Explain variable influence in black box models through pattern mining

Xiaoqi Ma
xiaoqi.ma@rwth-aachen.de
Matriculation number: 383420

Supervisor: Prof. Dr. Markus Strohmaier
Second Examiner: Prof. Dr. Bastian Leibe
Advisor: Dr. Florian Lemmerich

Chair of Computational Social Sciences and Humanities
RWTH AachenFaculty of Mathematics, Computer Science and Natural Sciences
RWTH Aachen University

This thesis is submitted for the degree of
M.Sc. Media Informatics

Aachen, Germany
December 18, 2019

Abstract

Understanding the decisions made by machine learning models is crucial for decision-makers and end-users. Enforced by GDPR, "right to explanation" demands businesses to provide understandable justifications to their users. Thus, it is of paramount importance to elucidate the model decision, which could be measured by model interpretability, the degree to which a human can understand the cause of a decision. In order to interpret black-box models, model-agnostic approaches could be applied, which provide flexibility in the choice of models, explanations and representation for models. From global interpretability viewpoint, feature importance and global surrogate are going to be explored. We also investigate the local model-agnostic methods, like LIME and Shapley value. After obtaining the designated feature contribution for each instance, we could use the subgroup discovery technique to figure out "interesting" patterns to provide more elaborate explanations. In this thesis, the aim is to build up a python package to provide a collection of tools to explain variable influence in black-box models through subgroup discovery.

Keywords: *Black box model interpretability; Model agnostic; Subgroup discovery*

e

Contents

Chapter 1

Introduction

Assume we had a scenario where an unemployed young wanted to get a loan from the bank to start a business. The bank manager served him politely and then entered his information into the banking system, but unfortunately, the system rejected his request. Of course, the young man desired to know why his loan was rejected, and the bank was obligated to provide justifications. Hopefully, the bank manager would receive some explanations from the system telling that the bank was not willing to take risks offering a loan because of his unemployment and no assert mortgage. However, the manager only obtained the final decision made by the banking algorithms rather than the explanations for this decision. If so, how can we trust banking algorithms without explanations and how can we ensure there are no mistakes while making the decisions since the model is a complete "black box" to customers. And that is why we would like to explore variable influence to assist decision-makers to explain model decisions.

In this chapter, I will provide an overview of how to interpret the results generated by black box models, how this topic caught our attention, what is the purpose of this master thesis and finally how to construct the model interpretation framework. This chapter shall give readers a complete plan for this thesis.

1.1 Background

Machine learning is a set of methods that are used to teach computers to perform different tasks without hard-coding instructions. Over the last decades, Machine learning area has gone through unprecedented growth. Due to the boosting computational power and the availability of "Big data", a myriad of classification or regression tasks could be solved by applying machine learning algorithms. For a simple classification scenario, like predicting the house prices based on the historical data, a traditional regression model might be adequate, like logistic regression. However, for tackling complex problems like language translation, more complicated models are required, such as neural networks.

When evaluating machine learning models, people have a tendency to focus more on

the performance by observing metrics like accuracy, precision, recall and etc., which are of course very fundamental to assess the model. Nevertheless, they neglect the importance of interpretability to the model, which shows "the degree for a human to understand model decisions and the ability to consistently predict the results"[1]. For those models with high interpretability, it is rather clear for human to understand the decisions, while for those are not easily interpretable by human, we should utilize interpretation methods to facilitate us to explain the outcomes. Therefore, to have a better understanding of the decisions made by the model, it is of paramount importance to investigate model interpretability.

From model interpretability perspective, models could be classified as white box models and black box models. Roughly to say, white box models have simple structures, a limited number of coefficients and can be understood by human, such as linear regression, logistic regression, and decision trees, since the prediction results could be interpreted by exploring into the model parameters. On the contrary, black box models usually have more complex structures and a substantial number of parameters which are not understandable. For example, ensemble models or neural networks could be regarded as black box models for the reason that decisions made by black box models cannot be understood by looking at their parameters, which is a major disadvantage for complex models. Typically, those complicated models could achieve better performance for the sake of less interpretability. However, proper interpretability is crucial to explain the choice made by the model and especially important for decision-makers. Besides, "right to explanation" meaning the right to be given an explanation for an output of automated algorithms was stated by General Data Protection Regulation(GDPR), which requires businesses to provide understandable justifications to their users [2], just like the scenario formerly described that the bank manager should be able to clarify the reason of loan rejection.

Except for the models with interpretable parameters that could be used directly to explain the decision, more model explanation methods should be explored to support the explanation for black box models as well. As defined in [3], an explanation lies the connections between the input feature values and its model output in a human-understandable way. Generally, two broad genres of explanation methods are often mentioned, one is the global interpretation methods and the other refers to local interpretation methods. As the name suggests, the global interpretation focus on the global view of the input variables, more specifically, it points out the most significant features that can affect predictions of the entire data set. After exploring the importance ranking of input features, we could at least obtain a general overview of variable influence. In contrast, our attentions are more inclined to local interpretation methods, which are more compatible with the scenario previously described, targeting at the instance level explanation. In this case, each instance should be supplied with a corresponding explanation specifying the cause to prediction results.

Though the above-mentioned methods could indeed give explanations to some degree, the global methods give a too broad interpretation view while local interpretation may become too sensitive to reveal the underlying cause due to the particularity of that instance, causing either inaccurate or unreliable explanations. Thus, explor-

ing the patterns of explanations could be a good amendment that comes into our mind, which aims to discover subgroups which share interchangeable explanations by applying pattern mining technique.

In this thesis, the huge effort would be made to investigate this novel technique that combines model interpretation methods and pattern mining technique.

1.1.1 Purpose of this thesis

Given the urgent need to obtain decent justifications for every decision made by the algorithms as well as the enforcement by GDPR, a feasible approach is to build a framework to provide explanations for each model regardless of model types. Undoubtedly, many interpretation methods have already come to the surface to facilitate model explanation, but they are merely limited to a single instance explanation, which might be unreliable and causing misunderstanding due to the excessive interpretation of that specific instance. Therefore, it is wiser to not only study the instance explanation but also investigate the groups of instances whose predictions are given similar explanations using subgroup discovery technique. And we noticed that we could generate more robust explanations by combining those two approaches.

Therefore, in this thesis, I would like to construct a robust framework combing the benefits of model interpretation methods and pattern mining technique to furnish us with good model interpretation.

1.1.2 Thesis Structure

The remainder of this thesis is structured as following.

Chapter 2 focuses on previous work on related fields, such as Model Interpretation methods and Subgroup Discovery field. In particular, it is dedicated to review the existing global interpretation methods and local interpretation methods. In addition, the fundamentals of subgroup discovery are discussed as well like the selection of interestingness measure.

Chapter 3 is concerned with the theoretical knowledge of local interpretation methods to explain black box models. It begins with simple approaches on specific scenarios, for example, to inspect the influence of binary feature using the binary flip approach. Then the methods becomes more general which can be applied on any type of features, like Shapley values. Finally, more attention is laid on the novel technique which combines the approach of model agnostic local interpretation and subgroup discovery.

Chapter 4 presents the detailed description of datasets that are collected and the set up of experiments. In experiments, the comparison of several local interpretation methods is covered. Additionally, we demonstrate case studies on specific datasets.

Finally, Chapter 5 concludes the work with a summary of results and ideas on future

work.

Chapter 2

Related Work

In this chapter, previous works in areas that are related to the topic of this thesis will be covered. Firstly, we will introduce the model interpretation methods, which facilitate us to interpret predictions from black box models. Two groups of methods, including global interpretation methods and local interpretation methods are explained respectively. The next part will investigate the decision trees, which could provide us a human-understandable explanation to the prediction. And the last part will give an overview of subgroup discovery technique, focusing on the target concept and interestingness measure.

2.1 Related works

2.1.1 Model interpretation methods

Recently, studies trying to understand why a model makes a certain prediction have been intensively investigated. The extent to which the model or its prediction are human-understandable is termed as Interpretability(comprehensibility), whose definition may be found in paper [1]. Various criteria can be used to classify types for machine learning interpretability. Intrinsic interpretability, for example, is one type of the interpretability, which refers to models that are intrinsic interpretable owing to their simple structures, such as linear models or decision trees. In contrast, post hoc interpretability is meant to analyze the model interpretability after model training. Particularly, post hoc interpretability is mainly considered. And to address the measurement of model interpretability, the complexity of the predictive model in terms of the model size is a determining factor as referenced from [4]. Evidently, it is much easier to explain model predictions when the predictive model has high comprehensibility, which are called interpretable models. Usually, the interpretable model could help us to directly explain the predictions with its parameter, and the commonly used models are liner regression, logistic regression and the decision trees. However, black box models cannot be interpreted through its inner parameters due to the low interpretability. Therefore, considerable research efforts have been devoted to model interpretation methods which could provide decent explanations

that are understandable to experts. Generally, two broad genres of explanation methods are often mentioned, one is the global interpretation methods and the other refers to local interpretation methods.

Global interpretation

The global interpretation method focus on the global view of the input variables, more specifically, it points out the most significant features that can affect predictions of the entire data set. As mentioned in [5], the Partial Dependence Plot (PDP) is a global interpretation method which shows the marginal effect of a feature on the predictions of the machine learning model. It can show the relationship between the selected feature and the target by adapting the values of the selected feature, and specifically to characterize the feature influence on model predictions.

Another popular approach is feature importance. There are many methods for assessment of feature importance. The default feature importance mechanism was proposed and implemented by the inventor of RandomForest algorithm, which was to add up the gini decreases for each individual variable over all trees in the forest and get the average. However, pointed by Strobl et al [6], this method was biased and was not reliable in scenarios that selected variable was biased in terms of the scale of measurement. Later, a novel strategy called permutation importance was described [7]. In this approach, the feature importance is estimated by the drop of prediction accuracy of the model after permuting the selected feature. A feature is regarded as "important" if prediction accuracy drops a lot after shuffling feature values as the model depends on the feature for the prediction. Conversely, a feature is "unimportant" if the accuracy is slightly dropped, which means the feature is hardly relied on for the model.

Local Interpretation

Local interpretation methods aims at the instance level explanation which means each instance should be supplied with a corresponding explanation identifying the cause to prediction. Following this idea, it leads us to the local surrogate methods, which are able to explain individual predictions of any black box models in a faithful way. As a concrete implementation of local surrogate models, Local interpretable model-agnostic explanations (LIME) was initially proposed in paper [22]. Another possible approach is to calculate the individual contribution of each feature in an instance to compose the final prediction as described in paper [25]. Inspired by this idea and the theoretical knowledge from the coalitional game theory, Shapely value was proposed to explain instance-level predictions with contributions of each feature values [26]. However, in this approach, the explanation for the prediction of a black box model is just a simple value, rather than an explanation model like LIME, which fails to make judgments about the connections between input change and prediction change. To address those problems, Lundberg and Lee [28] proposed a unified framework for explaining predictions, which is based on the Shapley value, and they named it SHAP(SHapley Additive exPlanations). This novel approach

unified existing explanation methods and brought more clarity to the methods space. They introduced the explanation model by treating the explanation of a individual prediction as a model.

2.1.2 Decision Trees

2.1.3 Subgroup discovery overview

Recent developments of the research filed in knowledge discovery in databases have attracted much attention, where a diverse of methods are proposed to extract local patterns from large volumes of data [8]. Apart from the methods for mining local patterns such as discriminative patterns [9] and emerging patterns [10], subgroup discovery (also called pattern mining) is established as a supervised and descriptive data mining technique. As defined in [11], in the subgroup discovery task, assuming we have a population of individuals and the corresponding property of interest, it aims to discover subgroups that are statistically "most interesting". Specifically, the interesting subgroups have the most unusual distributional characteristics with respect to certain property of interest given by the target variable [12].

In a formal definition, the fundamental concepts of subgroup discovery task could be summarized by a quadruple (D, Σ, T, Q) [13]. In the quadruple, D represents the dataset, which is formed by a group of instances. Σ means the search space, consisting of a set of selection expressions, and the search space covers all the patterns that are traversed through. T implies the target concepts being exploited in pattern mining task. Commonly, a single target concept, e.g. binary or numeric, is applied in the task, nevertheless, multi-target concepts are also enabled given by the exceptional model mining framework [14]. With respect to the quality measure criteria, symbolled as Q, it is specified depending on the target concept.

Considerable research efforts have been devoted to binary target concept. Normally the quality measure for binary target is relied on the parameters contained in a contingency table, which describes the distribution of positive/negative instances for the observed pattern and its complement, respectively. In paper [15], Kloesgen proposed a prevalent family of quality measure, relating to the size of the subgroup and the difference between the share of the target concept in the subgroup and the general population. Correspondingly, several approaches to measure the quality of numeric attributes have been proposed, and a listing of interestingness measures for numeric target concepts could be found in [16]. Since numeric attribute has certain characteristics, such as mean value or median value, therefore, the quality measure for numeric target could be formalized by slightly adapting the quality function for binary targets. In specific, the share of target in the subgroup and in the entire population could be replaced by the characteristic of the target. Generally, there are five categories of interestingness measure for numeric target, concluded in [13], which are mean-based measures, median-based measures, variance-based measures, distribution-based measures, and rank-based measures. Furthermore, as for multi-target concept, the quality function has been described in a number of studies. And a general framework for multi-target quality functions is given by exceptional model

mining [14], proposing a variety of model classes, which contains the correlation model, the regression model and the classification model.

In the subgroup discovery task, unlike the choice of applied quality measure which is mainly determined by the target concept, the mining algorithms are almost equivalent. And for a specific algorithm, the three algorithmic components should be defined, which are enumeration strategy, data structure and pruning strategy. Various enumeration strategies could be used, e.g. exhaustive methods, seeking to acquire the optimal subgroup by traversing through the whole search space. In contrast, heuristic approaches, normally a beam search strategy [17], is often used for subgroup discovery due to its efficiency, which aims to find interesting patterns but not necessarily the optimal patterns in a short time. Normally, data is stored in horizontal layout, e.g. tabular-formatted database. Instead, vertical data representations can also be used, which is covered in paper [18]. In addition, FP-tree structure is also applicable referring to the wide-spread FP-Growth algorithms [19]. Furthermore, considering the efficiency of algorithms, the pruning strategies if of critical importance. To determine the upper quality bounds and safely prune parts of the search space, optimistic estimates could be applied as proposed in [20] for binary target concept. In addition, to shrink the search space of subgroup discovery task, minimal support pruning strategy is useful by exploiting anti-monotone constraints.

Chapter 3

Approach

In this chapter, the details of approaches will be discussed. It starts with an overview of local interpretation methods and several variants of them are introduced respectively. Firstly, we present the binary feature flip idea which aims to characterize the impact of binary features by flipping the feature values. After that, we are interested in the effect of numeric features in the model by inspecting the outcome change of the model when the numeric feature values are perturbed to generate noises. Despite the "variable-specific" methods, we also focus on local interpretable model-agnostic explanations (LIME) which is able to explain individual predictions for any types of features and models. However, no theory can support why LIME can fit linear behavior locally on black box models. Therefore, we continue exploring Shapley value, which is a reasonable explanation method with well founded theory. In addition, the appealing approach assigns a contribution score for each feature value to smooth the path of interpreting the final prediction of individual instances by calculating the Shapley value.

Then the following section describes the novel technique which combines the local interpretation methods and pattern mining technique. Since the target concept during subgroup discovery in our situation is either prediction change or feature influence score, therefore, the focus shall attribute to numeric target. Later, the standard approaches to measure the interestingness of subgroups are discussed. Furthermore, methods to avoid redundancy in subgroups are explored.

3.1 Local interpretation methods

In comparison to Global interpretable methods which are dedicated to explain the global model output by comprehending the entire datasets, it is more interesting to examine the model prediction for an individual instance. Besides, it could be observed that the global interpretation methods are less sensitive to noises if we make some perturbations on feature values, however, it could lead to tremendous changes in the prediction for an instance. Therefore, the local explanations shall preserve high accuracy than global explanations. In the following, few local interpretation methods will be covered in detail.

3.1.1 Binary feature value flip

Binary feature implies that the feature only contains two unique values. In another words, if it is encoded as discrete numeric number, the feature value should be either 1 or 0. Thus, to flip binary feature value means to convert from 1 to 0 or the other way around. In practice, we could also use XOR operation to map from 1 to 0. For instance, gender is regarded as a binary feature which only holds value "male" and "female".

As mentioned previously, the assumption is that we hold the dataset and the corresponding model trained on that dataset. Initially, we could obtain the prediction from the model for a specific instance. Then, a binary value is flipped on a chosen feature and afterwards a new prediction is generated by applying the model to the modified instance. Therefore, as a simple measurement, the effect of this binary feature could be estimated by the difference of two outputs.

In practice, there are two variants to assess the variable influence. One way is to calculate the absolute difference of two predictions, and in this way we could ignore the bias of this binary feature on the original dataset. Literally to say, the binary feature is more influential when the difference becomes larger. In contrast, we could compute the difference for a defined direction, for example, we just care about the effect of gender changing from male to female. In this case, not only the magnitude of the effect is obtained, but also the positive or negative sign towards the prediction.

3.1.2 Numeric feature value perturbation

As the name suggests, this technique is applicable on features whose type is numeric. The idea is that we could apply binary operations to the input values to produce new values, which serves as injecting noises into the original dataset. In particular, only addition and subtraction are considered in this situation. For example, an instance includes a numeric feature called "age" and we could perturb this feature value by increasing or decreasing by a certain value to obtain the modified value.

The procedure of measuring the effect of a chosen input feature is similar to that in binary feature value flip approach. For classification or regression tasks, we could make predictions with the existing model on the instance we desire to explain. Afterwards, a new prediction is made on the adapted instance which is produced through perturbation on the selected numeric feature. And the impact of this numeric feature could be approximately evaluated by the absolute difference of two output predictions, which indicates that this particular feature plays an important role in this instance, causing unstable predictions. Roughly to say, larger prediction differences might imply the feature has stronger effect on the corresponding instance.

3.1.3 Local Surrogate: LIME

Various criteria can be used to classify types for machine learning interpretability. Intrinsic interpretability, for example, is one type of the interpretability methods, which refers to models that are intrinsic interpretable owing to their simple structures, such as linear models or decision trees. In contrast, post hoc interpretability is meant to analyze the model interpretability after model training. As introduced earlier, permutation feature importance is a post hoc interpretation method.

In this thesis, we would like to focus on post hoc interpretability, which indicates to explain model decisions after model has been trained. In particular, model agnostic interpretation methods, which extracts post hoc explanations by treating the original model as a black box, is highly valued. The model agnostic interpretation method is pretty flexible in terms of models, and it can work with any type of machine learning models, which provides a great advantage over model specific methods [21]. The principle behind is to learn an interpretable model on the decisions of the black box model and in return apply the interpretable model to those predictions that are expected to explain.

Following this idea, it leads us to the local surrogate methods, which are able to explain individual predictions of any black box models in a faithful way. As a concrete implementation of local surrogate models, Local interpretable model-agnostic explanations (LIME) was initially proposed in paper [22].

The key point behind LIME is pretty straightforward. It is intended to explain individual explanations by fitting a simple interpretable model to locally approximate the underlying black box model. The typical choice of interpretable model could be regularized linear models like Lasso or decision trees. To elaborate more intuition for LIME, the toy example is shown in 3.1. This is a binary classification task and the regions colored with blue or pink are regarded as two distinct decisions. Evidently, this decision function can not be easily interpreted by a linear model. As a clarification, we are interested in the individual instance explanation, which is marked with bold red cross. In order to fit a local interpretable model, some artificial points are created by perturbing the original data point. The learned local model, marked by the dashed line, could in principle provide a faithful explanation for the target instance.

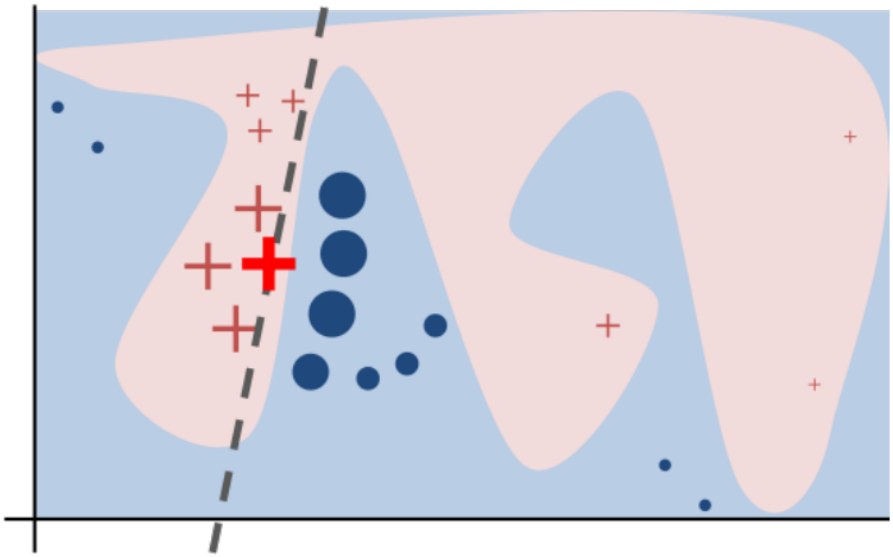

Figure 3.1: Binary classification task for a black box model. Decision regions are colored with blue or pink background. Instance to be explained are marked in bold red cross. Artificial points, marked as crosses and circles, are created by perturbing the instance of interest, whose size are weighted by the proximity to the instance. The dashed line expresses the fitted local interpretable model which could give faithful explanations.

Apart from the intuition, we could argue for the faithfulness from mathematical perspective and the constraint of LIME could be represented as equation 3.1.

$$\xi = \arg\min_{g \in G} L\left(f, g, \pi_x\right) + \Omega(g) \tag{3.1}$$

As formally defined in equation, f is the black box model, g is the local explanation model needs to be figured out, and G is a group of interpretable models, which includes linear models, decision trees, or falling rule lists [23]. As depicted in figure, the weight is measured by the proximity of instance of interest to the surrounding artificial instances, which is defined as $\pi_x(z)$. And the complexity of explanation model g is described as $\Omega(g)$. For example, the complexity could be estimated by the depth of trees for decision trees models or by constraining the maximum number of features in linear models. Thus, as seen from the formula, in order to obtain the local explanation model for instance x, the loss L (e.g. mean squared error) should be minimized while maintaining the complexity as low as possible.

In practice, the general procedure to train a explanation model is described as follows: First, select an individual instance that we desire to explain for its black box prediction. Then, generate artificial data points by perturbing the selected sample and make predictions for these new instances using original model. Afterwards, calculate the weights for new instances according to their proximity to the instance being explained. Next, fit a weighted, interpretable model on the obtained dataset. Finally, interpret the instance prediction by utilizing the trained local interpretable model.

After literature review, it is found that LIME is one of the few methods that works for tabular data, text and images, which is a very promising approach. The python implementation is current available in [24], which is still in active development and needs further exploring.

3.1.4 SHapley Additive exPlanations: SHAP

As we have seen, numerous approaches have been recently proposed to explain predictions for individual instances of black box models. As stated in [25], the presented approach is relied on the decomposition of a prediction for a single instance on individual contributions of each attribute, and the contribution for each feature value is measured as the difference between the output value and the average output over all perturbations of the corresponding feature. Nevertheless, this approach fails to work if the features are conditionally dependent.

Inspired by the coalitional game theory which instructs us to fairly distribute the "payout" among the "players", a general method for explaining black box models by taking into account interactions between features can be found in [26], whose fundamental concepts are borrowed to explain instance-level predictions with contributions of each feature values. Corresponding to the known concept in coalitional game theory, the contributions of individual feature values are called Shapley Value.

Despite from the abstract concept, an illustration taken from [3] might help us intuitively understand the Shapley value. Imagine there is a room and all feature values of a individual instance enter the room in a random order. All feature values, seen as players, need to collaborate with each other to participate the game, where each player contributes to receive the final prediction. And each order of feature values represents a coalition. Consequently, the Shapley value of a feature value corresponds to a difference in value of a coalition when the feature is added to it. In another words, the Shapley value is the average marginal contribution of a feature value across all possible coalitions.

Then, let us have a detailed look at the formal definition of Shapley value as expressed in equation 3.2, where S is the subset of the features in a individual instance, p is the number of features, and x is the vector of feature values of the instance to be interpreted. As for characteristic function val, it describes the contribution of feature j in each coalition.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(\text{val}\left(S \cup \{x_j\}\right) - \text{val}(S)\right) \tag{3.2}$$

Referred to [27], the Shapley value can provide the unique solution that adheres to he desirable properties, which are Efficiency, Symmetry, Dummy, and Additivity.

Efficiency: denoted as 3.3, which requires that the sum of feature contributions must equal to the difference of the final prediction and the average prediction over all coalitions.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X)) \tag{3.3}$$

Symmetry: The contributions of two feature values j and k are the same, which means equation 3.4 should be satisfied.

$$\begin{gathered} if\ val\left(S \cup \{x_j\}\right) = val\left(S \cup \{x_k\}\right) \\ then\ \phi_j = \phi_k \end{gathered} \tag{3.4}$$

Dummy: The contribution of feature j is 0 if it does not change the predictions when it joins into any coalitions. This properties can be demonstrated in equation 3.5.

$$\begin{gathered} if\ val\left(S \cup \{x_j\}\right) = val\left(S\right) \\ then\ \phi_j = 0 \end{gathered} \tag{3.5}$$

Additivity: For any pair of games v, w, the combined payouts should equal to the sum of two individual payouts, as shown in equation 3.6. For example, if we trained a random forest and the additivity axiom guarantees that we can calculate the Shapley value for each tree respectively then average them to obtain the final Shapley value.

$$\begin{gathered} \phi_j(v + w) = \phi_j(v) + \phi_j(w) \\ where\ (v + w)(S) = v(S) + w(S) \end{gathered} \tag{3.6}$$

Though classical Shapley value leads to a potentially promising result, this approach is too computationally expensive owing to computations for the exponential number of possible coalitions. Feasibly, approximation algorithms could be used to reduce the computational complexity, nevertheless, it inevitably will increase the variance for the calculation of Shapley value. What is worse, the explanation for the prediction of a model is just a simple value, rather than an explanation model like LIME, which fails to make judgments about the connections between input change and prediction change. To address those problems, Lundberg and Lee [28] proposed a unified framework for explaining predictions, which is based on the Shapley value, and they named it SHAP(SHapley Additive exPlanations). This novel approach unifies existing explanation methods and brings more clarity to the methods space. They introduced the explanation model by treating the explanation of a individual prediction as a model. Of course, the unique solution is guaranteed with the game theory. In addition, it provides a more human-understandable and intuitive explanation by user studies as they claimed.

In this case, SHAP values are introduced as a novel measure of feature contribution. Similar to classical Shapley value estimation methods, SHAP values provide the unique additive feature importance measure if the flowing properties are satisfied, which are Local accuracy, Missingness, and Consistency [28]. From another

perspective, SHAP method transforms the Shapley value approach into an optimization problem by using kernel function to measure proximity of instances. Within this domain, the novel approximation model agnostic method is called kernel SHAP, which is a combination of LIME and Shapley value. In order to use linear explanation model to locally approximate predictions, we should minimize the following objective function 3.1.

It is intended to obtain the unique solution of equation 3.1, which should also be in line with those three properties, the Shapley kernel is defined as [28]:

$$\begin{aligned} \Omega(g) &= 0 \\ \pi_{x'}\left(z'\right) &= \frac{(M-1)}{(M \text{ choose } |z'|)\,|z'|\,(M-|z'|)} \\ L\left(f, g, \pi_{x'}\right) &= \sum_{z' \in Z} \left[f\left(h_x\left(z'\right)\right) - g\left(z'\right)\right]^2 \pi_{x'}\left(z'\right) \end{aligned} \tag{3.7}$$

where $|z'|$ is the number of non-zero elements in z'

The SHAP framework recently is in active development and is accessible at [29]. Since it is seemed to be a very optimistic approach and we are quite interested in this novel model explanation method, therefore, further experiments will be conducted.

3.2 Decision trees with Local interpretation methods

To do

3.3 Pattern mining with Local interpretation methods

Up to now, we have seen a list of local interpretation methods which facilitate us to explain the individual prediction made by the black box model. It is observed that by applying interpretation methods, we obtain an explanation model for each instance regardless of the underlying black box model, meaning that even though the black box model is replaced, the instance-level explanation remain relative consistent, which is a big advantage in some degree. In principle, it is also assumed that two similar instances would lead to two similar explanation models, which are not supposed to provide two disparate explanations, nevertheless, situations like that could happen in practice. The cause might be attributed to the excessive interpretation of the selected instance, which is merely a sample instance rather than a representative of all instances. Therefore, it comes to our mind that we could discover some potentially useful and valid patterns in instances whose individual

explanations are similar, leading us the novel method that combines the subgroup discovery technique with local interpretation methods.

It could be reminded that pattern mining, whose interchangeable name is subgroup discovery, is a data mining technique which pursues to find subgroups of data instances that present interesting rules with respect to a predefined target variable [11]. And subgroup discovery is a descriptive technique, which tells enough details such that the results are understandable by human experts. Therefore, it is believed that the discovered patterns should provide a more robust explanation for the underlying subgroups which also covers the select instance of interest. In the following part, details of this novel technique will be elaborated.

Background: subgroup discovery

Firstly, the theoretical background of subgroup discovery will be introduced as follows.

Formally, four elements can be considered the most important part when a subgroup discovery technique is applied, whose task is defined by a quadruple (D, Σ, T, Q). These elements are defined below [30] [13]:

- D is a dataset and is formed by a set of instances.
- Σ constrains the search space, which is made up of subgroup descriptions (patterns). And patterns consist a set of selection expressions, also known as selectors.
- T represents the target variable for this discovery task. Various types of target concept could be identified, including binary, numeric or complex.
- Q defines the quality measure criteria. Different quality measures are specified for different types of target.

Using binary variable as the target of subgroup discovery is a more simple and general situation. Since the binary variable only contains two values (True or False), it is aimed to identify interesting subgroups for each of the possible value. Basically, the idea is to discover patterns whose target share is either remarkably high or remarkably low. However, pattern mining for numeric target is more complicated because the variable can be dealt with in a numerous ways such as numeric target discretization in a predefined number of intervals, or dividing the numeric domain in two ranges with respect to the average.

Recall from local interpretation methods, the contribution values for each variable of an individual instance could be obtained by using the explanation model to interpret the black box model. In this case, the contribution score of the chosen variable is our target, which is naturally to be numeric. Now, the next step goes to the traditionally pattern mining problem, aiming to discover subgroups of the population that are statistically interesting.

Interestingness measure for numeric target

Unlike binary target situation where interestingness measures are easy to investigate and research, the quality measure for numeric target becomes more complex. Nevertheless, a list of interestingness measure for numeric target could be found in [31]. And a substantial discussions about quality measure could also be traced in [16]. As could be summarized, numeric attribute has certain number characteristics, such as the mean value or the median value. From that perspective, it could be imagined to specify interestingness measure for numeric target with respect to those predefined data statistics. In this case, by comparing the statistics in the subgroups and in the entire population, the interesting groups will be discovered. Generally, there are five categories of interestingness measure for numeric target, concluded in [13], which are mean-based measures, median-based measures, variance-based measures, distribution-based measures, and rank-based measures. Among all of them, the details of mean-based measures will be extended since they are adopted in the experimental phase.

To elaborate a bit more, measuring the quality of a specific subgroup depends on the difference between the mean value in the subgroup and the mean value in the entire dataset. The general formulation is denoted in equation 3.8, where i_P is the size of the subgroup, a is a parameter which weights the subgroup size and deviations, and $\mu_P, \mu_\emptyset$ represent the mean value in the subgroup and the mean value in the dataset respectively. In particular, the choice of parameter a could be selected in an iterative process. For example, a is incremented if the subgroup size is too small to have a significant score, meanwhile, low parameter values for a is preferred with a high deviation of mean target values between the subgroup and the overall dataset. Therefore, after calculating the quality score for each subgroup, those subgroups with significantly higher or lower mean values are considered as interesting and the descriptions of them are our desirable interesting patterns.

$$q^a_{mean}(P) = i^a_P \cdot (\mu_P - \mu_\emptyset), a \in [0, 1] \tag{3.8}$$

Redundancy Avoidance

Though patterns could be discovered through the traditional interestingness measure presented above, the results are not ideal, which contains too many redundant patterns. In the quality measure, only the subgroup size and statistics difference between subgroups and entire dataset are considered, which might produce uninteresting patterns when ignoring the selector expressions of subgroups. For example, assume that the mean contributions of age for the entire dataset is at $M_\emptyset = 0.50$. And the mean value in the subgroup with the expression $age > 40 \cup gender = male$ is $M_{age>40 \cup gender=male} = 0.80$. It seems that the pattern should have a high quality score and is identified as an interesting pattern. However, it is probably not interesting enough if given the information that its generalization has nearly the same value, which means mean value does not deviate significantly from the mean value of its generalizations, e.g. $M_{age>40} = 0.78$.

To avoid that such subgroups are included in the result set, Generalization-aware interestingness measures could be applied to improve the traditional selection criteria for pattern mining by considering the statistics of the subgroup and also to its all

generalizations. In [32], Grosskreutz et al proposed to estimate the quality of a pattern P as the minimum of the quality of P with respect to the extension of all its generalizations. Denoted as equation 3.9, q^{Δ} is the incremental version of q, D is the dataset, P is the subgroup and H includes its all generalizations.

$$q^{\Delta}(DB, P) = \min_{H \subset P} q\left(DB\left[H\right], P\right) \tag{3.9}$$

Since the mean value of the target are mainly explored, the above equation could be formalized in a simpler way, as shown in 3.10. By doing so, redundant patterns are avoided and more interesting subgroups are discovered.

$$q^{a}_{\text{mean}}(P) = i^{a}_{P} \cdot \left(\mu_P - \max_{H \subset P} \mu_H\right), a \in [0, 1] \tag{3.10}$$

Search strategy Apart from the quality measure, the search strategy is critical since the dimension of the search space and time complexity if of great concern. Various strategies could be used, e.g. exhaustive methods, seeking to acquire the optimal subgroup by traversing through the whole search space. In contrast, heuristic approaches, normally a beam search strategy [17], is often used for subgroup discovery due to its efficiency, which aims to find interesting patterns but not necessarily the optimal patterns in a short time. The intuition behind is that it is assumed that the patterns are more likely to be interesting if their generalization are also interesting. Therefore, the search starts with a empty hypothesis, then it tries to find best patterns with size k (corresponding to beam width) by evaluating all selectors in the subgroup discovery task. Following that, at each search iteration, the hypotheses contained in the beam are expanded but only the currently best w hypotheses are kept using a hill-climbing greedy search [33].

Chapter 4

Experiments

Experiments

4.1 Datasets

4.1.1 Datasets description

4.1.2 Data preprocessing (experiment settings)

4.2 Experiments

4.2.1 Artificial Datasets

Before exploring the local interpretation methods, we would like to justify the concept that interesting subgroups could be recovered from the artificial dataset by inspecting variable influence. Presumably, there were hidden patterns in the synthetic dataset that were useful to provide reasonable explanation to the predictions. By interpreting the effect of a certain variable, e.g. gender, it was assumed that the interesting pattern could be recovered. The procedure to construct an artificial dataset and conduct experiments will be described as follows.

For simplicity, we constructed the artificial dataset relying on the popular adult dataset, but we only extracted partial information, which meant that only the information about age, education-num, sex, hours-per-week and income were included. As assumed, the synthetic data contained some interesting patterns, such as "age < 30". One exemplary case was that when "age < 30", the attribute "gender" had stronger effect on predictions while in its complementary subgroup, the effect of "gender" was slight. And the task was indeed to discover this pattern by exploring the effect of gender. For further experiments, one way to fabricate the interesting pattern was to modify the gender effect directly on the corresponding subgroups.

For instance, if the condition that "age < 30" was met, we could manually add 3 unit in terms of the scale of measurement on gender effect, and otherwise we could subtract 3 unit. Another idea was to establish two models that behaved differently when considering this condition. It is known that the coefficients in the logistic regression model have straightforward interpretation, indicating the influence level by the input features. Therefore, we could create two distinct models by changing the weights of the features in accordance with the previous defined patterns. In specific, we could assign larger weights to the model that was applied on the pre-described subgroup to maintain larger gender effect, while decreasing feature weights on the model that was applied on its complementary subgroup.

In this paper, we would like to adopt the latter method to make up the synthetic dataset and build the models. To measure the gender effect, we could simply use the binary flip approach described in previous chapter. By flipping the gender value, i.e. transform from "male" to "female" or the other way around, the prediction change denoted as probability was calculated and roughly it was regarded as the effect of gender. Then, treating the effect of gender as the target concept, subgroup discovery technique was applied on the artificial dataset to discover interesting subgroups. It could be observed that these interesting subgroups include the subgroup that were artificially generated in the dataset. The detailed results were left to the next chapter. In conclusion, it could be proved that subgroup discovery technique could indeed provide us patterns of explanations that facilitate us to understand the predictions.

4.2.2 Comparison of different local interpretation methods

datasets: adult, credit-g, housing binary flip: perturbation, LIME, SHAP

numeric perturbation: perturbation, LIME, SHAP

classification vs. regression

decision tree vs. subgroup discovery

4.2.3 Case Study

real world case study 2-3 datasets

Chapter 5

Conclusion and Future work

conclusion and future work...

5.1 Conclusion and Feature work

5.1.1 Factors to consider

Conclusion: Local surrogate models, with LIME as a concrete implementation, are very promising. But the method is still in development phase and many problems need to be solved before it can be safely applied.

5.1.2 Summary

5.1.3 Outlook

Bibliography

[1] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, 2016, pp. 2280–2288.

[2] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

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

[4] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD explorations newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[5] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[6] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[7] A. Fisher, C. Rudin, and F. Dominici, "Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective," *arXiv preprint arXiv:1801.01489*, 2018.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[9] H. Cheng, X. Yan, J. Han, and S. Y. Philip, "Direct discriminative pattern mining for effective classification," in *2008 IEEE 24th International Conference on Data Engineering*. IEEE, 2008, pp. 169–178.

[10] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. Citeseer, 1999, pp. 43–52.

[11] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and information systems*, vol. 29, no. 3, pp. 495–525, 2011.

[12] M. Atzmueller and F. Lemmerich, "Fast subgroup discovery for continuous target concepts," in *International Symposium on Methodologies for Intelligent Systems*. Springer, 2009, pp. 35–44.

[13] F. Lemmerich, "Novel techniques for efficient and effective subgroup discovery," 2014.

[14] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2008, pp. 1–16.

[15] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, 1996, pp. 249–271.

[16] B. F. Pieters, A. Knobbe, and S. Dzeroski, "Subgroup discovery in ranked data, with an application to gene set enrichment," in *Proceedings preference learning workshop (PL 2010) at ECML PKDD*, vol. 10, 2010, pp. 1–18.

[17] P. Clark and T. Niblett, "The cn2 induction algorithm," *Machine learning*, vol. 3, no. 4, pp. 261–283, 1989.

[18] M. J. Zaki, "Scalable algorithms for association mining," *IEEE transactions on knowledge and data engineering*, vol. 12, no. 3, pp. 372–390, 2000.

[19] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.

[20] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer, 1997, pp. 78–87.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[22] ——, "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*. ACM, 2016, pp. 1135–1144.

[23] F. Wang and C. Rudin, "Falling rule lists," in *Artificial Intelligence and Statistics*, 2015, pp. 1013–1022.

[24] S. S. Marco Tulio Ribeiro and C. Guestrin. (2019) Lime: Explaining the predictions of any machine learning classifier. [Online]. Available: https://github.com/marcotcr/lime

[25] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 589–600, 2008.

[26] I. Kononenko *et al.*, "An efficient explanation of individual classifications using game theory," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 1–18, 2010.

[27] L. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, pp. 31–40, 1953.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[29] ——. (2019) shap: A unified approach to explain the output of any machine learning model. [Online]. Available: https://github.com/slundberg/shap

[30] M. Atzmueller, F. Puppe, and H.-P. Buscher, "Towards knowledge-intensive subgroup discovery." in *LWA*. Citeseer, 2004, pp. 111–117.

[31] W. Klösgen, "Data mining tasks and methods: subgroup discovery: deviation analysis," in *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., 2002, pp. 354–361.

[32] H. Grosskreutz, M. Boley, and M. Krause-Traudes, "Subgroup discovery for election analysis: a case study in descriptive data mining," in *International Conference on Discovery Science*. Springer, 2010, pp. 57–71.

[33] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.