Hierarchical Clustering for Discrimination Discovery: A Top-Down Approach

Neda Nasiriani[†], Anna Squicciarini[‡], Zara Saldanha[†], Sanchit Goel[‡], Nicola Zannone[*]

[†]Bucknell University
[‡]Pennsylvania State University
[*]Eindhoven University of Technology

***Abstract*—Today, data is an essential part of many decision-making processes in businesses and social life through the use of various machine learning techniques. These methods can easily perpetuate human bias in the data and result in discrimination. Despite a growing interest in data discrimination discovery and removal, to date there is a lack of a general and robust framework to distinguish discriminatory decision-making processes from non-discriminatory ones. In this work, we present a generic framework that helps detect possible discrimination by analyzing historical data and associated decisions using a top-down unsupervised approach, which we refer to as hierarchical clustering. Our approach is highly adaptive as it gradually "learns" users' inherent groups, and clusters their records using cohesiveness and density of points in the dataset. Moreover, we propose a progressive attribute-selection method to choose statistically relevant attributes, thus reducing the effect of noise. Finally, we adopt a recursive notion of cluster profile that is *homogeneous* w.r.t. decision labels. This allows for deeper insights on the data and on the decision-making underlying the final user classification. Our framework is able to identify both positive and negative bias resulting in discrimination. We also highlight patterns of discrimination revealed by the homogeneous cluster centroids, which otherwise could not be captured.**

I. Introduction

Today, data is an essential part of many decision-making processes in businesses and social life, *e.g.*, credit card applications, life insurance and housing listing, through the use of various machine learning techniques. These methods can easily perpetuate human bias in the data and result in discrimination [3], [18], [19], [20], *i.e.*, the unfair treatment of an individual or a group based on its membership to a certain social category rather than on individual merits.

In order to alleviate built-in bias in data, we need approaches to detect the existence of discrimination (discrimination discovery) and remove this bias (discrimination prevention). This has attracted a growing interest in discrimination discovery and prevention, and several performance measures have been developed to define discrimination-aware machine learning models (*e.g.*, [8], [9], [12]). However, to date there is a lack of a general and robust framework to distinguish discriminatory decision-making processes from non-discriminatory ones.

Situation testing has been proposed as a legal means to determine the existence of discrimination in decision-making processes [9]. This is performed by conducting controlled experiments, where closely related records (*e.g.*, candidates having the same qualifications for the job) and belonging to protected and non-protected groups are created and tested against the decision process under analysis. However, situation testing can only be applied when the decision-making process is known. In many machine learning applications, only historical data and associated decisions are usually available. Accordingly, it is extremely challenging to determine whether any form of discrimination against specific individuals or specific protected groups occurred. Moreover, decision-making processes that rely on a large set of attributes can become hard to analyze because of the inter-dependencies among different attributes (sensitive and non-sensitive).

In this work, we present a top-down framework that detects possible discrimination by analyzing historical data and associated decisions using a novel unsupervised approach, which we refer to as *hierarchical clustering*. Given a set of sensitive attributes characterizing protected vs. non-protected groups, our framework is able to reliably identify the degree of both *discrimination* and *reverse discrimination* by measuring positive and negative bias within statistically similar data points. By learning patterns of discrimination directly from the data, with no direct use of the decision label, we aim to extract *reliable* evidence of data discrimination, considering possible correlations among sensitive and non-sensitive attributes describing users' demographics and other characteristics.

We are particularly interested in the detection of *non-explanatory* discrimination. We consider a decision process to be affected by non-explanatory discrimination if one of the following conditions holds: *(a)* there is strong statistical evidence that sensitive attributes were used for decision making or *(b)* there are other groups of users with similar non-sensitive attributes for which the decision is significantly different.

Distinguishing explanatory discrimination from non-explanatory discrimination is a challenging task, which we tackle in this work. For instance, Kamiran and colleagues [8] assume that explanatory attributes are specified by a domain expert, *i.e.*, they are known a priori. In our work, we relax any assumption of "knowledge" on the dataset, and use the data itself to extract explainable attributes and their level of discrimination. In particular, we propose a generic pre-processing approach to distinguish between explanatory and non-explanatory discrimination by studying the impact of

attributes on the decision label and the correlation between sensitive and non-sensitive attributes. As a result of this pre-processing phase, we can study the discrimination in the data based on a reliable set of attributes. An important assumption is that, at least to some extent, the class label decision *was made* based on the information in the dataset, particularly the non-sensitive important attributes, referred to as *explanatory attributes*.

Our approach performs an unsupervised learning over the input dataset to identify potentially discriminated user groups. We gradually "learn" users' inherent groups, and cluster their records using cohesiveness and density of points in the dataset. This avoids reliance on a fixed number of neighbors for clustering, which may affect the quality of the identified groups. Moreover, we propose a progressive attribute-selection method that uses statically relevant attributes to reduce the effect of noise. We evaluate our approach over three real-world datasets, and compare results with a known method as a baseline. We show that our approach is not only capable of identifying several instances of discrimination – even when numerically small number of records are considered – but also that we can do so efficiently, using only small amount of relevant information.

Our contributions are multi-fold:

- We present a general framework for discrimination discovery using a top-down hierarchical clustering of data. Our approach does not require any specific knowledge on the classification model and decision-making process used for classification.
- We propose an effective attribute-selection and pre-processing mechanism to choose relevant *explanatory* attributes for our analysis.
- As a byproduct of our approach, we are able to (statistically) profile individuals based on their non-sensitive features (*e.g.*, qualifications, history). This helps gain insights on the data and the possible discrimination that may have occurred.
- We are able to detect evidence of both positive and negative bias even for unbalanced datasets, provided that statistical significance is maintained.

The remainder of the paper is organized as follows. Section II provides an overview of existing discrimination measures and introduces the metrics used in this work. Section III presents our approach and Section IV presents its evaluation. Finally, Section V discusses related work and Section VI concludes the paper.

II. Discrimination Measures

From a data-centric point of view, the input of a decision making process is a dataset of n-tuples (records) describing users' demographics, with one or more sensitive attributes (the actual semantic of the attributes depends on the application domain), and a related decision label, which is generally either "favorable" or "non-favorable". Discrimination occurs when (a group of) individuals receive an unfair treatment because of their membership to a certain social category rather than on individual merits. Here, we are particularly interested in *targeted* discrimination as it provides specific reasons and insights on the nature of discrimination, as opposed to aggregate measures, *e.g.*, [3], [8], [18], [19]. Other forms of discrimination, *e.g.*, taste-based discrimination [8], are beyond the scope of this work.

Discrimination may or may not be explainable [8]. *Non-explanatory* discrimination occurs when a group of closely similar individuals are treated differently apparently only because of their membership to a protected group. On the other hand, *explanatory* discrimination is the result of high correlation between a sensitive attribute (or a set of sensitive attributes) and a non-sensitive attribute (or a set of non-sensitive attributes) that has a high impact on the decision making process (hereafter called *explanatory attribute*), *e.g.*, a larger portion of the male applicants satisfy the education requirements of the job. Even in legal procedures this type of decisions cannot be appealed and reversing it can result in *reverse discrimination*, *i.e.*, records in the non-protected group receive an unequal treatment (*e.g.*, lower qualified candidates from a protected group may experience higher rates of employment/acceptance compared to individuals with similar qualifications).

Our goal is to discover discrimination regardless of the size of the discriminated group, noting that a protected group may be a very small minority in a large dataset. Hence, majority rules that are extracted from the dataset might not represent mistreatment toward this group. Thus, we characterize the dataset based on user profiles and compare the information derived from the dataset against the decision labels for possible discrimination.

A. Preliminary Notions and Metrics

Let us assume a dataset D with N *records* defined over a set of n *attributes* $\mathcal{A} = \{A_1, \ldots, A_n\}$. A record r_i is a vector $[v_{i_1}, \ldots, v_{i_n}, d_i]$ where v_{i_j} denotes the value of attribute $A_j \in \mathcal{A}$ in the i-th record and d_i is the class label representing a favorable $(+)$ or a unfavorable $(-)$ decision.

Assume that within the set of attributes, there is one or more attributes that is defined as *sensitive*. Let $\mathcal{S}$ denote the set of sensitive attributes, with $\mathcal{S} \subseteq \mathcal{A}$. We assume that some values for attributes in $\mathcal{S}$ define a minority category, *e.g.*, *female* or $age > 60$. We say that a record is a member of a *protected group* when at least one of the sensitive attributes have a minority value; otherwise we say that the record belongs to a *non-protected group*. Hereafter, we use s to represent the protected group and $\neg s$ to represent the non-protected group.

A dataset D can be partitioned in four quadrants based on these groups and decision label. Fig. 1 provides a visual representation of the partition, where $n_{(x,z)}$ denotes the number of records in D that belong to subgroup z, with $z \in \{s, \neg s\}$, and to which decision x is assigned, with $x \in \{+, -\}$.

In existing legal frameworks, discrimination (also called group under-representation) is typically assessed based on the proportion of samples in the dataset from the protected group (p_1) and from the non-protected group (p_2), that were

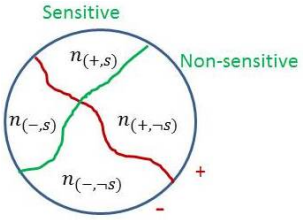


Fig. 1: Cluster representation

rejected, or more generally, received a negative decision. For example, group under-representation is measured as the difference $p_1 - p_2$ in the UK legislation [18], whereas in the US legislation the ration p_1/p_2 (called *selection lift*) is used as a measure of discrimination [20]. These measures have been formalized in [9] as follows:

$$p_1^- = \frac{n_{(-,s)}}{n_{(-,s)} + n_{(+,s)}} \tag{1}$$

$$p_2^- = \frac{n_{(-,\neg s)}}{n_{(-,\neg s)} + n_{(+,\neg s)}} \tag{2}$$

$$p^- = \frac{n_{(-,s)} + n_{(-,\neg s)}}{N} \tag{3}$$

$$diff_{(-,s)} = p_1^- - p_2^- \tag{4}$$

$$elift_{(-,s)} = \frac{p_1^-}{p^-} \tag{5}$$

Note that Eq. 5 is a normalization of p_1^- w.r.t. the proportion of samples that received a negative decision. Other notions (p^+, p_1^+, p_2^+, $diff_{(-,\neg s)}$, $elift_{(,\neg s)}$, etc.) are defined similarly.

A main limitation of the type of quantification of discrimination provided by the metrics above is the lack of information regarding smaller groups of records that are statistically similar. In other words, accounting only for the portion of negative decisions that are made for protected and non-protected groups can result in too small groups to be statistically significant, or otherwise hard to extract.

Accordingly, only relying on aggregate statistical measures may not be sufficient to detect discrimination. For example, consider a dataset of 400 records and 10 of them are in the protected group and labeled as negative. The sensitive records constitute only 2.5% of the records resulting in $p_1 = 1$, $p_2 = 0$ and $p = 0.025$, which can result in a very high value of $diff_{(-,\neg s)}$, but we really do not have any other statistically significant information about sensitive records. We cannot reliably conclude that discrimination is happening because of a high value of $diff$. Other explanations, ignored by these metrics, are possible, *e.g.*, all records from a protected group lack an attribute that is required for a positive decision.

Herein, we consider a set of *statistically similar* candidates irrespective of their membership to protected or non-protected groups for a more robust discrimination discovery (*cf.* Section III-B). To this end, we adjust these measures for a cluster-level discrimination measurement in the following section.

B. Cluster-level elift

Herein we introduce a more general set of discrimination measures in conjunction with statistical profiling of the instances. We adjust the general metrics defined in Eqs. 4-5 to account for records with a similar profile. Specifically, we introduce the notion of *cluster-level elift*. Given a cluster i, its cluster-level $elift$, denoted as $elift^{(i)}$, is defined as follows:

$$elift^{(i)}_{(+,s)} = \frac{\frac{n^{(i)}_{(+,s)}}{n^{(i)}_{(+,s)}+n^{(i)}_{(+,\neg s)}}}{r^{(i)}_s} \tag{6}$$

$$elift^{(i)}_{(-,s)} = \frac{\frac{n^{(i)}_{(-,s)}}{n^{(i)}_{(-,s)}+n^{(i)}_{(-,\neg s)}}}{r^{(i)}_s} \tag{7}$$

where the normalizing factor $r_s^{(i)}$ is defined as:

$$r^{(i)}_s = \frac{n^{(i)}_{(+,s)} + n^{(i)}_{(-,s)}}{N^{(i)}} \tag{8}$$

Intuitively, this normalizing factor measures the ratio of sensitive records in cluster i. $elift^{(i)}_{(+,\neg s)}$, $elift^{(i)}_{(-,\neg s)}$ and $r^{(i)}_{\neg s}$ are defined similarly.

The presence of discrimination in the decision making process is identified by comparing the difference in cluster-level $elift$ for each cluster. Specifically, by comparing the difference between positive decisions among protected and non-protected groups, we can have a proxy for amount of *positive bias* towards the non-protected group, as follows:

$$bias^{(i)}_{pos} = elift^{(i)}_{(+,\neg s)} - elift^{(i)}_{(+,s)}. \tag{9}$$

We also study the negative bias towards the protected group by comparing the ratio of negative decisions made for the protected vs. non-protected groups, as:

$$bias^{(i)}_{neg} = elift^{(i)}_{(-,s)} - elift^{(i)}_{(-,\neg s)}. \tag{10}$$

Positive values of $bias_{pos}$ ($bias_{neg}$) correspond to possible existence of positive (negative) bias toward non-sensitive (sensitive) records, which we count as *possible discrimination*.

To obtain reliable evidence of discrimination, discrimination should be assessed by considering clusters of statistically similar records. To this end, we cluster records in the dataset in homogeneous clusters w.r.t. the decision label. The level of homogeneity of a cluster i is defined as follows:

$$loh(i) = \frac{|n^{(i)}_{(+,\cdot)} - n^{(i)}_{(-,\cdot)}|}{N^{(i)}} \tag{11}$$

with $n^{(i)}_{(+,\cdot)} = n^{(i)}_{(+,s)} + n^{(i)}_{(+,\neg s)}$ and $n^{(i)}_{(-,\cdot)} = n^{(i)}_{(-,s)} + n^{(i)}_{(-,\neg s)}$. By imposing that clusters are homogeneous, we statistically assure that the set of records in each cluster received rather consistent treatment, and difference in their decision labels provides evidence of discrimination.

III. Data Discrimination Detection

We propose a principled approach consisting of attribute selection and top-down clustering to effectively emulate situ-

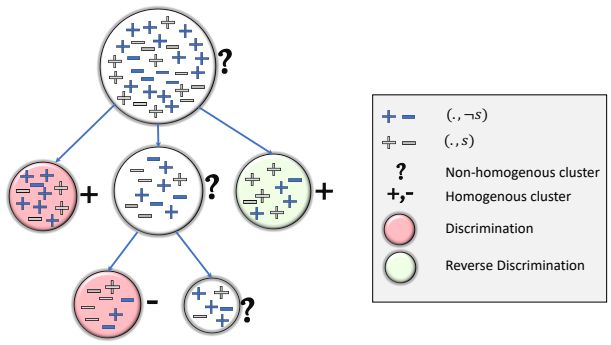


Fig. 2: Schematic of our top-down clustering approach (Hierarchical Clustering) for discrimination detection.

ational testing. A schematic representation of our hierarchical clustering approach is given in Fig. 2. We first find the high impact feature set and perform hierarchical clustering to find clusters that are homogeneously labeled (with majority decision, either positive or negative). Homogeneous clusters represent similar records that should have been treated similarly (because of the majority decision). By analyzing the proportion of different decisions based on membership to protected groups, we can mark a cluster with discrimination or reverse discrimination, respectively. We discuss both attribute selection and hierarchical clustering below.

A. Pre-processing: Attribute Selection

We first identify high-impact attributes with respect to the class label. Specifically, we study the *significance of attributes* (SoA) to be able to remove irrelevant attributes and, thus, limit noisy findings, which are likely in high-dimensional datasets. We define the significance of an attribute as the impact of the attribute on the classification of the records in a given dataset (w.r.t. class label $+, -$). Given a dataset D and a set of attributes $A \subseteq \mathcal{A}$, let $r_{|A}$ denotes the projection of a record $r \in D$ on A and $D_{|A}$ the multi-set including all $r_{|A}$ such that $r \in D$. The significance of an attribute A_i is defined as

$$\eta_{A_i} = \frac{accuracy_{D_{|A}}}{accuracy_{D_{|A \setminus \{A_i\}}}}$$

where $accuracy_{D_{|A}}$ refers to the accuracy of classification of any supervised method trained and tested on $D_{|A}$ with feature vector A. η_{A_i} equal to 1 indicates that the removal of attribute A_i does not affect the accuracy of the supervised classification and hence can be considered as a low significance attribute. Values larger (smaller) than one exhibit loss (gain) in accuracy when A_i is removed and hence higher (lower) significance for attribute A_i can be concluded based on this method. Moreover, to avoid redlining effects, which happen as a result of high correlation between non-sensitive and sensitive attributes, we study correlation among attributes using some correlation metric, *e.g.*, Normalized Mutual Information (NMI) [10]. If such dependencies are not isolated and removed, non-explanatory discrimination can be wrongly detected. Combining the two notions above, we define high-impact attributes as follows.

Definition 1 (High Impact Attributes): Given a dataset D, an attribute $A_i \in \mathcal{A} \setminus \mathcal{S}$ has *high impact* on D if the following conditions hold:

1) $\eta_{A_i} > \alpha$ with respect to D, where α is a threshold indicating the minimum level of significance for the dataset.
2) There does not exist a sensitive attribute $S_j \in \mathcal{S}$ such that $\sigma(A_i, S_j) > \beta$ with respect to D, where $\sigma(\cdot, \cdot)$ is a generic correlation function and β is a threshold indicating the minimum level of correlation required.

We extend this definition to attribute sets and call *high impact feature set*, denoted as $\mathcal{H}$, the set of all high impact attributes.

B. Discrimination Detection with Hierarchical Clustering

The use of the group under-representation measures, discussed in Section II-A, does not provide reliable evidence of discrimination and can be opposable in a court of law. In fact, non-sensitive but relevant variables can be used to explain the gap in the decisions (*i.e.*, the class label). For example, consider a set of candidates for a given job, where discrimination law prohibits the use of attributes such as gender and race for decision making. The aggregate measure *diff* in Eq. 4 considers the decisions made w.r.t. the membership of records to the protected or non-protected group, but does not reveal much about the discrimination among candidates with similar qualifications, *i.e.*, attributes that help assign the class label but that are *not* sensitive (*e.g.*, qualification).

In order to mitigate this issue, we adopt a *top-down* approach, by analyzing statistically close records (in terms of their attributes' values) and study cluster-level discrimination measures (*cf.* Section II-B). Herein, we present our hierarchical clustering algorithm for discrimination discovery, as shown in Algorithm 1. Upon selecting the high impact feature set $\mathcal{H}$ as explained in Section III-A, our hierarchical clustering method clusters the tuples $D_{|\mathcal{H}}$ recursively until homogeneous clusters (w.r.t. the class label, as defined in Eq. 11) are found. Intuitively, the level of homogeneity represents the existence of a majority decision for a specific cluster. This homogeneity is constrained by a threshold θ, which can be chosen based on different dataset-specific criteria. Homogeneity is used in Algorithm 1 as a stopping criteria, where recursion occurs until a sufficient level of homogeneity (greater than θ) is met (line 8). Note that the cluster size should be large enough to enable meaningful statistical similarity among the tuples. This is specified by threshold λ in Algorithm 1 (line 3), which can be adjusted based on the dataset at hand and the level of statistical reliance.

Recall from Section II, that we use the difference in the values of *elift* as a proxy for amount of possible mistreatment toward a given group. We calculate the difference in the ration of both positive and negative decisions made in each cluster with $bias_{pos}$ (Eq. 6) and $bias_{neg}$ (Eq. 7) measures, respectively. Positive values of $bias_{pos}$ ($bias_{neg}$) correspond to possible existence of positive (negative) bias toward non-sensitive (sensitive) records, which can be accounted for as *possible discrimination*.

Algorithm 1 Hierarchical Clustering

Global variables: θ, λ
procedure HCLUSTERING($data$)
  if size($data$) $\leq \lambda$ **then**
    terminate
  else
    $num_{cluster} \leftarrow$ CLUSTER-NUMBER($data$)
    for i from 1 to $num_{cluster}$ **do**
      if $loh(i) > \theta$ **then** ▷ Checks for homogeneity
        $bias^{(i)}_{pos} \leftarrow elift^{(i)}_{(+,\neg s)} - elift^{(i)}_{(+,s)}$
        $bias^{(i)}_{neg} \leftarrow elift^{(i)}_{(-,s)} - elift^{(i)}_{(-,\neg s)}$
        if $(bias^{(i)}_{pos} > 0)$ **then**
          Positive bias toward non-protected group
        if $(bias^{(i)}_{neg} > 0)$ **then**
          Negative bias toward protected group
        else
          HClustering($data^{(i)}$)

Algorithm 2 Cluster Number Finder

Global variables: M, τ
procedure CLUSTER-NUMBER($data$)
  $k \leftarrow 2, num_{cluster} \leftarrow 2, CoV_{final} \leftarrow -\inf$
  while $k \leq M$ **do**
    $clusters =$ K-Means$(data, k)$
    $sil_{val} = Silhouette(cluster)$
    for i from 1 to $|clusters|$ **do**
      $sil_i = sil_{val}[i]$
      if $std(sil_i) \leq \tau$ **then**
        $sil_{cov}[i] = \tau / mean(sil_i)$
      else
        $sil_{cov}[i] = std(sil_i)/mean(sil_i)$
    $CoV_n = std(sil_{cov})/mean(sil_{cov})$
    if $CoV_n \leq CoV_{final}$ **then**
      $num_{cluster} \leftarrow k$
      $CoV_{final} \leftarrow CoV_n$
    $k = k + 1$
  return $num_{cluster}$

C. Effective Number of Clusters

One challenge of unsupervised clustering is to find the effective number of clusters exhibiting different vicinity (group of similar records) with acceptable cohesiveness and separation. We deal with this problem at each level of the hierarchical clustering and hence propose a *general* and *automated* approach to find the effective number of clusters based on the data at hand. In particular, we use the *silhouette* measure [17], which is based on cohesion and separation of instances under a specified number of clusters. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A uniform distribution of silhouette values corresponds to a better clustering in terms of cohesiveness and separation. We use Covariance of Variation (CoV), a standardized measure of dispersion. The lower the CoV, the higher is the variation in respect to the average value. Therefore, we choose the number of clusters resulting in minimum CoV of silhouette values.

The actual number of clusters is computed using Algorithm 2. This algorithm tries different number of clusters up to M (pre-defined constant) and chooses the best number of clusters based on the CoV of the silhouette of different clusters. We define the silhouette values of cluster i as sil_i, and define the CoV of those values, denoted as sil_{cov} as follows:

$$sil_{cov}[i] = CoV(sil_i) = \frac{std(sil_i)}{mean(sil_i)}$$

Then, we apply the same principle to all clusters and calculate the CoV of their silhouette values. Algorithm 2 chooses the cluster number n with minimum CoV.

IV. Evaluation

This section presents selective experimental results using real-world datasets. We compare our results to an existing framework based on k-NN [9], which, similarly to ours, is unsupervised and can operate on high impact attribute sets.

A. Experiment Setup

Datasets & Settings: For our experiments, we use three datasets: (i) German Credit (GC), (ii) Communities and Crime (CC), and (iii) Census Income (CI) datasets. These datasets present high discrimination potential, as studied in prior works [8], [9], [11]. Summative statistics of each dataset are given in Table I, along with the experimental settings.

The GC dataset contains credit applicants in Germany with 20 attributes where 7 are numerical and the others are categorical. The dataset includes two sensitive attributes, *i.e.*, PersStat and Age. For our analysis, we consider records with PersStat = divorced, separated, married female as sensitive. The class label is whether credit is considered Good or Bad.

The CC dataset contains 128 attributes, which are all numerical except one. The dataset comprises 1994 records describing the crime rate of different communities in the US. We used the discretized version of the dataset given in [8]. An attribute race is added to the dataset based on a threshold of 0.06 for attribute racepctblack. The class label is defined based on the data belonging to a high or low crime rate community. We considered attribute race as sensitive for our analysis.

The CI dataset comprises 14 attributes out of which 6 are numeric and the others are categorical. The class label is used to denote whether an individual's annual income is lower than $\$50K$ or greater than $\$50K$ per year. Four attributes are sensitive (*i.e.*, race, age, gender, native-country). Among them, we choose gender for our analysis.[1]

Baseline: We use the framework proposed by Luong et al. [9] as a baseline. This approach labels a record as discriminated if a significant difference of treatment (in the decision label) is observed among its k nearest neighbors from the protected groups vs. k nearest neighbors from non-protected group.

[1]Note that all sensitive attributes were considered in the pre-processing step. In the discrimination detection step, we only consider one sensitive attribute, but our hierarchical clustering approach can be easily generalized to deal with more than one attribute by marking a record as a member of protected group if it has at least one sensitive value.

Dataset	Dataset Characteristics				Parameters Settings					
	Size	$\|\mathcal{A}\|$	$\|\mathcal{H}\|$	*Sensitive*	α	β	θ	λ	M	τ
GC	1000	20	6	PersStat	1.00	0.2	0.5	40	8	0.005
CC	1994	128	24	race	1.00	0.4	0.5	40	8	0.005
CI	16 282	13	6	gender	1.005	0.2	0.5	200	8	0.005

TABLE I: Dataset characteristics and parameter settings

This method can be used to assess discrimination with respect to positive and negative decisions. Luong's approach is bottom-up, and searches for the k nearest neighbors of each record, and then analyzing the potential discrimination based on *diff* values as defined in Eq. 4. $diff_{(-,s)} \geq 0$ shows that the negative decision is more frequent among the sensitive neighbors compared to non-sensitive neighbors. This represents potential mistreatment of sensitive records based on their group membership, rather than explainable differences in explanatory attributes. As in the original experiments [9], we set the number of neighbors to 16 (*i.e.*, $k = 16$).

Evaluation Framework: We evaluate our hierarchical clustering for discrimination detection by measuring the positive ($bias^{(i)}_{pos}$) and negative ($bias^{(i)}_{neg}$) bias toward sensitive and non-sensitive records in a distant-based neighborhood. We compare our approach with the k-NN approach from [9]. However, a strait comparison between the approaches is not possible as the k-NN approach computes biases $diff_{(+,s)}$ and $diff_{(-,s)}$ with respect to a fixed number of neighbors whereas $bias^{(i)}_{pos}$ and $bias^{(i)}_{neg}$ are normalized versions of these metrics (*cf.* Section II-A). Therefore, for an accurate comparison, we calculate the bias obtained using k-NN for each record as follows:

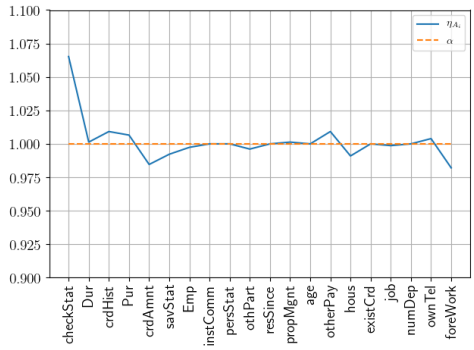
$$bias_{pos} = \frac{diff_{(+,\neg s)}}{p^+} \quad (12) \qquad bias_{neg} = \frac{diff_{(-,s)}}{p^-} \quad (13)$$

Eqs. 12-13 can be easily derived from the ones in Section II.

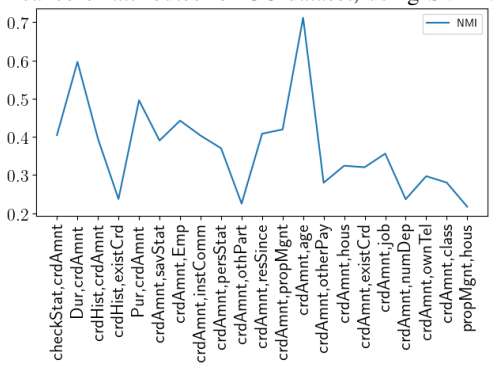
For our approach, we also compute the level of homogeneity (loh) of each cluster as a confidence bound on the reported discrimination measures.

B. Pre-processing

We apply the pre-processing step for attribute selection presented in Section III-A to the three datasets. The size of the obtained high impact feature sets ($\mathcal{H}$) is reported in Table I. As an illustrative example, Fig. 3 provides detailed results for the GC dataset. Fig. 3a shows the significance of the attributes. The highest η is observed for attribute checkStat, as it results in the largest drop in accuracy when this attribute was omitted. We also compute the correlation between attributes using Normalized Mutual Information (NMI), as shown in Fig. 3b. We chose this measure due to the several categorical attributes in the dataset. The obtained NMI values are used to determine whether high impact attributes have a high correlation with any sensitive attribute. The results reveal that PersStat is only correlated with crdAmnt, which however is not in $\mathcal{H}$. Accordingly, we select six attributes for discrimination discovery, *i.e.*, $\mathcal{H}$ ={checkStat, Dur, CrdHist, Pur, otherPay, ownTel}.



(a) Significance of attributes for GC dataset, using SVM classifier.



(b) Normalized Mutual Information (NMI) among attributes. For the sake of space, only combinations with a value greater than 0.1 are reported.

Fig. 3: Selection of high impact attributes for GC dataset.

C. Discrimination Discovery

In this section, we present the result of the hierarchical clustering (HC) vs. k-NN approach by looking at the discrimination measures $bias_{pos}$ and $bias_{neg}$ calculated over the three datasets and the high impact feature set $\mathcal{H}$, see Table II. These metrics are calculated for each homogeneous cluster (for HC method) and for k nearest neighbors of each record (for k-NN method) to assure similarity of records in terms of their high impact feature set ($\mathcal{H}$). However, note that the size of each cluster for k-NN approach is fixed and equal to $2k$ nearest neighbors of that record (k from sensitive and k non-sensitive) plus the record itself. Hierarchical clustering results in different sizes of homogeneous clusters, as reported in Table II. Also, in hierarchical clustering, the cluster size can be controlled by parameter λ such that larger clusters are allowed for larger or denser clusters for efficiency and faster convergence.

The results show that all three datasets suffer from some discrimination as all clusters have positive $bias_{pos}, bias_{neg}$ values for at least some records. For the GC dataset, the HC method indicates that 75% of the records have positive (negative) bias greater than or equal to -0.1 (0.14) while 25% of the records have positive (negative) bias greater than or equal to 0.07 (0.64). As reported, the negative bias is more severe, *i.e.*, a significant number of sensitive records are rejected because of their membership to the protected group. These results are consistent with known findings from the same dataset [9], [11].

Dataset	Approach	Affected records in clusters				loh			$bias_{pos}$			$bias_{neg}$		
		# clusters	*avg. size*	*min size*	*max size*	*avg.*	*min*	*max*	*avg.*	25^{th}	75^{th}	*avg.*	25^{th}	75^{th}
GC	HC	8	17	2	64	0.69	0.54	0.86	0.07	−0.1	0.07	0.31	0.14	0.64
	k-NN	699	33	33	33	n.a.	n.a.	n.a.	0.16	0.06	0.31	0.59	0.09	1
CC	HC	19	19	2	36	0.74	0.54	0.9	0.06	0.16	1.9	0.06	0.5	1.54
	k-NN	1529	33	33	33	n.a.	n.a.	n.a.	0.46	0.06	0.72	0.68	0.08	1.14
CI	HC	11	427	11	1444	0.61	0.52	0.84	0.88	0.68	1.02	0.21	0.15	0.2
	k-NN*	1336	33	33	33	n.a.	n.a.	n.a.	−0.11	−0.5	0.29	0.02	−0.09	0.15

* partial results (3000 records) due to resource consumption

TABLE II: Experiment results

Note that our detected bias values differ from the baseline, albeit showing a similar trend. We speculate that the results from the baseline approach suffer from the choice of a fixed number of neighbors, ignoring the actual strength of similarity.

A visual representation of the clusters obtained using the HC approach on the GC dataset is given in Fig. 4. Bubbles represent clusters that are possibly discriminated against. The position of each bubble is indicative of the amount of bias for that specific cluster ($bias_{pos}, bias_{neg}$), while its size (which is also labeled on the bubble) represents the number of affected records within the cluster. Positive values of $bias^{(i)}_{pos}$ and $bias^{(i)}_{neg}$ indicate discrimination against sensitive records while negative values show reverse discrimination where the sensitive records are being favored. For instance, the largest cluster (195 records) is associated with relatively small $bias_{pos}$ and $bias_{neg}$ against its 64 sensitive members (33% of the cluster total members). The small bias values are expected, as discrimination typically occurs among smaller portions of the population (affected communities are usually minorities). The high level of homogeneity of this cluster ($loh = 0.9$) provides high confidence in the results. We also note that negative bias values are observed for three relatively small clusters with centroids representing small loan amounts, which is for small home furniture mostly (per the Pur attribute that represents the purpose of the loan). On the other hand, the clusters with high values of $bias_{pos}$ have centroids representing large loan amounts, for business and education. These observations show one of the main advantages of our approach: we are able to profile users in each cluster, and gain more insights about the amount, confidence, and characteristics of discriminated records. In contrast, the k-NN approach only provides statistics on the amount of discrimination in a dataset as a whole.

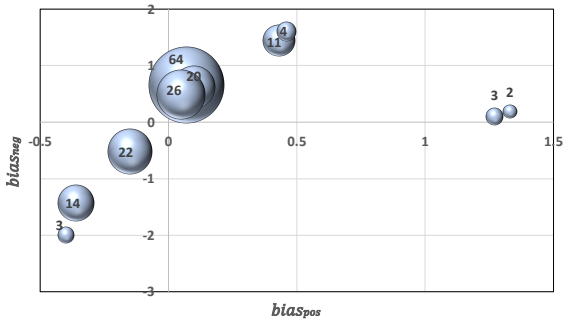


Fig. 4: Clusters from the GC dataset analysis. The size of each bubble represent the number of sensitive records in each cluster. Overlapping clusters are clusters that have the same discrimination rate (not the same records).

Computational complexity The recurrence equation[2] for the hierarchical clustering is

$$T(n) = mT(n/k) + O(Mknt)$$

where n is the number of records in the dataset, k is the number of clusters, and t is a constant representing the number of iterations of k-Means. Here, M denotes the number of times k-Means is executed for best clustering (Algorithm 2). Note that M is fixed and small (*e.g.*, 8 in our experiments) and thus $k \ll n$. Accordingly, each stage of the algorithm (including clustering) is linear in the input size n. We do not account for the complexity of the pre-processing step, as it is a one-time operation. The resulting time complexity from direct Master theorem is $T(n) = O(n \log n)$. For the baseline approach, we use a common brute-force technique for k-NN clustering[3] to find the k closest neighbors of each node from the protected and non-protected group, and calculate p_1^x and p_2^x (with $x \in \{+, -\}$). The resulting complexity is $O(n^2 \log n)$.

V. Related Work

A large body of research has explored the causes and risks of discrimination in decision processes (see, e.g., [15] for a survey). Among the others, data mining has been recognized as a possible source of discrimination [1]. Research on anti-discrimination in data mining has developed across two main trends: *discrimination prevention* and *discrimination discovery*. Discrimination prevention [4], [5], [7], [21], [22] aims to build discrimination-free classifiers, starting from training data biased by discriminant choices. Our work is positioned in the area of discrimination discovery.

The goal of discrimination discovery is to detect the presence of (possibly hidden) discriminatory decisions. A large body of research [9], [11], [12], [13], [16] models the problem of discrimination in terms of association rules mining and formalizes the notion of discrimination by means of a set of metrics able to quantify the level of discrimination in the rules discovered by association rules mining algorithms.

[2]Average case for k-Means implementation from sklearn package [6].

[3]We are not aware of a more efficient implementation of the k-NN approach.

These approaches, however, might result in misleading findings due to correlations between attributes. This problem is partly addressed by Qureshi et al. [14], who proposes the use of propensity score analysis to filter out the effects of so-called "confounding variables", i.e. variables (attributes) linked to both the class and other variables, which might introduce bias in the discovered association.

Our approach to discover discriminatory decisions significantly extends and refines early work in this area. Prior work has focused on rule and association extraction as a way to extract data patterns and reconstruct the underlying decision making process. Similarly, other works have looked into statistical approaches (*e.g.*, fitting a regression model to the data) to discover evidence of discrimination [2]. Yet, these works suffer from several limitations, including, an unmanageable set of rules, lack of a comprehensive view of the discriminatory actions and a relative high false-positive rate.

A work similar to ours is from Luong and colleagues [9], which relies on distance-based techniques. Luong's goal is to identify differences in predictions across groups of people due to non-protected characteristics. Here, a key notion is given by "similarity of treatment" and by how these groups are determined. These groups are pre-defined and fixed, leading to potential misinterpretation of the data and exclusion of any discriminated group that was not known a-priori. Our work shares a similar goal, *i.e.*, detecting similarity of treatment, but aim at a more generic analysis of potential evidence of discrimination in that we dynamically identify the groups that are homogeneous with respect to highly relevant attributes (with respect to the decision label) with no pre-established assumption on the number of groups to be identified (or their composition). Our approach also significantly differs in its method, in that clustering is dynamic and recursive, and possible discrimination – both positive and negative – is detected. Importantly, we are also able to efficiently pinpoint potentially discriminated groups even if these groups are severely underrepresented in the dataset. Finally, Luong's method may produce biased results due to the way the k neighbors are selected. In particular, it only considers the k nearest neighbors and not the amount of closeness.

VI. Discussion and Conclusions

In this paper, we presented an unsupervised approach for detection of discrimination in decisional processes. Our approach focuses on identifying profiles of alike individuals, wherein their treatment (*i.e.*, decision label) is "unexplainable", other than the use of sensitive attribute values. While not privy of limitations, we believe that our approach takes a significant step toward addressing data discrimination, even when there is little known ground truth on the type of discrimination that may have occurred and the protected groups of individuals. Particularly, it is worth noting that our approach is able to identify potentially discriminated groups even if these groups are underrepresented in the dataset.

We envision several directions for further improving our work. First, currently we rely on pre-defined sensitive attributes, and cannot easily detect hidden biases. That is, while it is possible to study each cluster's difference of treatments within groups according to any attribute or combination of attributes, our decision on whether a group is discriminated or not is currently based on the attributes identified a-priori. Further, in our empirical analysis we considered one condition (or combination of attributes) at a time. We will implement ways to efficiently analyze multiple sensitive attribute conditions, for increased fine-grain analysis. Of note, our approach for clustering is inherently distance-based. This may be a potential limitation in case of non-numerical attributes or attributes where a distance measure is semantically hard to determine. Finally, we currently ignore strength of membership of a cluster. We will study ways to account for soft memberships to better account for records that are not strongly linked to any cluster.

Acknowledgements: This work is funded by the ECSEL project SECREDAS (783119) and the ITEA3 project APPSTACLE (15017). Dr Squicciarini's work was partly funded by CSRE grant from the Pennsylvania State University.

References

[1] C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it, 2003. Tutorial at KDD.
[2] B. Edelman and M. Luca. Digital Discrimination: The Case of Airbnb.com. Harvard Business School Working Papers 14-054, Harvard Business School, 2014.
[3] European Union Legislation. http://ec.europa.eu/social.
[4] S. Hajian and J. Domingo. A methodology for direct and indirect discrimination prevention in data mining. *TKDE*, 25(7):1445–1459, 2013.
[5] A. A. Hintoglu, A. Inan, Y. Saygin, and M. Keskinoz. Suppressing data sets to prevent discovery of association rules. In *Proc. of ICDM*, pages 645–648. IEEE, 2005.
[6] k-means Time Complexity. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
[7] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, 2012.
[8] F. Kamiran, I. Žliobaitė, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644, 2013.
[9] B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of KDD*, pages 502–510. ACM, 2011.
[10] Normalized Mutual Information. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html.
[11] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of KDD*. ACM, 2008.
[12] D. Pedreschi, S. Ruggieri, and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In *Proc. of ICAIL*. ACM, 2009.
[13] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proc. of SDM*. SIAM, 2009.
[14] B. Qureshi, F. Kamiran, A. Karim, and S. Ruggieri. Causal discrimination discovery through propensity score analysis. arXiv preprint arXiv:1608.03735, 2016.
[15] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.
[16] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *TKDD*, 4(2), 2010.
[17] Scikit-learn Silhouette. http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.
[18] U.K. Legislation. https://www.legislation.gov.uk/.
[19] United Nations Legislation. https://www.ohchr.org/EN/pages/home.aspx.
[20] U.S. Legislation. https://www.usdoj.gov.
[21] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *TKDE*, 16(4):434–447, 2004.
[22] K. Wang, B. Fung, and P. Yu. Template-based privacy preservation in classification problems. In *Proc. of ICDM*, pages 466–473. IEEE, 2005.