2-2021

Visual analysis of discrimination in machine learning

Qianwen WANG
Hong Kong University of Science and Technology

Zhenghua XU
Hong Kong University of Science and Technology

Zhutian CHEN
Hong Kong University of Science and Technology

Yong WANG
Singapore Management University, yongwang@smu.edu.sg

Shixia LIU
Tsinghua University

See next page for additional authors

Visual Analysis of Discrimination in Machine Learning

Qianwen Wang, Zhenhua Xu, Zhutian Chen, Yong Wang, Shixia Liu, and Huamin Qu

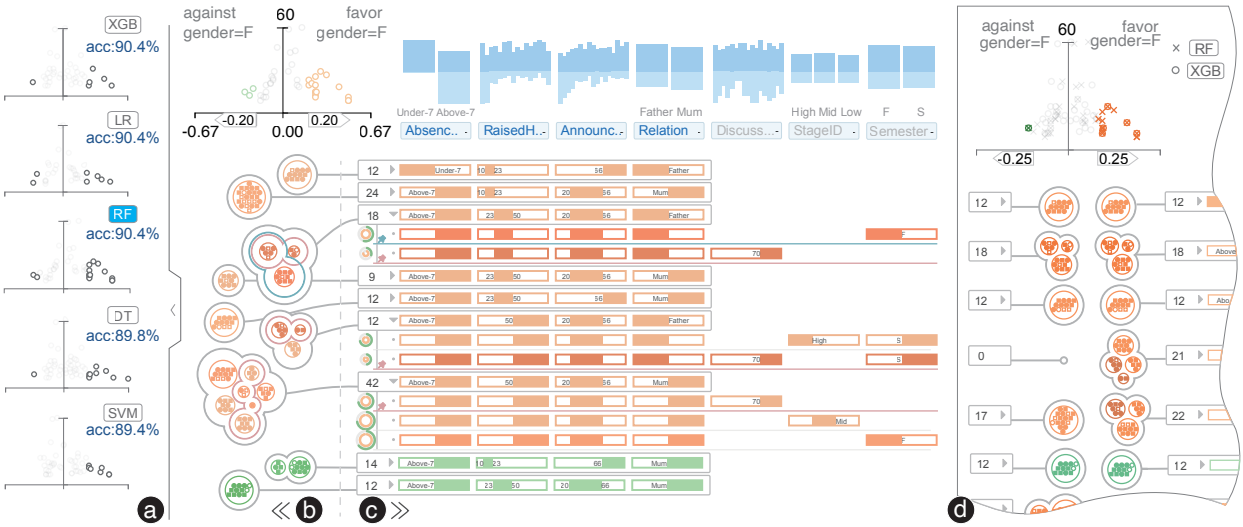

Fig. 1. DiscriLens facilitates a better understanding and analysis of algorithmic discrimination: (a) scatter plots offer an overview of the discriminatory itemsets; (b) RippleSets reveal the intersections among these itemsets; (c) the attribute matrix represents the details of each discriminatory itemset; (d) the comparison mode enables users to compare two models side by side.

Abstract—The growing use of automated decision-making in critical applications, such as crime prediction and college admission, has raised questions about fairness in machine learning. How can we decide whether different treatments are reasonable or discriminatory? In this paper, we investigate discrimination in machine learning from a visual analytics perspective and propose an interactive visualization tool, DiscriLens, to support a more comprehensive analysis. To reveal detailed information on algorithmic discrimination, DiscriLens identifies a collection of potentially discriminatory itemsets based on causal modeling and classification rules mining. By combining an extended Euler diagram with a matrix-based visualization, we develop a novel set visualization to facilitate the exploration and interpretation of discriminatory itemsets. A user study shows that users can interpret the visually encoded information in DiscriLens quickly and accurately. Use cases demonstrate that DiscriLens provides informative guidance in understanding and reducing algorithmic discrimination.

Index Terms—Machine Learning, Discrimination, Data Visualization.

1 Introduction

Machine learning (ML) has progressed dramatically in recent decades and become a useful technique in a variety of applications, including credit scoring [31], crime prediction [20], and college admission [50]. Since decision-making in these areas may have ethical or legal issues [14, 46], it is crucial for model users to go beyond model accuracy and consider the fairness of ML models.

Consider the following scenario. When reviewing loan applications, loan officers need to estimate the risk of default (*i.e.*, the probability of failing to repay the loans), which is usually time-consuming and error-prone. A machine learning model trained on historical credit data can estimate the creditworthiness of applicants and thus facilitate the decision-making. However, this model can unintentionally make discriminatory predictions in the social sense, even though the training data describes objective facts and includes no human discrimination. For example, this model may treat two applicants unequally based on gender despite their same repayment capacity. To avoid making decisions based on protected attributes (attributes such as gender and race that are legally protected by laws [46, 58]), a straightforward method is to hide these attributes. But this method not only decreases the model accuracy but has also been proven ineffective since models are able to learn protected attributes from other non-protected attributes (*e.g.*, predict gender based on address and occupation) [21, 49, 67].

To further promote the adoption of ML models and prevent potential negative social impacts, discrimination in ML is drawing increasing research attention. Many methods have been proposed to assess and mitigate discrimination from three main aspects: pre-process methods that investigate the discrimination in training data [27, 35, 63], in-process methods that adjust the model learning process [29, 41, 62], and post-process methods that modify the discriminatory model predictions [23, 65]. However, these studies usually formalize discrimination as summary statistics and may hinder a detailed assessment. Meanwhile, these studies simply assume that the representation of discrimination has been clearly defined, which usually does not hold in practice [23, 48]. Due to the complex nature of discrimination, it has no clear and uniform definition and its representation varies a lot in different domains. In this study, we develop a visual analysis tool that enables the involvement of domain knowledge and supports a systematical

- *Qianwen Wang, Zhenhua Xu, Zhutian Chen, Yong Wang, and Huamin Qu are with Hong Kong University of Science and Technology. E-mail: {qwangbb, zxubg, zhutian.chen, ywangct}@connect.ust.hk, huamin@ust.hk.*
- *Shixia Liu is with Tsinghua University. E-mail: shixia@tsinghua.edu.cn.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

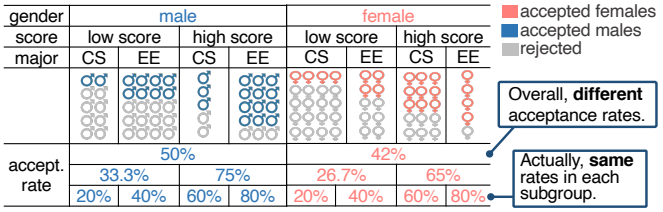

Fig. 2. It is not easy to distinguish reasonable differences from discriminatory treatment. Here is a toy example of college admission. The overall low acceptance rate for females can be explainable by their tendency to apply to the more competitive major.

assessment of discrimination, thus further benefiting the analysis and mitigation of discrimination [24].

To analyze discrimination in machine learning, it is important to access whether differential treatment is discriminatory (*e.g.*, based on race) or reasonable (*e.g.*, based on the qualification of the applicant). The term *fairness* used in this paper refers to the principle that any two individuals who are *similar with respect to a particular task* should be *classified similarly* [17, 32]. The first question is **which individuals should be regarded as similar for the task at hand**. The definition of similar people varies a lot among tasks and is important for the analysis of discrimination. Fig. 2 shows an example of college admission, where `gender` is a protected attribute. If all the applicants are treated as similar, there seems to be gender discrimination since the acceptance rate is 42% for females but 50% for males. Unequal treatment still exists if we group applicants based on the `test score`. But if we define similar people using the combination of {`major`, `test score`}, it then shows no unequal treatment in any of the sub-groups. Such a phenomenon is also mentioned in Simpsons paradox [26]. The second question is **how to present discrimination among these similar individuals effectively**. We treat a group of similar people as an **itemset** defined by a series of attributes values (*e.g.*, {`test score=low, major=CS`}). The interpretability of these itemsets is severely weakened when the definition is long and complex. Furthermore, the number of these itemsets can be large and these itemsets are often intricately intertwined together. Therefore, it is non-trivial to assist users in perceiving these itemsets and interpreting discrimination.

To tackle thechallenges, we design and implement DiscriLens, an interactive visualization tool that facilitates an easy interpretation, evaluation, and comparison of the algorithmic discrimination. A demo is available at `http://discrilens.hkustvis.org` (recommend opening in Chrome). We develop a three-stage pipeline to identify a collection of potentially discriminatory itemsets based on causal modeling and classification rules mining. A set of user interactions are provided to incorporate human domain knowledge in discrimination analysis. A novel Euler-based visualization, RippleSet (Fig. 1(b)), is proposed to provide an effective presentation of discrimination. RippleSet represents one set as several adjacent circles instead of one convex shape, thereby preventing the overlaps in traditional Euler diagram. We further combine the RippleSet with a matrix-based visualization (Fig. 1(c)) to support users in examining the discriminatory itemsets from multiple aspects. We demonstrate the effectiveness of DiscriLens in analyzing discrimination through a user study and use cases.

The main contributions of this work are as follows:

- The design and development of DiscriLens, an interactive visual analysis tool with a set of novel visualization techniques for analyzing discrimination in machine learning.
- A user study and a series of use cases that assess the utility and usability of DiscriLens.

2 Related Work

2.1 Discrimination in Machine Learning

Existing studies mainly investigate algorithmic discrimination from two aspects: the data and the model training process [21].

Training data may include human discrimination and then influence the trained model [35]. Various approaches have been proposed to discover discrimination existing in the training data [10, 27, 41, 47, 67]. A pragmatic solution is to examine different treatments among similar items [27, 47]. For example, Luong et al. [40] used the k-nearest neighbor algorithm to group similar items and identify the discrimination among them. Pedreshi et al. [49] employed classification rules to reveal itemsets that may lead to unequal treatment. Recently, to provide an interpretable notation of discrimination, several studies considered the relationships among item attributes and examined discrimination from a causal modeling perspective [35, 63].

Even though human discrimination can be identified and eliminated from the training data, a model can still make discriminatory predictions due to the data distribution and the model learning mechanism (*e.g.*, over-fitting) [39, 48, 65, 67]. To remove discrimination introduced in the training process, researchers developed regularizers that penalize discriminatory predictions [29, 62] and methods that directly modify the model predictions [23, 65]. Many studies have defined different criteria for model discrimination to provide meaningful [48, 65] and interpretable [23] notations. Meanwhile, to support users in measuring and comparing discrimination in different models, initial efforts have been made to quantify model discrimination [55].

These studies shed light on discrimination analysis and form the foundation of this study. However, the discrimination mined by these methods can be complex and thus hard to interpret (*e.g.*, a large number of rule lists), requiring the support of effective visual presentation. More importantly, the analysis of discrimination is by nature subjective and varies based on the application domains. This paper provides interactive analysis to help incorporate human domain knowledge in discrimination analysis.

2.2 Visual Analysis for ML Discrimination

Visualization, an effective tool for information communication, has been employed to present and analyze discrimination in ML [1, 8, 9, 25, 36, 56]. For example, Google Big Picture developed an interactive visualization demo [8] to illustrate the discrimination in ML decisions. Other tools, such as IBM AI Fairness 360 [25], Tensorflow Fairness Indicators [56], and FairLearn, enable the easy computation of fairness metrics and support visualizations of these fairness metrics. However, most tools only support segmenting groups based on one protective attribute and only provide basic visualizations (*e.g.*, bar charts). Therefore, they cannot be directly applied to investigate and present the potentially complicated discrimination between different groups.

Recently, advanced visualization tools have been developed to enables more comprehensive analysis of discrimination, including FairSight [1] and FairVis [9]. FairSight represents a workflow that supports the four fairness-aware actions (*i.e.*, understand, measure, identify, and mitigate) required in decision making. FairSight shares the same high-level goal as DiscriLens and also supports the analysis of individual fairness. However, FairSight either a) identifies discriminated instances based on an KNN algorithm or b) provides a global-level measure by aggregating over all instances. As a result, FariSight fails to uncover the discriminatory itemsets and to reveal *"when and where will a model yield discriminatory predictions"*. FairVis is more related to DiscriLens due to its focus on the analysis of discriminatory itemsets. This tool helps users to audit the fairness among different itemsets. However, FairVis focuses on suggesting similar discriminatory itemsets during analysis. The possible complex definition of the itemsets are presented using tables, which is not an optimal representation method. No special visualizations are designed to reveal the intertwining relationships among the subgroups.

In DiscriLens, we propose novel visualizations for better understanding and analysing ML discrimination.

2.3 Set Visualization

The core of discrimination analysis is to understand how a set of items, grouped by attribute values, are treated unequally. Due to the ubiquity of set-based analysis, a large body of literature has been developed. Below we review some visualizations that are germane to our work. A comprehensive review of set visualizations can be found in the work of Alsallakh et al. [4].

The most common set visualizations are Euler and Venn diagrams [19, 37], which represent sets as convex or concave shapes and show set intersections as overlapping shapes. Previous work has proposed many extensions of Euler diagram with varying design goals [12, 15, 52, 54]. For example, Riche et al. [52] simplified a sophisticated collection of intersecting sets into a strict hierarchy, thereby untangling Euler diagrams and improving its readability. Bubble Sets [12] delineate set memberships and minimize cluster overlap while retaining the layout that reflects the semantic spatial organization. However, methods based on Euler and Venn diagrams often impose severe limitations on the number of sets and the complexity of set intersections [3].

Considering the inherent limits of Euler and Venn diagrams, researchers have developed many alternative presentations of set data, including curve-based presentation [2], treemap-based presentation [5], parallel coordinate-based presentation [34], matrix-based presentation [33, 37], and chord diagram-based presentation [3]. The most relevant ones to our work are Upset [5] and Conset [33], which employed matrix to communicate the properties of the set aggregations and intersections. However, Upset and Conset are specified for visualizing binary set data and cannot be directly applied to the categorical set data in discrimination analysis. Meanwhile, in existing work, the comparison of two collections of set data has not been adequately explored and requires further investigation.

In this work, to facilitate the analysis and comparison of categorical set data, we propose a novel set visualization that combines an extended Euler diagram, RippleSet, with a matrix-based visualization.

3 Discrimination: A Mathematical Notation

This section provides background information and a mathematical notation of discrimination. In this paper, we focus on the analysis of a single protected attribute, denoted by A. We denote other non-protected attributes by X, and the outcome by Y. Note that the outcome Y can be the labels in the training data or the predictions of ML models. The beneficial class (*e.g.*, a loan approval) is denoted by $Y = 1$, and the protected group (*e.g.*, females) is denoted by $A = 1$. Unobserved attributes are not considered in this study.

A natural notation of fairness is demographic parity [23, 35], which states that each segment of a protected class (e.g. blacks and non-blacks) should receive positive outcomes at equal rates (*e.g.*, same admission rate), regardless of the value distribution of other attributes (*e.g.*, score). In other words, this notation treats all individuals as similar and requires the decisions to be independent of the protected attribute, as shown in Fig. 3(a)(b). Demographic parity is easy to examine and thus has been widely used in legal practice, including the US Equal Employment Opportunity Commission [57] and the UK Sex Discrimination Act [45]. However, this notion can be flawed when the correlation between attributes is strong and needs to be considered, which is the situation we consider in this paper.

As shown in Fig. 3(c), the protected attribute `race` are correlated with other attributes and can influence the `admission` from two paths: **justified difference** (`race`→`major`→`admission`) and **indirect discrimination** (`race`→`postcode`→`admission`). This first path is called justified difference since `major` offers objective information about the `admission` and can explain the outcome differences. *E.g.*, the overall low admission rate for black students can be explained by their tendency to apply for more competitive majors. Such attributes that can justify decision differences are called as **resolving** attributes [32]. The second path is called indirect discrimination because the `postcode` is correlated with `race` (*e.g.*, majority-black neighborhood) but should not affect the `admission`. These attributes that act as a proxy for protected attributes are called **proxy** attributes [32]. As a result, resolving and proxy attributes should be both considered for the analysis of discrimination [32, 64]. In general, there is no ground truth for resolving or proxy attributes. Their definitions depend on the task (*e.g.*, college admission) and require human domain knowledge. Therefore, an interactive visual analysis of discrimination is needed.

In this study, we use **risk difference** (RD), a widely used metric in anti-discrimination literature [48, 53, 63, 64], for the measurement of discrimination. $RD = P(Y = 1|A = 1, S = s_i) - P(Y = 1|A = 0, S = s_i)$,

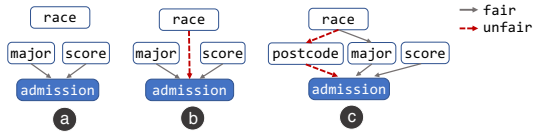

Fig. 3. In a toy example of college admission, different types of relationship exist between the protected attribute `race` and the result `admission`. (a) `race` doesn't influence `admission`. (b) `race` directly influences `admission`. (c) `race` indirectly influences `admission` through other attributes.

where S includes all resolving attributes and no proxy attributes, ensuring that justified difference is separated from discrimination. Following the practice in [63], non-discrimination is claimed when $RD < \tau$ holds for each itemset s_i. Note that DiscriLens is not metric specific and can be easily extended to other metrics such as true positive rate difference, error rate ratio, and odds ratio [25]. We would like to emphasize that the notation used in this paper is just one of many possible choices. The choice of notation depends on the context of analysis.

4 Designing DiscriLens

4.1 Design Goals

The goal of DiscriLens is to enable model users to flexibly interpret and analyze the discrimination in ML. The users have a certain level of expertise in ML and wish to apply ML models to a real-world application. They want to analyze discrimination to better avoid the potential ethical or legal issues. Based on the previous literature on algorithmic discrimination and discussions with two ML experts, we identify the following design goals.

G1 Customize the definition of discrimination.

Discrimination is a complex concept and can have different representations at different domains [44]. For example, a decision based on age is discrimination in certain domains but not in others. Also, the resolving attributes to a decision significantly vary from domain to domain. Therefore, the domain knowledge of the users is essential for the analysis of discrimination. Users should be allowed to customize the definition of discrimination (*i.e.*, protected attribute, proxy attributes, resolving attributes, the threshold τ) based on their domain knowledge as well as the results of discrimination discovery algorithms.

G2 Measure the degree of discrimination.

For the analysis of discrimination, model users need to measure the degree of algorithmic discrimination [48, 55], which can be evaluated by two indicators: the *risk difference* and *size*. The *risk difference* answers the question: *"To which extent are the protected group ($A = 1$) and non-protected group ($A = 0$) treated unequally?"* The *size* answers the question: *"How many people are influenced by algorithmic discrimination?"* High *risk difference* and large *size* indicate severe discrimination.

G3 Identify the condition of discrimination.

Apart from measuring the degree of discrimination, the condition of discrimination should be identified to reveal when a model will yield discriminatory predictions. This information can be conveyed by a set of attribute values that specify the discriminatory itemsets, *i.e.*, a group of individuals who are similar to the task but are treated differently. The condition of discrimination is crucial since it can provide detailed guidance on applying models (*e.g.*, identify the failure mode [49]), improving the training data (*e.g.*, modify inappropriate data labels [63], collect new data [66]), and adjusting the training process (*e.g.*, add regularization to penalize discriminatory predictions [41, 62]).

G4 Depict the distribution of discrimination. The degree and the condition of discrimination can vary a lot among itemsets [49, 67]. For example, in income prediction, the degree of discrimination towards blacks may vary with other attribute values, such

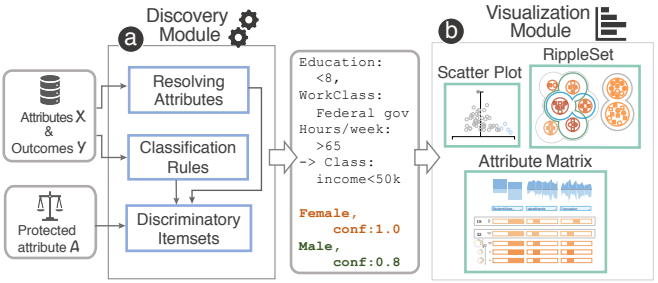

Fig. 4. DiscriLens consists of two main modules: a discovery module (a) and a visualization module (b).

as `working hours`, `education level`, and `work class`. As a result, algorithmic discrimination can hardly be comprehensively described by summary statistics or easily interpreted by humans [55]. The difficulty is significantly increased when the intersections between itemsets are complex. To fully capture the discriminatory behavior of a machine learning model, the distribution of discrimination should be offered to help users understand how discrimination varies among and inside different itemsets.

G5 Compare discrimination. Summary statistics can be ineffective for the analysis of ML model behavior. Models with the same accuracy can assign conflicting predictions for the same input data [42], leading to discriminatory predictions against different groups of data items. For example, one model may discriminate equally against all female students, while another model discriminates severely against females with low scores but mildly against other female groups. Therefore, in order to choose an appropriate model, it is important to compare models and analyze these conflicting predictions.

Apart from model-level comparison, it is also important to compare discrimination among different itemsets. Several studies [13, 48] have observed that model discrimination cannot be completely removed. Therefore, different itemsets should be compared to prioritize the most severe discrimination, whose identification is objective and requires human domain knowledge.

4.2 System Overview

DiscriLens consists of two main modules: a discovery module and a visualization module (Fig. 4). **The discovery module** takes the training data, the model, and the user-defined protected group as input. It then goes through a three-stage pipeline and produces a collection of potentially discriminatory itemsets. **The visualization module** serves as an interface that helps understand the discrimination, as well as a tool that provides guidance on applying and improving the model. Details of the discovery and the visualization module are introduced in the following two sections.

5 Discrimination Discovery

We view the task of discovering discrimination as a problem of mining discriminatory itemsets [49, 63]. Given the outcomes $\hat{Y} = f(X)$ and a threshold τ, the goal is a collection of itemsets $\{s_1, s_2, ..., s_n\}$ that $|P(\hat{Y} = 1|s_i, A = 0) - P(\hat{Y} = 1|s_i, A = 1)| > \tau, i \in n$, where the itemset s_i is a group of items that are similar to the task. As shown in Fig. 4(a), the discovery module contains three main components: mine resolving attributes, mine classification rules, and mine discriminatory itemsets.

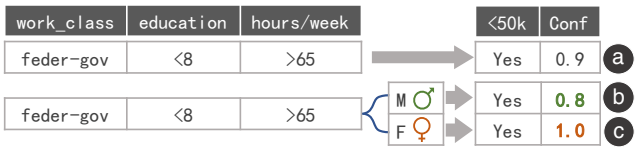

Fig. 5. (a) is an example of classification rules. The confidence difference between rule (b) and (c) might indicate a discrimination towards females and requires further examination.

5.1 Resolving Attributes

In this step, we mine resolving attributes following the causality-based method proposed in [63]. Causal modeling is a complicated task and requires strong prior knowledge. Fortunately, when analyzing discrimination, we only need to mine local causality and identify the parents of Y (*i.e.*, $(Pa(Y))$ without building the complete causal graph. We refer the readers to [32, 63] for more details and the complete mathematical proof. The resolving attributes can be then represented as $Par(Y) \setminus (A \cup P)$, where P indicates the proxy attributes. The identification of proxy and resolving attributes requires users' domain knowledge. Therefore, we allow users to interactively modify resolving attributes, remove proxy attributes, and analyze the results in the visualization module. In this implementation, we use the Fast Greedy Equivalence Search [11] for local casual discovery.

5.2 Classification Rules

In this stage, we can extract a list of classification rules to explain the outcome Y. Take a support vector machine trained on Adult dataset [16] as an example. The task is to predict whether the annual income of a person is over 50k USD. One of the classification rules is shown in Fig. 5(a). The support of this rule $supp(s \rightarrow \hat{Y} = 1) = supp(s, \hat{Y} = 1) = 16$ and the confidence of the rule $conf(s \rightarrow \hat{Y} = 1) = P(\hat{Y} = 1|s) = supp(s, \hat{Y} = 1)/supp(s) = 0.9$. In this implementation, we use minimum description length [18] to discretize continuous attributes and FP-Growth [22] to mine the classification rules. Alternative methods can also be used, and we refer the reader to [61] for a detailed discussion.

5.3 Discriminatory Itemsets

Given two classification rules: $(s, A = 0) \rightarrow \hat{Y} = 1, (s, A = 1) \rightarrow \hat{Y} = 1$, we can treat s as a discriminatory itemset if $|P(\hat{Y} = 1|A = 0, s) - P(\hat{Y} = 1|A = 1, s)| > \tau$ and s includes all resolving attributes and no proxy attributes. We do not specify the sign due to reverse discrimination, *i.e.*, discrimination against the non-protected group.

As shown in Fig. 5, if `[workclass, education, hours per week]` include all resolving attributes and no proxy attributes, the confidence difference between rule (b) and (c) can be treated as discrimination towards females. In other words, in the itemset s =`[workclass=feder-gov, education<8, hours per week>65]`, males are less likely to be predicted to have low incomes compared with their female counterparts (0.8 vs. 1.0). Therefore, this itemset is called discriminatory itemset. To enable interactive modification, we do not specify resolving and proxy attributes when mining classification rules. Note that s may contain attributes which are neither proxy nor resovling attributes. We call these attributes as **context** attributes. In real-world, a discriminatory itemset can be defined by a long list of attribute values, and the itemsets are usually intertwined together. As a result, effective visualization is needed to assist users in the analysis of discrimination.

6 Visual Interface

The visualization module consists of three visualization components (Fig. 1). The **scatter plot** (a) offers an overview of the discriminatory itemsets and allow users to filter itemsets. **RippleSet** (b) and **attribute matrix** (c) supplement each other to support a detailed analysis of discriminatory itemsets. RippleSet reveals the discrimination among items that belong to intricately intertwined sets, while attribute matrix illustrates the details of each itemsets.

6.1 RippleSet

When analyzing model discrimination, it is important to effectively represent discrimination among itemsets that have complex intersection and inclusion relationships (**G4**). Existing set visualization methods often impose severe limitations on the complexity of set intersections [3, 4], and thus can hardly satisfy the design goals of DiscriLens. To better support the analysis in DiscriLens, we extend the widely-used Euler diagram and design RippleSet.

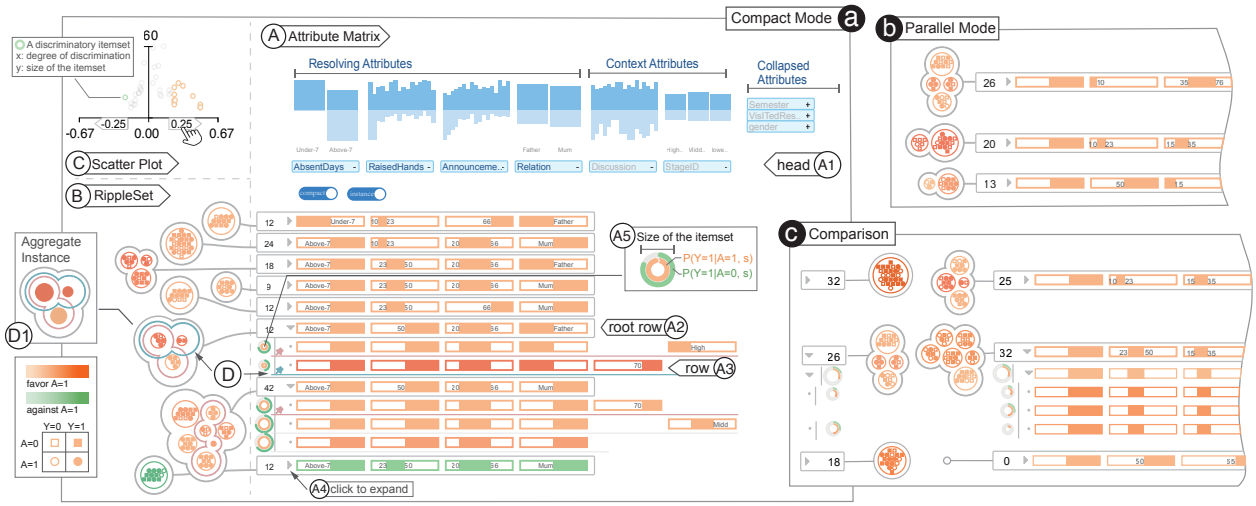

Fig. 6. The visual interface consists of three key components: the attribute matrix (A), the RippleSet (B), and the scatter plot(C). Three modes of analysis are supported: the compact mode (a), the parallel mode (b), and the comparison mode (c).

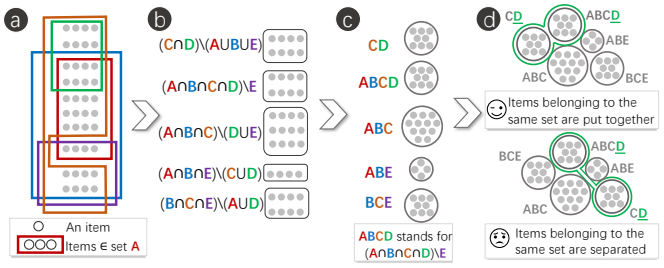

Fig. 7. RippleSet avoids the complex overlapping shapes in the traditional Euler diagram (a) by representing the maximal inseparable subset (b) as circles (c). Subsets belonging to the same set are put adjacent to one another to better present the composition of one set (d).

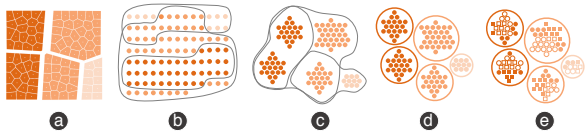

Fig. 8. Alternative designs of the set visualization: the weighted Voronoi-based design (a) contains irregular shapes and makes it difficult to compare the size of subsets; the Euler-based design (b) introduces significant visual clutter. We improved the design in (b) through an iterative design process((c)-(e)).

6.1.1 Visual Encoding

As shown in Fig. 7, the main idea of RippleSet is to represent one set using several adjacent shapes, instead of one shape, to avoid the complex overlapping shapes in the traditional Euler diagram. In RippleSet, each shape stands for one maximal inseparable subset, *i.e.*, a subset that will not be further segmented based on the sets they belong to (Fig. 7(b)). Therefore, more shapes in the RippleSet indicate more complex intersections among itemsets. During the analysis, a set can be highlighted using an outline that surrounds all the subsets of the set.

In RippleSet, we also offer instance-level visualization. Each data item is represented as a dot and placed inside the subset it belongs to, as shown in Fig. 6(C). Inside the subset, we use the algorithm proposed by Wang et al. [60] to layout the data items, which are inserted based on their original order in the dataset. Three visual channels, color, shape, and solid/hollow, are used to display the attributes of data items. Color, the most effective visual channel, is used to present the most important information, the direction and the degree of the discrimination. The color encoding is consistent with that used in the attribute matrix: color saturation indicates the absolute value of the risk difference while color hue indicates the discriminated group (protected group or non-protected group). A solid item represents a data item classified as the beneficial class ($\hat{Y} = 1$) whereas a hollow item represents a data item classified as the non-beneficial class($\hat{Y} = 0$). Shapes of the items encode their groups: circles for the protected group ($A = 1$) due to their metaphor of softness and squares for the non-protected group ($A = 0$) due to their metaphor of hardness [38]. We admit that shape is a less effective visual channel. We plan to double encode the group information through item positions in the future by modifying the layout algorithm. For a large number of items, RippleSet supports the aggregation of items, as shown in Fig. 6(D1). More details about the layout algorithm are provided in the supplementary material.

6.1.2 Design Alternatives

We tested several design alternatives (Fig. 8) for the set visualization, during which we interviewed domain experts and discussed with visualization experts. We first tried a weighted Voronoi-based design, as shown in Fig. 8(a). This design was unsuccessful since the irregular shapes posed challenges to users for comparing the size of subsets. We then implemented a Euler-based design, as shown in Fig. 8(b). From Fig. 8(b) to (e), we clustered items belonging to the same subset to offer a clear organization, replaced irregular outlines with circles for aesthetics, and added detailed information of each item.

6.2 Attribute Matrix

The attribute matrix (Fig. 6(A)) helps users investigate each discriminatory itemset in detail. The attribute matrix employs a similar matrix-based layout as RuleMatrix [43]. This design has been proven effective in presenting classification rules, which is used to identify discriminatory itemsets in this paper. In the attribute matrix, each row stands for an group of individuals and each column stands for an attribute. The head (A1) represents the distribution of data items on different attributes, a root row (A2) represents a collection of discriminatory itemsets with the same resolving attribute values, and a row (A3) represents one discriminatory itemset.

The head (Fig. 6(A1)) of the attribute matrix consists of a row of charts. Histograms are used for visualizing the distribution on continuous attributes, and bar charts are used for discrete attributes. In each chart, the x-axis represents the attribute value, and the y-axis represents the number of items. We use dark blue to indicate the items classified as the beneficial class (*e.g.*, a loan approval) and light blue for the non-beneficial class (*e.g.*, a loan decline). To distinguish resolving attributes,

the names of resolving attributes are highlighted. Users can modify the resolving attributes based on their domain knowledge as well as their observation of the data distribution (**G1**). Discriminatory itemsets will dynamically update according to the modified resolving attributes.

To facilitate the interpretation and exploration, we group and aggregate discriminatory itemsets to offer a clear summary. We first group the discriminatory itemsets based on their resolving attribute values. All itemsets sharing the same resolving attribute values are put in one collection, which is represented as a bordered rectangle called root row (Fig. 6(A2)). The number of items in a collection is labeled at the left end of the bordered rectangle. Then, in each collection, we organize itemsets, which are represented as rows (Fig. 6(A3)), according to their inclusion relations. If an itemset A is included in itemset B (*i.e.*, $A \in B$), A can be opened by clicking the expand icon on B.

For each discriminatory itemset, a simple glyph (Fig. 6(A5)) is put at the left end to indicate the degree of discrimination (**G2**). The radius of the glyph encodes the size of the itemset. The angle of the outer arc is determined by the probability that the non-protected group is labeled as beneficial class(*i.e.*, $P(\hat{Y} = 1|A = 0, s)$) and the angle of the inner arc is determined by the probability that the protected group is labeled as beneficial class (*i.e.*, $P(\hat{Y} = 1|A = 1, s)$).

The condition of discrimination (*i.e.*, a series of attribute values) is represented by an array of rectangles, whose solid parts represent the specified ranges/values of the aligned attributes (**G3**). Since we discretize continuous attributes for mining classification rules, all attributes here are actually categorical and are visualized using similar encodings with continuous attributes. The color saturation of these rectangles is determined by the absolute value of the risk difference and the color hue is determined by the sign of the risk difference (**G2**). To better compare the condition of discrimination (**G5**), itemsets are vertically arranged in an order that maximizes the Jaccard similarity of adjacent itemsets.

6.3 Interactions

Filter Itemsets & Modify Resolving Attributes

The scatter plot (Fig. 6(C)) allows the filtering of itemsets by offering a summary of all discriminatory itemsets (**G4**). In the scatter plot, one point represents one discriminatory itemset, with the x-axis indicating the *risk difference* (the extent of the unequal treatment) and the y-axis indicating the size of the itemset (**G2**). The scatter plots help users measure the discrimination from a coarse level and guide them in the filtering operation. By moving the sliders on the x-axis, users can change the threshold of risk difference and select itemsets for further exploration. The scatter plot also supports a coarse level comparison of models (**G5**) through two ways: juxtaposition (*i.e.*, side by side) and superposition (*i.e.*, overlay).

In the head of the attribute matrix (Fig. 6(A1)), users can modify resolving attributes (**G1**) through drag and drop. Such modification changes the mined discriminatory itemsets. Other visual components, *i.e.*, the scatter plot, the RippleSet, and the rows in the attribute matrix, will be updated dynamically.

Coordinate RippleSet with Attribute Matrix

The attribute matrix and the RippleSet are coordinated to support a comprehensive analysis. Each root row of the attribute matrix is linked to a RippleSet through a curve. Clicking on a RippleSet can expand the rows belonging to the root row. Users can hover over a row to highlight the corresponding itemset in RippleSet. By clicking on a row, users can draw an outline for this itemset (Fig. 6(D)) to facilitate their understanding of the discrimination distribution (**G4**).

Three modes of analysis are supported: the compact mode, the parallel mode, and the comparison mode. In the compact mode (Fig. 6(a)), rows in the attribute matrix are compact, and each RippleSet is placed to minimize the overall L1 distance between the RippleSets and the corresponding root rows. The compact mode provides space efficiency and can offer an overview of a relatively large number of itemsets. A "focus+context" technique is employed in the compact mode. The focused RippleSet will be horizontally aligned with the corresponding rows (Fig. 6(D)). In the parallel mode (Fig. 6(b)), RippleSets are placed vertically and each root row is horizontally aligned with the corresponding RippleSet. The parallel mode provides clear alignment and thus can guide users on the detailed exploration of a relatively small number of itemsets. The compact mode and the parallel mode help users smoothly switch between the overview and the details they are interested in. In the comparison mode (Fig. 6(c)), users can conduct a side-by-side comparison of the discriminatory itemsets of two machine learning models (**G5**). As pointed out by Marx et al. [42], real-world datasets can admit ML models that have equal accuracy but assign conflicting predictions to the same input data. To help user identify these conflicting predictions and compare the fairness of two models, two RippleSets and their corresponding root rows are horizontally aligned if they have the same resolving attribute values.

Table 1. Tasks in the user study.

Task	Goals	Questions
T1	**G1**,**G5**	Compare the degrees of discrimination in the following settings of resolving attributes.
T2	**G2**,**G3**	Which item will face the largest discrimination?
T3	**G2**,**G4**	Which itemset has a varying degree of discrimination?
T4	**G3**,**G5**	Which item will be discriminated against by both model $m1$ and model $m2$?

Table 2. Comparing different visualizations to choose a proper baseline.

	G1	G2	G3	G4	G5
Euler	✗	✓	implicit	✓	✗
BubbleSets	✗	✓	implicit	✓	✗
ParallelSet [34]	✗	✓	only categorical attributes	✗	✗
UpSet [37]	✗	✓	only binary attributes	✓	only between itemsets
DiscriLens	✓	✓	✓	✓	✓

7 Laboratory Study

We conducted a laboratory study to assess participants' ability in using DiscriLens to analyze machine learning discrimination. Even though it is also important to separately evaluate RippleSet, we evaluate DiscriLens as a whole due to our focus on discrimination analysis.

7.1 Experimental Settings

We recruited a total of 16 participants (10 males, 4 females, age 21-30) through university mailing lists. All participants are postgraduate students in the school of engineering and have experiences in machine learning.

A within-subject design was employed for the user study. We designed four tasks to assess participants' ability to conduct discrimination analysis (**G1**-**G5**), as shown in Table 1. Each task contains 2-4 different questions in the same format. We trained models and mined discriminatory itemsets for two datasets: an adult income dataset [16] with a small number of discriminatory itemsets and a student performance dataset [6] with a large number of discriminatory itemsets. For the user study baseline, we first considered existing fairness visualization tools, such as FairVis [9], FairSight [1], FairLearn [36], and TensorFlow Fairness Indicator [56]. However, it is unfair to compare these tools with DiscriLens since they focus on different aspects and have different design goals, as discussed in Sect. 2.2. On the contrary, we compare DiscriLens with other visualizations that can illustrate set intersections based on the survey results in [30]. The result is shown in Table 2. DiscriLens, as a combination of Euler-based and Matrix-based

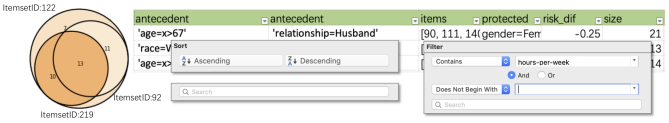

Fig. 9. The baseline system used in the user study: Euler diagram + table (with search, sort, and filter functions).

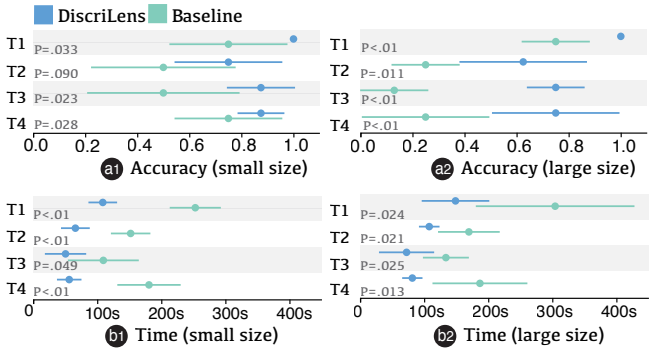

Fig. 10. Accuracy and completion time under different conditions. Error bars indicate 95% confidence intervals.

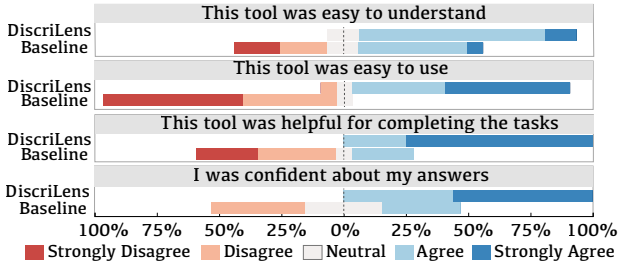

Fig. 11. Comparison of DiscriLens and the baseline based on the post-study questionnaires.

set visualization, supports a more detailed analysis of discrimination. Since no existing set visualization can meet all the requirements of discrimination analysis, we used Euler diagram + table (with search, sort, and filter functions) as the baseline method (Fig. 9).

For each participant, two datasets were randomly associated with the two conditions (DiscriLens and baseline) and were presented in a counter-balanced order. Before the formal study, each participant received a 20-minutes tutorial, when they learned the tool, completed trial tasks, and freely asked questions. In each condition, participants completed the four tasks. Participants are randomly ordered in this user study. Finally, each participant completed a post-study questionnaire and received a short informal interview.

We set a threshold at 0.05 for p value and hypothesized that, compared with the baseline, **H1**) DiscriLens is harder to understand; **H2**) DiscriLens doesn't decrease the accuracy of tasks; **H3**) DiscriLens decreases the completion time of tasks.

7.2 Results

We summarized the results of the performed tasks in Fig. 10 and the results of the post-study questionnaire in Fig. 11.

Contrary to our expectation, participants thought DiscriLens was easier to understand than the baseline (Fig. 11), rejecting **H1**. Although most (10/16) participants expressed that *"the tool seems complicated at first glance"*, they stated that *"the visual encodings are actually easy to understand and remember once you explained them"* and *"the tool offers a better support to the analysis than the baseline"*. This highlighted the importance of designing visualizations suitable for the analysis. Complicated visual designs are sometimes inevitable for the analysis of complicated data. A simple visualization can be hard to understand when it fails to support the analysis.

As shown in Fig. 10(a), except T2 in small size ($p = 0.09$), participant gave significantly more accurate answers ($p < .05$) when using DiscriLens. This result supports **H2**. Note that the overlap between accuracy confidence intervals decreased with the increase of data size. This phenomenon might indicate that DiscriLens has a more significant advantage in accuracy when analyzing a large volume of discrimination.

As shown in Fig. 10(b), the time cost of DiscriLens is significantly less than that of the baseline ($p < .05$), which supports **H3**. Unlike accuracy, the overlap between time confidence intervals increased with the increase of data size. This was because participants tended to give up some tasks when they found these tasks were difficult to complete using the baseline, which led to short completion time.

Meanwhile, we observed that participants employed different strategies with the two tools. When using the baseline to answer a question, most participants (12/16) switched back and forth between different question options and compare them carefully using the baseline tool. When using DiscriLens, most participants (10/16) first read the question, then examined the visualization to conclude an answer, and finally chose an option directly in the questionnaire.

8 Use Cases

In addition to the laboratory study, we further demonstrated the effectiveness of DiscriLens in analyzing algorithmic discrimination through use cases. The cases were conducted in collaboration with two machine learning experts (E1 and E2) and one domain experts: a professor with more than ten years of teaching experience (E3).

We mainly used the xAPI dataset [6] for demonstration and more use cases are available in the supplementary material. Each data point in the xAPI datasets has 9 attributes (*e.g.*, `raised hands`, `absence days`) of a student and a binary label indicating whether the test score of this student is over 69. We set `gender=female` as the protected group and $\tau = 0.25$. Six different types of ML models were trained: XGBoost, k-nearest neighbor (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), and decision tree (DT). The hyperparameters of all six models have been tuned to maximize the 5-fold cross-validation accuracy using AutoML [7, 59]. More implementation details are available in the supplementary material.

Identify Discrimination. E3 conducted analysis on the xAPI dataset and set the resolving attributes at first. Based on his domain knowledge and the observation of item distribution on different attributes (Fig. 6(A1)), E3 was satisfied with the resolving attributes suggested by the discovery module set `[announcements view,raised hands,absence days, relationship]` and did not modify them. He then analyzed the XGBoost model, which had the highest 5-fold cross-validation accuracy.

In the training data, female and male students had a different proportion of high scores. However, E3 observed no discriminatory itemsets in the scatter plot. He deduced this phenomenon can be explained by the correlation between `gender` and resolving attributes. For example, female students might have lower values of `absence days`. He then verified his conjecture by dragging `absence days` and `relationship` from the resolving attributes. Discriminatory itemsets then appeared in the scatter plot, which confirmed his conjecture.

E3 then used the trained models to predict unforeseen samples. As shown in Fig. 6, a number of discriminatory itemsets were identified and presented, indicating the model discrimination appeared in unforeseen samples. E3 also found that the majority of the discriminatory itemsets were orange (Fig. 12(c)), which represented discrimination towards the non-protected group (*i.e.*, male students).

Moreover, E2 conducted the same analysis on the Adult dataset [16]. The Adult dataset contains 45,222 data items. The task of this dataset is to predict whether the annual income of a person is over 50k USD. Interestingly, for models trained on the Adult dataset, discrimination did not increase in test data. One possible explanation of this phenomenon is that models trained on large datasets are more stable and make more consistent predictions. Another possible explanation is that the test data and training data had a very similar distribution of the Adult dataset, while the opposite goes for xAPI dataset.

Understand Discrimination among Itemsets. E2 was interested in the XGBoost result, which had the highest 5-fold cross validation accuracy. He examined the discrimination in different itemsets at XGBoost. In the attribute matrix, the solid parts in rectangles represented the condition of discrimination. E2 found that all discriminatory items had the condition `absence days:>7` (Fig. 12A). In other words, if a student was absent less than 7 days, this student was less likely to be treated

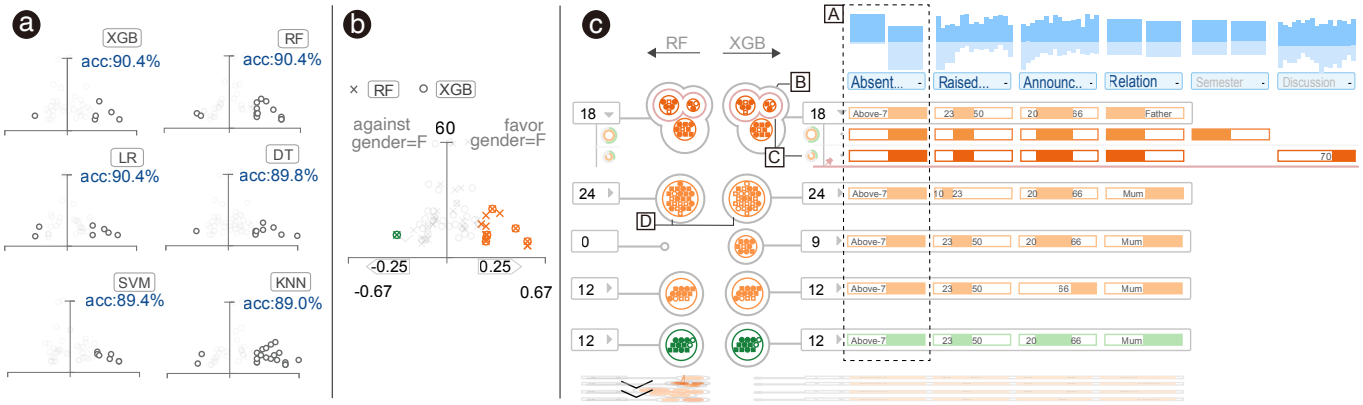

Fig. 12. Use DiscriLens to analyze discrimination: (a) analyze the discriminatory itemsets of different models using the scatter plots; (b) compare RF with XGBoost in a scatter plot; (c) compare RF with XGBoost in the comparison mode.

differently based on gender. This observation can be easily explained by the histogram of `absence day`, which indicated that almost all student whose `absence day<7` were predicted as the high-score class.

E2 then compared the RippleSets of different itemsets. As shown in Fig. 12B, among all RippleSets, the RippleSet of `[absence days:>7, announcements view:20-66, raised hands:23-50, relationship=father]` had the highest color saturation, indicating the largest value of RD. Meanwhile, this RippleSet consisted of three clusters of dots with different saturation, indicating that the RD value varies. For a detailed examination, E2 clicked on this RippleSet and expanded the corresponding root row. As shown in Fig. 12(C), the row with higher saturation indicated that, inside the itemset `[absence days:>7, announcements view:20-66,raised hands: 23-50, relationship:father]`, male students whose `discussion>70` were more likely to be treated unequally.

These observations reveal when the model predictions are potentially discriminatory and should be checked and modified by human experts or discrimination-removal methods.

Compare Discrimination among Models. Instead of directly applying XGBoost, the model with the highest accuracy, E1 compared this model with other models from a fairness perspective. According to the scatter plots (Fig. 12(a)), DT and KNN had itemsets with higher values of risk difference (*i.e.*, points with larger y-position) than XGBoost. Therefore, these two models were first excluded.

E1 then compared XGBoost with RF, the model with the second highest accuracy. Overall, RF had more discriminatory itemsets than XGBoost. An interesting finding was that four of the five discriminatory itemsets in XGBoost also existed in RF: coincident points in the scatter plot corresponded to itemsets with the same size and the same degree of discrimination (Fig. 12(b)); aligned rows in attribute matrix indicated the same condition of discrimination (Fig. 12(c)). For these four discriminatory itemsets, the only difference was in the itemset `[absence days:>7, announcements view:20-66, raised hands:10-23, relationship:mom]`, where the RippleSet of RF has more hollow items than XGBoost (Fig. 12(D)). In other words, XGBoost was more likely to predict students in this itemset as the low-score class. Considering that most discriminatory itemsets in XGBoost also existed in RF and that RF had more discriminatory itemsets, E1 concluded that XGBoost was a better choice than RF even from the fairness perspective.

E1 also compared XGBoost with LR and SVM. He found these models had discrimination towards different itemsets, as most of their RippleSets were not aligned. Therefore, comparing the discrimination of these models depended on which groups of students were more important for the model user. E1 commented, *"Without visualization, all these differences might just be obscured in a summary statistic."*

Remove Discrimination. E1 adopted Reject Option [28], a popular

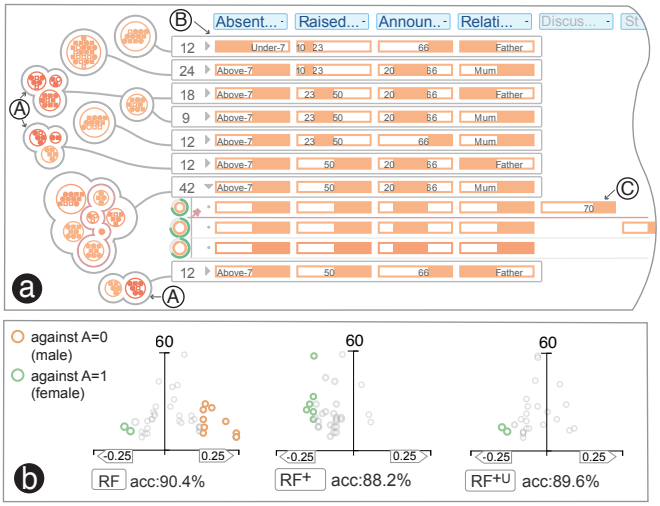

Fig. 13. (a) Users can identify discrimination with high values of RD (A) and discrimination towards critical groups (BC). (b) Compare the discriminatory itemsets of the original RF model, the RF model using Reject Option (RF^{+}), and the RF model using the user-defined Reject Option (RF^{+U}).

discrimination-removal algorithm [1], to remove discrimination in RF, a model with relatively more complex discriminatory itemsets.

One solution is to apply Reject Option to all discriminatory itemsets whose $RD > \tau = 0.25$. However, like most discrimination-removal methods [35, 55], this operation led to unexpected reverse discrimination (*i.e.*, towards female students) and a decrease in accuracy. One possible solution to this problem is to increase τ and apply Reject Option to discrimination with higher RD values.

However, in practice, severe discrimination is not only those with high RD values but also those towards critical groups. For example, E3 was concerned with the discrimination happened to hard-working students (*e.g.*, low absenteeism, active discussion). Therefore, E1 employed DiscriLens to identify severe discrimination. Discriminatory itemsets with high RD can be easily identified through their colors (Fig. 13(A)). E1 then checked the discrimination towards critical groups. He first examined the students with low absenteeism and selected the itemset whose `absent days<7` (Fig. 13(B)). He did not open the row for further examination since the corresponding RippeleSet, with only one cluster of items, indicated there was no complex intersection. E1 then checked the students with active class participation by examining the values of `raised hands`. Among the three RippleSets with high values (> 50) of `raised hands`, two had high values of RD and only

[1]The implementation by IBM Fairness 360 [25] is used here.

one (Fig. 13(C)) was left to be examined. E1 expanded the rows of this RippleSet and selected an itemset (Fig. 6(F)) that also had a high value (> 70) of discussion.

Reject Option was then applied to the user-defined discriminatory region. For convenience, we denoted RF using default discriminatory region by RF^{+}, and RF using the user-defined discriminatory region by RF^{+U}. As shown in Fig. 13(b), compared with RF^{+}, RF^{+U} reduced both the accuracy loss and the reverse discrimination ○. By providing more detailed information, DiscriLens enables users a precise removal of discrimination.

Expert Interview All the experts expressed great enthusiasm for DiscriLens and offered useful suggestions for its improvement. They commented that it helped them *"understand ML models from a novel perspective"* (E3) and could *"contribute to the adoption of ML to real-world applications"* (E1). E1 suggested combining the analysis of model accuracy with the analysis of discrimination. E2 was curious about the comparison between algorithmic discrimination and human discrimination, *"Even though a model makes discriminatory predictions, it can still be less biased than humans."* E3 suggested adding a *"what-if"* function to check whether and how a model would discriminate against a given data item.

The issue of trust was mentioned and discussed by the experts. E3 suspected that the training data might omit some attributes related to the students performance. He also commented that the results of discrimination analysis were influenced by the choices of the analysts (*e.g.*, the setting of key attributes). *"Two analysts might draw opposite conclusions."* E1, the machine learning expert, suggested the comparison between algorithmic discrimination and human discrimination to increase user confidence. *"Even though a model makes discriminatory predictions, it can still be less biased than humans and thus be helpful."*

9 Discussion

9.1 Scalability

For the discrimination discovery, it takes about six minutes to run the four-stage pipeline on 4,000 samples with 14 features on a PC (2.3GHz dual-core, Intel Core i5 processor). The major bottlenecks lie in the FEGS algorithm (two minutes), the FP-Growth algorithm (one minute), and the discriminatory rules mining (three minutes).

The scalability of Rippleset is mainly limited by the number of sets and the number of items. According to our users study, Rippleset effective reduces the visual clutter in traditional Euler diagram, which has difficulties to handle more than three sets. However, the readability of RippleSet hasn't been validated with more than seven intertwining sets. For a large number of items, RippleSet supports the aggregation of items as shown in Fig. 6(D1). In the future, we plan to support more grouping strategies, such as those discussed in Squares [51], and provide more statistical summaries in the aggregation view of RippleSet.

9.2 Complicated Visualization Design

We acknowledge that the novel and complicated visualization designs in DiscriLens can pose challenges to users, especially those who have no prior knowledge in visual analytics. Even though we have provided interactive tutorials and legends to alleviate this issue, the steep learning curve still exists and can lead to low accessibility to novices. For example, RippleSet and the attribute matrix can be overwhelming when there are too many itemsets. Users may stop at the scatter plot, blindly trying to optimize RD without conducting detailed visual analysis of discrimination. On the other hand, complicated visual designs are sometimes inevitable for the analysis of complicated data. Even with easy-to-understand encodings, a simple visualization can be hard to use when it fails to support the required analysis. How to strike a good balance between designing intuitive visualizations and accomplishing complex analysis tasks is still an open question in the field of visual analytics and requires further investigation.

9.3 Evaluation of RippleSet

In this paper, we proposed a novel set visualization, RippleSet, to assist the analysis of discrimination. RippleSet supplements the attribute matrix by illustrating the discrimination distribution among data items, especially these belong to intricately intertwined sets. While the current evaluation shows the utility of DiscriLens as a whole, it fails to demonstrate the contribution of RippleSet to the effectiveness of DiscriLens. Apart from discrimination analysis, RippleSet also shows the potential to stand alone and be applied for more general tasks. Unfortunately, due to page limits and our focus on discrimination analysis, we are unable to provide a stand-alone evaluation of RippleSet in this paper.

9.4 Subjectivity in Analysis

In DiscriLens, we allow users to customize the definition of discrimination and support the integration of human domain knowledge. While this feature is regarded as a strength by the interviewed experts, we also admit that user customization can act as a double-edged sword in discrimination analysis.

On the one hand, user customization enables the integration of domain knowledge. Since the definition of discrimination differs across domains, the involvement of domain knowledge enables a more comprehensive analysis. On the other hand, additional bias can be potentially imposed by the subjective choices of the analysts. Different customization may lead to different analysis results. One possible solution is to support what-if analysis and cross-validate the analysis results of different customization. Analysts can test and compare different customization before drawing the final conclusion. Similar methods are commonly used in verifying subjective analysis.

9.5 More Intelligent Discrimination Mining

In this work, we do not consider the hidden attributes in the causal graph(*i.e.*, attributes that cannot be explicitly observed). We assume the key attributes can capture the main model decisions and set a threshold of the risk difference to allow oscillations caused by hidden attributes. However, the protected or resolving attributes may not be included in the dataset and this assumption sometimes can fail. In future work, we plan to take the hidden attributes into consideration and offer more comprehensive explanations of the model predictions.

Meanwhile, the current version of DiscriLens only supports the analysis of one protected attribute and requires users to define the protected group as input. When there are many groups (*e.g.*, females, blacks, LGBTs) qualified for special protection by a law, users cannot analyze multiple protected attributes at the same time and need to examine different attributes separately one by one.

Compared with other explainable models (*e.g.*, decision tree), our study provides limited support in explaining the identified discrimination. Therefore, an interesting future direction is to integrate discrimination analysis with model interpretation and diagnosis techniques, which can help us track the origin of algorithmic discrimination and better understand model behaviors.

10 Conclusion

In this work, we designed and developed DiscriLens, an interactive visualization tool that facilitated a better understanding and analysis of algorithmic discrimination. A four-stage pipeline was developed for the discovery of discriminatory predictions. For an effective presentation, a novel set visualization was designed by combining an extended Euler diagram with a matrix-based set visualization. Two case studies demonstrated the usability and utility of DiscriLens in understanding and removing algorithmic discrimination. Context-aware Reject Option, a post-processing method, was proposed to better remove discrimination while reducing the accuracy loss. We also reported insights into algorithmic discrimination that were obtained during the development and evaluation of DiscriLens.

References

[1] Y. Ahn and Y.-R. Lin. Fairsight: Visual analytics for fairness in decision making. *arXiv preprint arXiv:1908.00176*, 2019.

[2] B. Alper, N. Riche, G. Ramos, and M. Czerwinski. Design study of linesets, a novel set visualization technique. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2259–2267, 2011.

[3] B. Alsallakh, W. Aigner, S. Miksch, and H. Hauser. Radial sets: Interactive visual analysis of large overlapping sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2496–2505, 2013.

[4] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. In *Eurographics Conference on Visualization, EuroVis*, pp. 1–21. Eurographics Association, Swansea, UK, 2014.

[5] B. Alsallakh and L. Ren. Powerset: A comprehensive visualization of set intersections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):361–370, 2017.

[6] E. A. Amrieh, T. Hamtini, and I. Aljarah. Mining educational data to predict students academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136, 2016.

[7] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[8] G. Big Picture. Attacking discrimination in ML. `http://research.google.com/bigpicture/attacking-discrimination-in-ml/`, 2017. Accessed: 2019-06-30.

[9] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. *arXiv preprint arXiv:1904.05419*, 2019.

[10] C. Chen, J. Yuan, Y. Lu, Y. Liu, H. Su, S. Yuan, and S. Liu. Oodanalyzer: Interactive analysis of out-of-distribution samples. *IEEE Transactions on Visualization and Computer Graphics*, 2020. to be published. doi: 10.1109/TVCG.2020.2973258

[11] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[12] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1009–1016, 2009.

[13] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, Halifax, NS, Canada, 2017.

[14] Council of European Union. Equal Employment Opportunity Commission, 1978. `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32006L0054`.

[15] K. Dinkla, M. J. van Kreveld, B. Speckmann, and M. A. Westenberg. Kelp diagrams: Point set membership visualization. *Computer Graphics Forum*, 31(3pt1):875–884, 2012.

[16] D. Dua and C. Graff. UCI machine learning repository, 2017.

[17] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226. ACM, New York, NY, USA, 2012.

[18] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *13th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1022–1027. Morgan Kaufmann, Chambéry, France, 1993.

[19] W. Freiler, K. Matkovic, and H. Hauser. Interactive visual analysis of set-typed data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1340–1347, 2008.

[20] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[21] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2125–2126. ACM, San Francisco, CA, USA, 2016.

[22] J. Han and J. Pei. Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD Explorations Newsletter*, 2(2):14–20, 2000.

[23] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323. Curran Associates, Barcelona, Spain, 2016.

[24] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 600. ACM, Glasgow, UK, 2019.

[25] IBM. AI fairness 360. `https://aif360.mybluemix.net/`, 2018. Accessed: 2019-06-30.

[26] P. H. G. Ivan Koswara, Ananya Aaniya. Simpson's paradox. `http://https://brilliant.org/wiki/simpsons-paradox/`. Accessed: 2019-12-30.

[27] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[28] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, Brussels, 2012.

[29] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, Bristol, UK, 2012.

[30] Keshif. Visual techniques for analyzing set-typed data. `https://gallery.keshif.me/setvis`, 2010. Accessed: 2019-07-30.

[31] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11):2767–2787, 2010.

[32] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 656–666. Curran Associates, Long Beach, California, USA, 2017.

[33] B. Kim, B. Lee, and J. Seo. Visualizing set concordance with permutation matrices and fan diagrams. *Interacting with Computers*, 19(5-6):630–643, 2007.

[34] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

[35] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076. Curran Associates, Long Beach, CA, USA, 2017.

[36] F. Learn. Fair learn. `https://github.com/fairlearn/fairlearn`, 2019.

[37] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014.

[38] C. H. Liu. Form symbolism, analogy, and metaphor. *Psychonomic Bulletin & Review*, 4(4):546–551, 1997.

[39] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):91–100, 2017.

[40] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 502–510. ACM, San Diego, USA, 2011.

[41] K. Mancuhan and C. Clifton. Combating discrimination using bayesian networks. *Artificial intelligence and law*, 22(2):211–238, 2014.

[42] C. T. Marx, F. d. P. Calmon, and B. Ustun. Predictive multiplicity in classification. *arXiv preprint arXiv:1909.06677*, 2019.

[43] Y. Ming, H. Qu, and E. Bertini. Rulematrix: visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, 25(1):342–352, 2018.

[44] K. C. Naff. Subjective vs. objective discrimination in government: Adding to the picture of barriers to the advancement of women. *Political Research Quarterly*, 48(3):535–557, 1995.

[45] Parliament of the United Kingdom. Sex Discrimination Act 1975, 1975. `https://www.legislation.gov.uk/ukpga/1975/65`.

[46] Parliament of the United Kingdom. Equality Act 2010, 2010. `http://www.legislation.gov.uk/ukpga/2010/15`.

[47] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 581–592. SIAM, Nevada, US, 2009.

[48] D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 126–131. ACM, Riva, Italy, 2012.

[49] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568. ACM, Las Vegas, USA, 2008.

[50] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi, and A. I. Madbouly.

A comparative analysis of classification algorithms for students college enrollment approval using data mining. In *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments*, p. 106. ACM, Albacete, Spain, 2014.

[51] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):61–70, 2017.

[52] N. H. Riche and T. Dwyer. Untangling euler diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1090–1099, 2010.

[53] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

[54] P. Simonetto, D. Auber, and D. Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum*, 28(3):967–974, 2009.

[55] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2239–2248. ACM, London, UK, 2018.

[56] TensorFlow. Fairness indicators. `https://github.com/tensorflow/fairness-indicators`, 2019.

[57] The United State. Equal Employment Opportunity Commission, 2010. `https://www.law.cornell.edu/cfr/text/29/1607.4`.

[58] U.S. Federal Legislation. Civil Rights Act of 1991, 1991. `https://www.eeoc.gov/eeoc/history/35th/1990s/civilrights.html`.

[59] Q. Wang, Y. Ming, Z. Jin, Q. Shen, D. Liu, M. J. Smith, K. Veeramachaneni, and H. Qu. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems, CHI*, p. 681. ACM, Glasgow, Scotland, UK, 2019.

[60] W. Wang, H. Wang, G. Dai, and H. Wang. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 517–520. ACM, Montréal, Québec, Canada, 2006.

[61] X. Yin and J. Han. Cpar: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pp. 331–335. SIAM, San Francisco, CA, USA, 2003.

[62] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[63] L. Zhang, Y. Wu, and X. Wu. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344. ACM, Halifax, NS, Canada, 2017.

[64] L. Zhang, Y. Wu, and X. Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935. AAAI Press, San Francisco, USA, 2017.

[65] L. Zhang, Y. Wu, and X. Wu. Achieving non-discrimination in prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3097–3103. ijcai.org, Stockholm,Sweden, 2018.

[66] Q. Zhang, W. Wang, and S. Zhu. Examining CNN representations with respect to dataset bias. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence*, pp. 4464–4473. AAAI press, New Orleans, Louisiana, USA, 2018.

[67] I. Žliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *11th International Conference on Data Mining (ICDM)*, pp. 992–1001. IEEE, Vancouver, BC, Canada, 2011.