and proxy model.

Keywords

Interpretability, Explainability, blackbox Models, Deep Neural Network, Machine Learning, Formal Concept Analysis

1. Introduction

Explainability is an important aspect for an AI system in order to increase the trustworthiness of its decision-making process. Many blackbox deep learning models are being developed and deployed for real-world use (an example is Google's Diabetic Retinopathy System [1]). For such blackbox systems neither the model details nor its training dataset is made publicly available. Explanations of the predictions made by such blackbox systems has been a great challenge.

Apart from post-hoc visualization techniques [2] (e.g., feature dependency plots), feature importance techniques based on sensitivity analysis, there have been three main approaches for explainability of AI systems: i) Proxy or Shadow model approaches like LIME, SHAP ii) Model inspection approaches like Class Activation maps (CAM), Grad-CAM, Smooth-Grad-CAM, etc. and iii) Data based approaches like Decision sets and Formal Concept Analysis [3, 4, 5, 6, 7]. Most of the research work on explainability has followed one of the above approaches [8]. However, each of these approaches have limitations in the way the explanations are generated. In the *proxy model* approach, the data corpus needs to be created by perturbing the inputs of the target blackbox model and then an interpretable shadow model is built, while in the *model inspection approach* the model architecture needs to be available for inspection to determine the activations, and in the *data-based approach* the training data needs to be available.

Local shadow models are interpretable models that are used to explain individual predictions of blackbox machine learning models. LIME (Local Interpretable Model-agnostic Explanations [9]) is a well-known approach where shadow models are trained to approximate the predictions of the underlying blackbox model. LIME focuses on training local shadow models to explain individual predictions, wherein a prediction of interest y_i of the target blackbox deep learning model $\mathcal{B}$ is considered and its related input features x_i's are perturbed within a neighborhood proximity to measure the changes in predictions. Based on a reasonable sample of such perturbations a dataset is created and a locally linear explainable model is constructed. To cover the decision-making space of the target model $\mathcal{B}$, Submodular Pick-LIME (SP-LIME) [9] generates the global explanations by finding a set of points whose explanations (generated by LIME) are varied in their selected features and their dependence on those features. SP-LIME proposes a sampling way based on sub-modular picks to select instances such that the interpretable features have higher importance.

CIKM 2020 Workshops, October 19-20, 2020, Galway, Ireland

EMAIL: amit.sangroya@tcs.com (A. Sangroya); mouli.r@tcs.com (M. Rastogi); c.anantaram@tcs.com (C. Anantaram); lovekesh.vig@tcs.com (L. Vig)
ORCID:

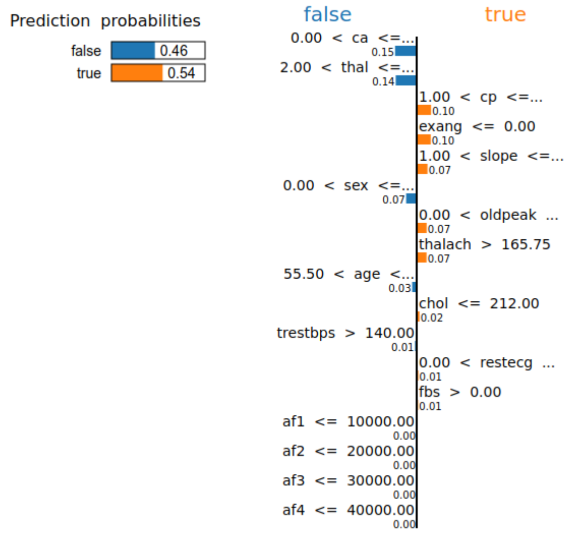


Figure 1: Example output of LIME after adding noisy features in the Heart Disease dataset

Figure 1 shows a sample explanation output of LIME for a binary classification problem on Heart Disease dataset. The prediction probabilities are shown in the left using different colors and prominent features that are important for classification decision are shown in the right. Important features are presented in a sorted manner based on their relevance. Note that some noisy features are also injected in the dataset and therefore are present in the explanation (af1, af2, af3 and af4) as well. In an ideal scenario, noisy features should not be the most relevant features for any ML model and therefore should be least important from an explanation point of view. However, due to proxy model inaccuracies and unreliability, sometimes these noisy features can also come as the most relevant features in explanations. In figure 2, we show an example scenario that compares the calibration level of two proxy models with a machine learning model. The x axis in this figure is the confidence of model and y axis is the accuracy. Assuming that we have a blackbox machine learning model and a proxy model that explains this model, we argue that these models should be closer to each other in terms of their calibration levels.

Ideally, a proxy model which is used for explaining a machine learning model should be as close as possible to the original model

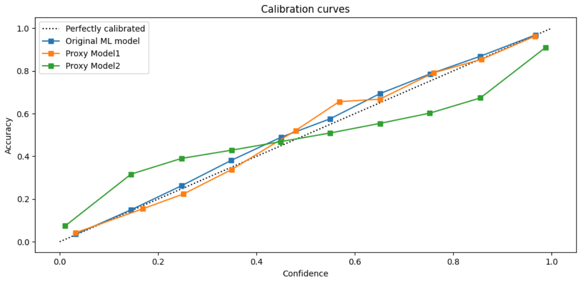


Figure 2: Example for Calibration of ML model and proxy explanation models

Motivated by the design of an optimized explanation model, we design a hybrid approach where we combine the shadow model approach proposed by LIME with the data-based approach of Formal Concept Analysis to explain the outcomes of a machine learning model. We use LIME to interpret locally by using a linear shadow model of the blackbox model, and use Formal Concept Analysis to construct a concept lattice of the training dataset, and then extract out implication rules among the features. Based on the implication rules we select relevant samples for the global instances that we feed to SP-LIME. Therefore, rather than using all instances (which is very costly for deep networks) or random sampling (which never guarantees optimal behavior), we use a FCA guided approach for selecting the instances. Therefore, we call our framework as Guided-LIME.

We show that Guided-LIME results in better coverage of the explanation space as compared to SP-LIME. Our main contributions in this paper are as follows:

- We propose a hybrid approach based on LIME and FCA for generating explanation by exploiting the structure in training data. We demonstarte how FCA helps in structured sampling of instances for generating global explanations.
- Using the structured sampling, we can choose optimal instances both in terms of quantity and quality to generate explanations and interpret the outcomes.Thereafter, using *calibration error* metric we show that Guided-LIME is a closer approximate of the original blackbox ML model.

2. Background and Preliminaries

2.1. Blackbox Model Outcome Explanation

A blackbox is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans. Given a blackbox model solving a classification problem, the blackbox outcome explanation problem consists of providing an interpretable explanation for the outcome of the blackbox. In other words, the interpretable model must return the prediction together with an explanation about the reasons for that prediction. In this context, local interpretability refers to understanding only the reasons for a specific decision. In this case, only the single prediction/decision is interpretable. On the other hand, a model may be completely interpretable when we are able to understand the global prediction behavior (different possible outcomes of various test predictions).

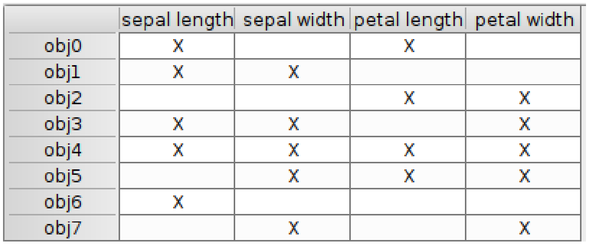
2.2. LIME Approach for Global Explanations

SP-LIME algorithm provides a global understanding of the machine learning model by explaining a set of individual instances. Ribeiro et al. [9] propose a budget B that denotes the number of explanations to be generated. Thereafter, they use *Pick Step* to select B instances for the user to inspect. The aim of this is to obtain non-redundant explanations that represent how the model behaves globally. This is done by avoiding instances with similar explanations. However, there are some limitations of this algorithm [10]:

- The SP-LIME algorithm is based on a greedy approach which does not guarantee an optimal solution.
- The algorithm runs the model on all instances to maximize the coverage function.

Data points are sampled from a Gaussian distribution, ignoring the correlation between features. This can lead to unlikely data points which can then be used to learn local explanation models. In [11], authors study the stability of the explanations given by LIME. They showed that the explanations of two very close points varied greatly in a simulated setting. This instability decreases the trust in the produced explanations. The correct definition of the neighborhood is also an unsolved problem when using LIME with tabular data. Local surrogate models e.g. LIME is a concrete and very promising implementation. But the method is still in development phase and many problems need to be solved before it can be safely applied.

2.3. Formal Concept Analysis



	sepal length	sepal width	petal length	petal width
obj0	X		X	
obj1	X	X		
obj2			X	X
obj3	X	X		X
obj4	X	X	X	X
obj5		X	X	X
obj6	X			
obj7		X		X

Figure 3: Example of a formal context using samples from IRIS dataset

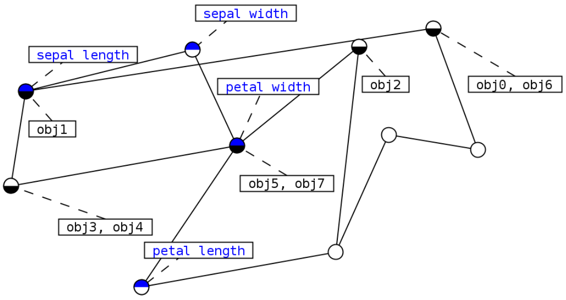


Figure 4: Example of a concept lattice related to formal context in Figure 3

Formal Concept Analysis (FCA) is a data mining model that introduces the relation among attributes in a visual form. It was introduced in the early 80s by Wille (1982) to study how objects can be hierarchically grouped together according to their common attributes. FCA deals with the formalization of concepts and has been applied in many disciplines such as software engineering, machine learning, knowledge discovery and ontology construction during the last 20-25 years. Informally, FCA studies how objects can be hierarchically grouped together with their common attributes. A formal context $K = (G, M, I)$ consists of two sets G and M and a relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes of the context. A formal concept of a formal context $K = (G, M, I)$ is a pair (A, B). The set of all formal concepts of a context K together with the order relation I forms a complete lattice, called the concept lattice of K.

Figure 3 and 4 are examples from IRIS dataset (more details in Section 4). In figure 3, we show a collection of some objects and their attributes. For simplicity, we choose only those objects where a particular attribute is present or not. In real-world objects can have very complex relationships with fuzzy values. Figure 4 is an example concept lattice generated using this sample data.

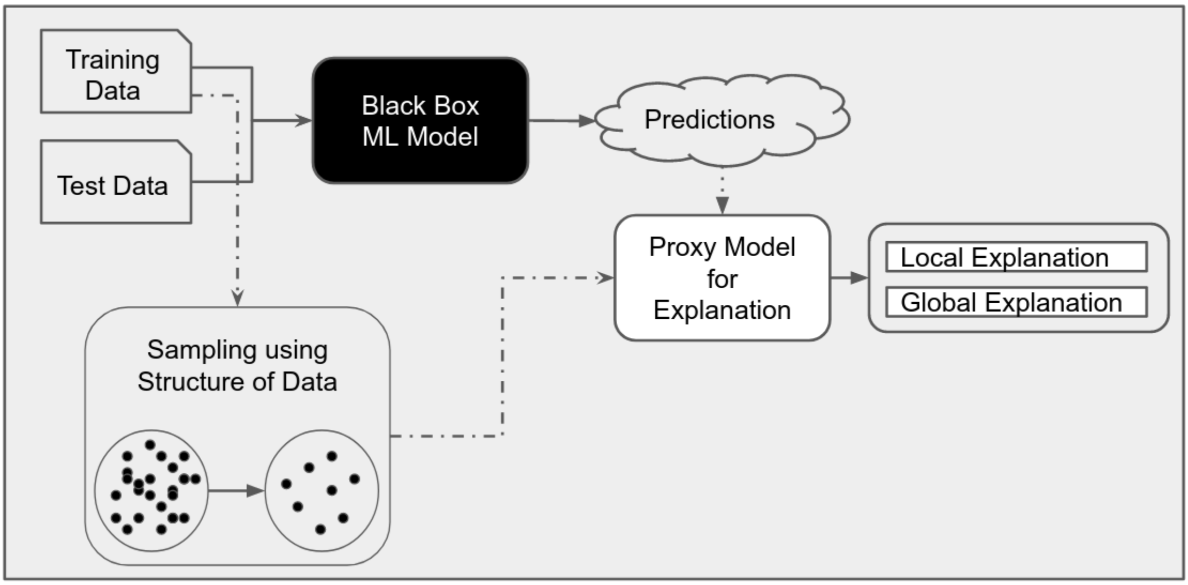


Figure 5: Overall workflow of Guided-LIME

3. Guided-LIME Framework: Guiding sampling in SP-LIME using FCA extracted concepts

In [9] SP-LIME has been used to generate global explanations of a blackbox model. SP-LIME carries out submodular picks from a set of explanations generated for a given set X of individual data instances. The SP-LIME algorithm picks out explanations based on feature importances across generated explanations. However, the data instances X from which explanations are generated, are either the full dataset (called Full-LIME) or data points sampled from a Gaussian distribution (SP-LIME random), and ignore the correlation between features in the dataset. Carrying out SP-LIME for the full dataset (Full-LIME) is very time consuming especially when the dataset is large. Carrying out SP-LIME random on the dataset may end up considering data points that are implied by other data points in the explanation space. Thus it is important to analyze the full data set and choose only those points for SP-LIME such that the selected data points are representative of the data space. In this work, we propose a mechanism to determine the implication of features to guide the selection of the instances X from the training dataset. We use Formal Concept Analysis (FCA) to analyze the training data and discover feature implication rules. Using these feature implication rules, we pick appropriate instances to feed into SP-LIME. SP-LIME then uses these instances to generate a set of local explanation models and covers the overall decision-making space. FCA provides a useful means for discovering implicational dependencies in complex data [12, 13].

In previous work, FCA-based mechanism has been used as an approach to explain the outcome of a black-box machine learning model through the construction of lattice structure of the training data and then using that lattice structure to explain the features of predictions made on test data [4]. In this proposed hybrid approach, we use the power of FCA to determine implication rules among features and using that to guide the submodular picks for LIME in order to generate local explanations. It provides the benefits of using data-based approach and proxy model based approach in a unified framework.

3.1. FCA-based selection of Instances

The goal of our FCA-based instances selection is to take advantage of the underlying structure of data to build a concise and non-redundant set of instances. We hypothesize that most of the state-of-the-art approaches do not consider this information (to the best of our knowledge). We shortlist sample instances using the following process:

1. We first binarize the training data in an ad-hoc way. The binarization technique is applied to discretize the continuous attribute values into

only of two values, 0 or 1. The binarization process can be done in a more formal manner e.g. chiMerge algorithm [14] which ensures that binarization method does not corrupt the generated lattice. In the scope of current work, we keep this process simple enough. Thereafter, we generate concept lattice using standard FCA-based approach. Each concept in the lattice represents the objects sharing some set of properties; and each sub-concept in the lattice represents a subset of the objects.

2. We use **ConExp** concept explorer tool to generate lattice from the training data [15].

3.1.1. Generating Implication Rules from Training Data

In order to find an optimal subset of samples, we generate implication rules from the given training data. One of the challenge in generating implication rules is that for a given domain and training data, the number of rules can be very large. Therefore, we shortlist rules based on their expressiveness e.g. we select the subset of rules that have the highest coverage and lowest redundancy.

When we generate association rules from the dataset, conclusion does not necessarily hold for all objects. However, it is true for some stated percentage of all objects covering the premise of rule. We sort the rules using this percentage and select the top k rules. The value of k is emperically calculated based on a given domain.

3.1.2. Generating Lattice Structure and selecting Instances

Using the lattice structure and implication rules, we select instances for guiding SP-LIME. We identify all the instances that follow the implication rules. For each rule in the “implication rules list", we calculate if a given sample “pass" or “fail" the given criteria i.e. if a particular sample s follows implication rule r or not. Finally, we produce a sorted list of the instances that are deemed more likely to cover maximally and are non-redundant as well.

3.2. Guided-LIME for Global Explanations

We propose structured data sampling based approach Guided-LIME towards a hybrid framework extending SP-LIME. SP-LIME normally has two methods for sampling: *random* and *full*. In the random approach, samples are chosen randomly using a Gaussian distribution. On the other hand, *full* approach make use of all the instances. We extend the LIME implementation to integrate another method “*FCA*" that takes the samples generated using lattice and implication rules.

Algorithm 1 explains the steps to perform structured sampling using training data and pass to SP-LIME for generating global explanations. The input to Guided-LIME is training data used to train the blackbox ML model. Data processing for finding the best samples for Guided-LIME involves binarization of data. Thereafter, a concept lattice is created based on FCA approach [4]. Using the concept lattice, we derive implication rules. These rules are then used to select test instances for Guided-LIME.

Algorithm 1 Sample selection algorithm using FCA for Guided-LIME

```
Require: Training dataset $D$
Ensure: Samples and their ranking
  for a given Training dataset $D$ consisting of data samples $s$ do
    Binarize numeric features
    Generate concept Lattice using FCA
    Find implication rules $r$
    Generate samples and their ranking
    Select top $k$ samples from each rule
  end for
  for all top $k$ samples from each rule do
    Select samples using redundancy and coverage criteria
  end for
```

As we mentioned previously, there are various examples of using a single approach for explanation. This can be done using any of the proposed techniques i.e. proxy model, activation based or perturbation based approach. However, we argue that none of these approaches provides a holistic view in terms of outcome explanation. Whereas, if we use a hybrid approach such as a combination of proxy model and data-based approach, it can provide a better explanation at a much reduced cost.

One of the question that arise in our hybrid approach is whether the approach is still model agnostic such as LIME. We argue that sampling step do not affect the model agnosticity in any manner. It just adds a sampling step which helps in choosing the samples in a systematic manner.

Dataset	Classes	# of instances	# of features	Features
IRIS	3	150	4	sepal length, sepal width, petal length, petal width
Heart Disease	2	303	14	age of patient, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, peak exercise ST segment, number of major vessels colored by fluoroscopy, Thal, Diagnosis of heart disease
Adult Earning	2	30000	14	age, workclass, fnlwgt, education, education-num, marital status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country

Table 1
Summary of Datasets

4. Experiments and Results

4.1. Experimental Setup

We use the following publicly available datasets to evaluate the proposed framework: IRIS, Heart Disease and Adult Earning dataset (See Table 1). IRIS dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant [16]. There are a total of 150 samples with 5 attributes each: sepal length, sepal width, petal length, petal width, class (Iris Setosa, Iris Versicolor, Iris Virginica). Similarly, Heart Disease dataset contains 14 attributes; 303 samples and two classes [17]. Adult Earning dataset contains 48000 samples, 14 features across two classes. The machine learning task for all three datasets is classification. We use random forest blackbox machine learning model in all our experiments.

4.2. Results

The goal of this experiment is to compare the proposed Guided-LIME approach with random sampling of SP-LIME. In the scope of this work, we do not compare the proposed hybrid approach with full sampling of SP-LIME. We perform a case study to find out which approach is better in selecting important features for a given blackbox model. As shown in Table 1, we maintain ground truth oracle of important features as domain knowledge [18, 19]. We train random forest classifier with default parameters of scikit-learn. In this experiment, we add 25% artificially “noisy” features in the training data. The value of these features is chosen randomly. In order to evaluate the effectiveness of approach we use FDR (false discovery rate) metric which is defined as the total number of noisy features selected as important features in the explanation.

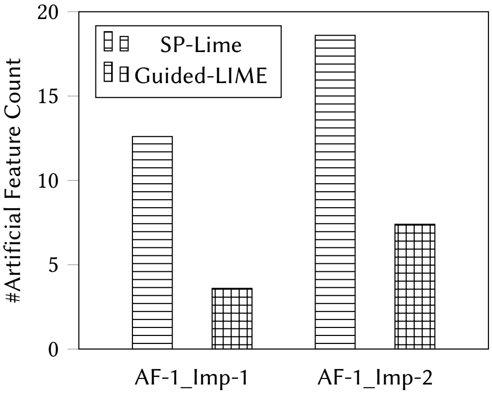


Figure 6: FDR (False discovery rate) for IRIS dataset

We calculate the occurrence of noisy features in the generated explanations. Ideally, the noisy features should not occur among the important features. Therefore a lower FDR suggests a better approach for explanation. We present the discovery of number of noisy features for each explanation averaged over 100 runs. Each explanation consists of a feature importance vector that shows the importance of a particular feature. As we see in Figures 6, 7, and 8, y axis is the number of noisy features and x axis is index of noisy feature. We include the cases where a noisy feature is at first or second place in the feature importance vector. AF-1_Imp-1 represents artificial/noisy feature occurring at first place in feature importance vector whereas AF-1_Imp-2 represents artificial/noisy feature occurring at second place. Guided-LIME sampling approach is consistently better than basic SP-LIME.

Table 2
Expected Calibration Error of Blackbox Model and proxy models

Datasets	With artificial features			Without artificial features			
	blackbox	**Guided-LIME**	**SP-LIME**	**blackbox**	**Guided-LIME**	**SP-LIME**	**Full-LIME**
Adult Earning	0.061	0.065	0.041	0.056	0.065	0.041	0.059
Heart Disease	0.149	0.167	0.216	0.125	0.165	0.169	0.136
IRIS	0.106	0.042	0.033	0.038	0.006	0.08	0.031

Table 3
Maximum Calibration Error of Blackbox Model and proxy models

Datasets	With artificial features			Without artificial features			
	blackbox	**Guided-LIME**	**SP-LIME**	**blackbox**	**Guided-LIME**	**SP-LIME**	**Full-LIME**
Adult Earning	0.187	0.372	0.428	0.19	0.353	0.219	0.344
Heart Disease	0.428	0.485	0.326	0.681	0.546	0.475	0.297
IRIS	0.307	0.311	0.178	0.134	0.009	0.406	0.408

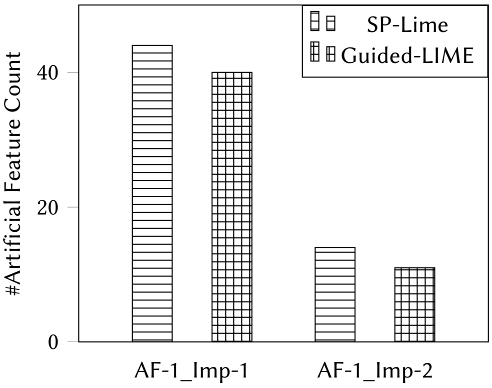


Figure 7: FDR (False discovery rate) for Heart disease dataset

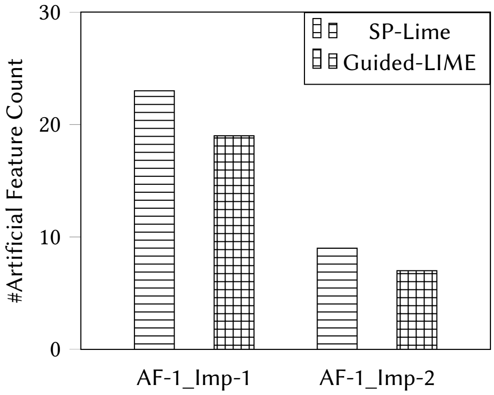


Figure 8: FDR (False discovery rate) for Adult earning dataset

4.3. Validating Guided-LIME using calibration level

The objective of this experiment is to validate which proxy model is a closer approximation to original blackbox model with respect to the prediction probabilities of each model. In order to measure this closeness, various distance metric can be used e.g. KL divergence, cross entropy etc. We use the well established ECE (*expected calibration error*) and MCE (*maximum calibration error*) as the underlying metric to detect the calibration of both the models [20]. Calibration error provide a better estimate of reliability of ML models [21, 22]. Moreover, the focus of our experiment is to estimate the proximity of the shadow model w.r.t the original blackbox model. Calibration error values are therefore used to compare which model is the better approximation of the original model. We hypothesize that the proxy model with a ECE closer to the original blackbox ML model shall be a closer approximate.

We perform experiment in two settings: 1) with original data 2) by adding noisy features in the data. As shown in Tables 2 and 3, in both scenarios, ECE and

MCE of Guided-LIME is closer to the original ML model in comparison to the random SP-LIME. This justifies the benefit of structured sampling. We also run experiments with full samples of LIME. Although, this can be a better approximate of original model, but taking all the samples in the proxy model is not a practical and economic choice for real world huge datasets. Guided-LIME has a closer ECE to the original blackbox model. Hence, Guided-LIME is a better choice as a proxy model to explain the original ML model.

5. Related Work

Various approaches for explainability of blackbox models have been proposed [8]. Broadly the existing techniques can be classified into Model Explanation approaches; outcome Explanation approaches; Model Inspection approaches. There are also example of works that focus on designing transparent design of models.

In this work, we focus only on the outcome explanation approaches. In the category of outcome explanation, CAM, Grad-CAM, Smooth Grad-CAM++, SHAP, DeepLIFT, LRP and LIME are the main approaches [23, 24, 25, 9, 26, 27, 28]. These methods provide a locally interpretable shadow model which is able to explain the prediction of the blackbox in understandable terms for humans.

Most popular shadow model approaches for blackbox ML model explanations are Local Interpretable Model-Agnostic Explanations (LIME) and SHAP. LIME can explain the predictions of any classifier in "an interpretable and faithful manner, by learning an interpretable model locally around the prediction. In order to make the predictions easily interpretable, LIME have two design goals: **Easy to interpret** and **Local fidelity**: This means that outcomes of shadow model are easily interpretable and the explanation for individual predictions are locally faithful, i.e. it correspond to how the model behaves in the vicinity of the individual observation being predicted.

In contrast, SHAP (SHapley Additive exPlanations) is distinctly built on the Shapley value. The Shapley value is the average of the marginal contributions across all permutations. The Shapley values consider all possible permutations, thus SHAP is a united approach that provides global and local consistency and interpretability. However, its cost is time — it has to compute all permutations in order to give the results. SHAP approach has speed limitations as it has to compute all permutations globally to get local accuracy whereas LIME perturbs data around an individual prediction to build a model. For generating a global explanation, SHAP need to run for every instance. This generates a matrix of Shapley values which has one row per data instance and one column per feature. We can interpret the entire model by analyzing the Shapley values in this matrix.

In CAM and Grad-CAM approaches, explanation is provided by using a Saliency Mask (SM), i.e. a subset of the original record which is mainly responsible for the prediction. For example, as salient mask we can consider the part of an image or a sentence in a text. A saliency image summarizes where a DNN looks into an image for recognizing their predictions. Although these solutions are not just limited/agnostic to blackbox NN, but it requires specific architectural modifications.

Feature importance is well known approach to explain blackbox models. More recently, instance-wise feature selection methods are proposed to extract a subset of features that are most informative for each given example in deep learning network. [29]. In [30] authors make use of a combination of neural networks to identify prominent features that impact the model accuracy. These approaches are based on subset sampling through back-propagation.

Ribeiro et. al. [9] present the Local Interpretable Model-agnostic Explanations (LIME) approach which does not depend on the type of data, nor on the type of blackbox b to be opened. In other words, LIME can return an understandable explanation for the prediction obtained by any blackbox. The main intuition of LIME is that the explanation may be derived locally from the records generated randomly in the neighborhood of the record to be explained. As blackbox the following classifiers are tested: decision trees, logistic regression, nearest neighbors, SVM and random forest.

In [31], authors find the global importance introduced by Local Interpretable Model-agnostic Explanations (LIME) unreliable and present approach based on global aggregations of local explanations with the objective to provide insights in a model's global decision making process. This work reveal that the choice of aggregation matters regarding the ability to gain reliable and useful global insights on a blackbox model. We find this work as motivation to propose an hybrid approach where aggregations can be generated using knowledge of data through FCA-based system.

In contrast to model explanation approaches such as LIME and SHAP [9, 26], our approach is complementary which can guide these approaches for selecting the optimal instances for explanation. Extracting rules from neural networks is also a well studied problem [32]. These approaches depend on various factors such as: Quality of the rules extracted; Algorithmic complexity; Expressive power of the extracted rules; Portability of the rule extraction technique etc. Our approach also uses the knowledge of structure in data however it is not dependent on the blackbox model. Moreover, formal concept analysis based data analysis provides a solid theoretical basis.

6. Conclusions and Future Work

In this paper,we proposed a hybrid approach for evaluating interpretability of blackbox ML systems. Although Guided-LIME do not guarantee an optimal solution, yet we observe that a single approach like LIME is not sufficient to explain the AI system thoroughly. There are limitations of deciding an optimal sampling criteria in SP-Lime algorithm. Our approach combines the benefits of using a data-based approach (FCA) and proxy model based approach (LIME). Overall, our approach is complementary to SP-LIME as we provided a structured way of selecting right instances for global explanations. Our results on real world datasets shows that false discovery rate is much lower with Guided-LIME in comparison to random SP-LIME. Moreover, Guided-LIME has a closer ECE and MCE to the original blackbox model. In future, we would like to perform extensive experiments with diverse datasets and complex deep learning models.

References

[1] M. Miliard, Google, verily using ai to screen for diabetic retinopathy in india, 2019. URL: https://www.healthcareitnews.com/news/google-verily-using-ai-screen-diabetic-retinopathy-india.

[2] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, S. Behnke, Interpretable and fine-grained visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] H. Lakkaraju, S. H. Bach, J. 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, Association for Computing Machinery, New York, NY, USA, 2016.

[4] A. Sangroya, C. Anantaram, M. Rawat, M. Rastogi, Using formal concept analysis to explain black box deep learning classification models, in: Proceedings of the 7th International Workshop "What can FCA do for Artificial Intelligence"? co-located with IJCAI 2019, Macao, China, ????

[5] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, K. Saenko, Black-box explanation of object detectors via saliency maps, 2020. arXiv:2006.03204.

[6] J. Pfau, A. T. Young, M. L. Wei, M. J. Keiser, Global saliency: Aggregating saliency maps to assess dataset artefact bias, 2019. arXiv:1910.07604.

[7] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, K. Sycara, Transparency and explanation in deep reinforcement learning neural networks, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 144–150. URL: https://doi.org/10.1145/3278721.3278776. doi:10.1145/3278721.3278776.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.

[10] C. Molnar, Interpretable Machine Learning, 2019. https://christophm.github.io/interpretable-ml-book/.

[11] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods (2018). URL: http://arxiv.org/abs/1806.08049, cite arxiv:1806.08049Comment: presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.

[12] S. O. Kuznetsov, Fitting pattern structures to knowledge discovery in big data, in: P. Cellier, F. Distel, B. Ganter (Eds.), Formal Concept Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 254–266.

[13] X. Benavent, A. Castellanos, E. de Ves, A. García-Serrano, J. Cigarrán, Fca-based knowledge representation and local generalized linear models to address relevance and diversity in diverse social images, Future Generation Computer Systems 100 (2019) 250 – 265. URL: http://www.sciencedirect.com/science/article/pii/S0167739X18307271. doi:https://doi.

org/10.1016/j.future.2019.05.029.

[14] R. Kerber, Chimerge: Discretization of numeric attributes, in: Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI'92, AAAI Press, 1992, p. 123–128.

[15] S. Yevtushenko, J. Tane, T. B. Kaiser, S. Obiedkov, J. Hereth, H. Reppe, Conexp - the concept explorer (2000-2006). URL: http://conexp.sourceforge.net.

[16] R. Fisher, Iris data set, 2019. URL: https://archive.ics.uci.edu/ml/datasets/iris.

[17] D. W. Aha, Heart disease data set, 2019. URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[18] R. El-Bialy, M. A. Salamay, O. H. Karam, M. E. Khalifa, Feature analysis of coronary artery heart disease data sets, Procedia Computer Science 65 (2015) 459 – 468. URL: http://www.sciencedirect.com/science/article/pii/S1877050915029622. doi:https://doi.org/10.1016/j.procs.2015.09.132, international Conference on Communications, management, and Information technology (ICCMIT'2015).

[19] D. D. Sarkar, Hands-on machine learning model interpretation, 2018. URL: https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608.

[20] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 1321–1330.

[21] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, D. Tran, Measuring calibration in deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[22] V. Kuleshov, N. Fenner, S. Ermon, Accurate uncertainties for deep learning using calibrated regression, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 2796–2804. URL: http://proceedings.mlr.press/v80/kuleshov18a.html.

[23] B. Zhou, A. Khosla, L. A., A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization., CVPR (2016).

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: The IEEE International Conference on Computer Vision (ICCV), 2017.

[25] D. Omeiza, S. Speakman, C. Cintas, K. Weldermariam, Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models, 2019. arXiv:1908.01224.

[26] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[27] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, 2017. arXiv:1704.02685.

[28] A. Binder, G. Montavon, S. Bach, K.-R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, 2016. arXiv:1604.00825.

[29] J. Chen, L. Song, M. J. Wainwright, M. I. Jordan, Learning to explain: An information-theoretic perspective on model interpretation, 2018. arXiv:1802.07814.

[30] J. Yoon, J. Jordon, M. van der Schaar, INVASE: instance-wise variable selection using neural networks, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=BJg_roAcK7.

[31] I. van der Linden, H. Haned, E. Kanoulas, Global aggregations of local explanations for black box models, ArXiv abs/1907.03039 (2019).

[32] R. Andrews, J. Diederich, A. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems 6 (1995) 373–389. doi:10.1016/0950-7051(96)81920-4.