THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Advanced Machine Learning Algorithms for Discrete Datasets

Sameen Mansha

Master of Computer System Engineering

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2019

School of Information Technology and Electrical Engineering

Abstract

Despite recent works in the area of machine learning, there remains the need for robust, yet easily usable, methods. In this thesis, we focus on the design, performance, and improvement of well-known clustering and classification algorithms for discrete datasets with application in different domains. In the first section of the thesis, we formulate an optimization problem for clustering interesting itemsets to extract a sparse representation of itemsets and show that their discrete nature makes it NP-hard. An efficient approximation algorithm is presented which greedily solves maximum set cover to reduce overall compression loss. Furthermore, we incorporate our sparse representation algorithm into a layered convolutional model to learn nonredundant dictionary items. Following the intuition of deep learning, our convolutional dictionary learning approach convolves learned dictionary items and discovers statistically dependent patterns using chi-square in a hierarchical fashion; each layer having a more abstract and compressed dictionary than the previous. In the second section for fairness aware classification, we utilize reject option in different classifiers, a general decision-theoretic framework for handling instances whose labels are uncertain, for modelling and controlling discriminatory decisions. Specifically, this framework permits a formal treatment of the intuition that instances close to the decision boundary are more likely to be discriminated in a dataset. We propose three different solutions for discrimination-aware classification problems. The first solution invokes probabilistic rejection in single or multiple probabilistic classifiers while the second solution relies upon ensemble rejection in classifier ensembles. The third solution integrates one of the first two solutions with situation testing which is a procedure commonly used in the court of law. We evaluate our proposed clustering and discrimination-aware classification solutions on relevant benchmark real-world datasets and compare their performance with previously proposed state of the art approaches. The results demonstrate the superiority of our solutions in terms of performance and flexibility of applicability.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

- **Sameen Mansha**, Hoang Thanh Lam, Hongzhi Yin, Faisal Kamiran, and Mohsen Ali, Layered Convolutional Dictionary Learning for Sparse Coding Itemsets, World Wide Web Journal 2018.

Contributor	Statement of contribution
Sameen Mansha (Candidate)	Designed Algorithms, Implementation and Paper Writing (60%)
Hoang Thanh Lam	Designed Algorithms, Implementation and Paper Writing (60%)
Hongzhi Yin	Discussions, Analysis of the Algorithm and Paper Writing (30%)
Faisal Kamiran	Discussions, Analysis of the Algorithm and Paper Writing (30%)
Mohsen Ali	Discussions, Analysis of the Algorithm and Paper Writing (30%)

- Faisal Kamiran, **Sameen Mansha**, Asim Karim and Xiangliang Zhang, Exploiting Reject Option in Classification for Social Discrimination Control, Information Sciences 2017.

Contributor	Statement of contribution
Faisal Kamiran	Designed Algorithms, Implementation and Paper Writing (60%)
Sameen Mansha (Candidate)	Designed Algorithms, Implementation and Paper Writing (60%)
Asim Karim	Discussions, Analysis of the Algorithm and Paper Writing (40%)
Xiangliang Zhang	Discussions, Analysis of the Algorithm and Paper Writing (40%)

Other publications during candidature

Zaheer Babar, Md Zahidul Islam and **Sameen Mansha**, Rank Forest: Systematic Attribute Sub-spacing in Decision Forest, The 15th Australasian Data Mining Conference, 2017, Australia.

Contributions by others to the thesis

My Principal advisor Dr. Hongzhi Yin has significantly contributed towards the research presented in this thesis. Besides, all co-authors of published papers included in this thesis have been involved in the design, development and revision of research contributions.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research involving human or animal subjects

No animal or human subjects were involved in this research.

Acknowledgments

I am thankful to my advisors: Prof. Shazia Sadiq and Dr. Hongzhi Yin for their precious time, fruitful discussions and valuable suggestions. I also want to thank the co-authors of my papers: Prof. Asim Karim, Dr. Faisal Kamiran, Dr. Hoang Thanh Lam, Dr. Mohsen Ali and Dr . Xiangliang Zhang for their hard work and knowledge sharing. I am thankful to my coworkers at Data and Knowledge Engineering (DKE) group for their support. Last but not least, I am grateful to my family to encourage me to travel overseas to pursue higher degree research program.

Financial support

“This research was supported by an Australian Government Research Training Program Scholarship”.

Keywords

Interesting Itemset Mining, Convolutional Dictionary Learning, Reject Option in Classifiers, Social Discrimination Control.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 170203, Machine Learning, 100%.

Fields of Research (FoR) Classification

FoR code: 0806, Information Systems, 100%.

Contents

List of figures

List of tables

List of Algorithms

Chapter 1

Introduction and Motivation

1.1 Introduction

With the growth in data generation, the effective arrangement and recovery of massive data has become progressively troublesome. The prime purpose behind the systematic arrangement, is to discover knowledge from generated data for future planning and directives. Massive research has been conducted and also in progress to devise more promising solutions for efficient data arrangement. Till now, significant advancements have been made in the field of data mining and machine learning that provide promising ways to process such large datasets [8]. Specifically, two learning tasks (classification and clustering) have become dominant ways to perform knowledge discovery and pattern mining. In clustering, the objective is to divide the data into groups (clusters) on the basis of calculated similarity or minimal distance among data instances. While classification refers to the assignment of a category (from previously defined categories) to newly generated instances, on the basis of identified patterns in previously categorized data.

In data mining operations, we use to represent a discrete dataset as a two dimensional array, as a combination of m data instances/records/rows $D = \{R_1, R_2, R_3, \cdots, R_m\}$. It can also be represented as a combination of n attributes/features/columns $R = \{A_1, A_2, A_3, \cdots, A_n\}$. Clustering segregate groups of similar instances R considering A attributes to place them in a cluster. In this way, dissimilar instances are separated. While in classification operations, a data instance contains values from n non-class attributes that provide a base to get a value for class attribute C (values from predefined classes/categories), where $C = \{c_1, c_2, c_3, \cdots\}$. Each data instance gets a category label (class label) from predefined categories. There are two types of datasets that have been defined to perform classification tasks (1) Existing dataset (termed as Training data) contains value of class attribute (class label) for each data instance (2) Unseen dataset (termed as Testing data) does not contain value for class attribute. In classification task, we build a classifier on training data (which contains predefined classes) and predict the class values for testing data. Clustering algorithms do not require class labels C while classification does. Selecting appropriate number of clusters and understanding unlabelled data is also a challenging task.

1.2 Motivation

In what follows, we provide a brief overview of motivation, contribution and related work in Interesting Itemset Mining (clustering) and Social Discrimination Control (classification) fields.

1.2.1 Interesting Itemset Mining

Itemset mining deals with clustering for a compact set of itemsets to summarize a given transaction dataset in the most effective and efficient way. For example, in market basket analysis, a dataset contains a number of items and transactions. Each transaction is a list of items a customer has purchased. We can examine which items are sold together to analyze user behavior, increase sales and make predictions. Early works in this domain focus on finding frequent items that satisfy minimum support thresholds for the analysis of datasets. Apriori is the first work introduced for finding frequent itemsets whose minimum support is above a user-specified threshold [9] and it has been applied extensively in numerous applications since then. Apriori and many similar algorithms, e.g., Eclat [10] and FPGrowth [11] suffer from the *pattern explosion*, i.e., high minsup thresholds lead to return a small number of well-known patterns. Additionally, these methods return an incredibly large number of patterns for small values of minimum support threshold, many of which are only variations of the same theme. For example, if we learn from the transactions that bread and butter are often purchased together and many people buy milk, then it is entailed by redundancy to inspect if these three items are purchased together [12]. A few other works tried to solve this problem [13–15] but they do not fully resolve the problem of pattern explosion [16].

This field is introduced as *'Interesting Itemset Mining'* by advanced itemset mining community which focus on finding the non-redundant and self-sufficient summary of data [6, 7, 12, 17–21]. These works have achieved comparatively interesting and nonredundant patterns than the frequent itemset mining works. We summarize a few recent works who mine small, high quality and non-redundant patterns that yield the best lossless compression of the database. Interesting itemset mining is already proved NP-hard [22]. A few clustering based approaches are used to create frequent feature value pairs belonging to a specific cluster. The compression ratios are dependent on the number of clusters. Outliers detection and compaction gain are bottlenecks in this work [23]. MTV [24] uses Minimum Description Length (MDL) principle together with the maximum entropy distribution to directly calculate the expected frequencies of itemsets and identify interesting contents. KRIMP [21] applies MDL principle to create a simple two column translation based code table that optimally describes the data. The candidate itemset is selected w.r.t. the standard candidate order. It uses a cover algorithm to select a smaller compressed sized encoding. KRIMP candidate generation technique requires high running time and selecting right threshold values for larger databases or candidate collections is challenging. SLIM [25] addressed this issue by directly mining descriptive patterns from the data. It uses MDL along with an accurate heuristic to greedily construct patterns in a bottom-up fashion. OPUS Miner [26] is a branch and bound approach which deploys two pruning mechanisms considering itemset values and statistical significance levels. It finds top k productive and nonredundant itemsets to

identify small sets of key associations ultimately leading to self-sufficient itemsets. Interesting Itemset Miner (IIM) [6] uses a generative model over itemsets in the form of Bayesian networks. The greedy approximation based weighted set cover approach infers interesting itemsets. These approaches help to intelligently analyze the data-driven problems from the domain of finance, graph search [27–29], recommendation systems [30–33] and data engineering [34–37] etc. however, existing works do not consider sparsity constraints of the encoding. In some applications, e.g., compression of transaction databases, sparsity constraint might be preferred to limit the maximum size of each selected itemset. Additionally, these works do not learn a convolutional hierarchical representation of data. In Chapter 2, we propose *Layered Convolutional Dictionary Learning for Sparse Coding of Itemsets (CDSI)* that draws inspiration from the field of sparse dictionary learning and convolutional sparse coding for clustering interesting itemsets. A concise description of related work in these fields is given as follows:

Sparse Dictionary Learning: Sparse coding is an unsupervised algorithm which is widely used in signal and image processing to compress images or signals using a compact set of basis learned from data. It discovers basis functions called *dictionary* to adapt it to specific data, an approach that has recently proven to be very effective for signal reconstruction and classification in the audio and image processing domain [38, 39]. A dictionary consisting of image edges can give a better representation of images than the pixel intensity values. *Sparsity constraint* is enforced to restrict the size of the basis for sparse coding image hence the dictionary is overcomplete. Sparse dictionary learning mainly deals with the continuous data while in practice many datasets are discrete. Continuing this highly promising line of work, we explore how to represent itemsets under sparsity constraint and learn dictionary. Though the idea of coding binary data is not new, handling of discrete data for the sparse coding problem is still challenging. It is in high demand to study sparse coding techniques for discrete data.

Layered Convolutional Sparse Dictionary Learning: A sparse feature vector is computed to reconstruct the original input vector by minimizing an energy function. A highly redundant representation of an image is produced if patches are processed independently, as these features can be correlated. The sparse coding algorithm cannot capture dependencies alone. To address this problem, a variety of convolutional sparse coding methods have been introduced in the image processing domain [40, 41]. These techniques are based on the convolutional decomposition of input data to learn dictionary under a sparsity constraint. It is a top-down approach seeking to generate the input signal by summing up the convolutions of the feature maps with learned filters. Sparsity limits the representation by imposing size restriction at each layer, which facilitates assembling parsimonious features into more complex structures. A convolutional sparse coded dictionary contains rich information which many existing feature detectors cannot detect.

1.2.2 Social Discrimination Control

Social discrimination is said to occur when a decision in favor of or against a person is made based on the group, class, or category to which that person belongs to rather than on merit. Discriminatory

practices suppress opportunities for members of deprived groups in employment, income, education, finance, and other benefits/services on the basis of their age, gender, skin color, religion, race, language, culture, marital status, economic condition, and other non-merit factors. Today, discrimination is considered unacceptable from social, ethical, and legal perspectives. Many anti-discrimination laws [42–45] have been enacted and many anti-discrimination organizations (e.g., ENAR [46]) are working for the eradication of discrimination. The consequences of discriminatory practices can range from legal prosecution to a variety of social problems like high unemployment rate, frustration, low productivity, and social unrest.

Data mining can help control discrimination arising from discriminatory or biased historical data. In particular, discrimination-aware classification problem studies the construction and application of classifiers learned from discriminatory or biased data. The do-nothing approach of simply using a classifier learned from discriminatory data will propagate, if not exacerbate, discriminatory decisions, which is undesirable for decision makers at financial institutions, hiring agencies, and social service providers. Thus, this do-nothing approach can lead to litigations and penalties.

In recent years, several methods have been proposed for discrimination-aware classification. However, these methods have one or both of the following shortcomings. First, they require that either the discriminatory data is processed to remove discriminatory patterns before learning a classifier or a specific classifier's learning algorithm is modified to make it discrimination-aware. Second, they are usually 'brute force' techniques with limited control over overall and illegitimate (unexplainable) discrimination removal.

These shortcomings of existing methods have hindered their adoption by practitioners. A direct consequence of the first shortcoming is that whenever discrimination w.r.t. a different sensitive attribute needs to be addressed, the historical data or classifier needs to be processed again. Experiments reported with the *Dutch Research and Documentation Center* (WODC) associated with the Ministry of Security and Justice and *Statistics Netherlands*, the national census body, confirm the importance of tackling discrimination w.r.t. multiple factors including age, gender, and race [47]. Being restricted to a specific discrimination-aware classifier (e.g., naive Bayes [2], decision tree [3]) is also an issue because that classifier may not be the best performing classifier for a given dataset. The second shortcoming can lead to reverse discrimination whereby deprived group individuals are favored without a legitimate or plausible explanation. This issue has been studied by the authors of [48]. They split overall discrimination into legal and illegal parts and claim that if the discrimination (e.g., high income of male employees as compared to female employees) can be explained by some reasonable factors (e.g., longer working hours of males), then it is acceptable and legitimate 'discrimination' rather than illegal discrimination. (i.e., higher salary of males can be explained by the higher work hours of males). On the other hand, it would be illegal to discriminate on the basis of sensitive factors (e.g., gender, race) without any plausible explanation. The current state-of-the-art methods either deal with the overall discrimination or illegal discrimination and are not flexible enough to prevent both overall and illegal discrimination simultaneously.

In Chapter 3, we develop and evaluate a methodology for making single and ensembles of classifiers

discrimination-aware w.r.t. overall and illegal discrimination. This methodology is based on the decision theoretic notion of reject option where instances with highly uncertain labels are not given one in classification (i.e., they are given the reject label). Previously, it has been hypothesized that discriminatory decisions are often made close to the decision boundary because of decision maker's bias [1]. Our proposed methodology formalizes this into practically usable solutions for discrimination-aware classification. Furthermore, the rejected instances represent potentially discriminated or favored instances in the biased dataset. Thus, our methodology also serves as a model-based discrimination discoverer in biased datasets.

Related Work

Discriminating against individuals based on their membership to specific segments of society is ethically and legally undesirable. Data mining techniques can assist with the discovery of discriminatory patterns from data and with preventing discriminatory decisions based on biased data. The topic of social discrimination in data mining was introduced by Pedreschi et al. in 2008 [49], and was further explored in [50–52]. They focused on discovering discriminatory classification rules from biased datasets following a frequent itemset mining approach coupled with a measure of discrimination. Since then many researchers have focused on discrimination detection and prevention in data mining [1, 4, 53–60]. A multidisciplinary survey of discrimination analysis methods is given by [61] while an edited book provides a summary of the research works for discrimination discovery and prevention [62]. The book also deals with the legal and ethical issues of discrimination and profiling.

Proposed methods for discrimination prevention are either based on data preprocessing or algorithm/model tweaking. Data preprocessing methods modify the biased data to remove discriminatory patterns from it before learning a prediction model from it. In works on discriminatory rule protection [53–55], data transformations are performed for making discriminatory classification rules discrimination-free according to a discrimination measure. The key limitation of these methods is their applicability to classification rules only which may not be the best classifier for a given problem. The authors of [57] propose a method of finding an intermediate representation of the given biased data that best encodes the data while obfuscating the membership of instances to the protected group. In [1], data sampling and [1, 63, 64] massaging techniques are presented for removing discrimination w.r.t. a single sensitive attribute. Although these methods can support the learning of any classifier, they are restricted to a single sensitive attribute at a time. In general, data preprocessing methods require that the data (preprocessed or original) is made available which may not be appropriate for privacy reasons or the released data need to be transformed to suppress the private informations.

Proposed methods for discrimination prevention requiring learning model adaptation include those for decision trees [3], naive Bayes classifiers [2], logistic regression [56], and support vector machines (SVM) [59]. All these methods require that the learning model or algorithm is tweaked, and these methods are specific to their respective classifiers. For example, in [3], the authors propose a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free while in [59] fairness constraints are introduced to control discrimination in discriminative classifiers like SVM.

Direct discrimination arises when sensitive attributes are utilized in learning and prediction. Nonetheless, it has been shown that discrimination is not removed by simply removing these attributes from the dataset [1]. That is, discriminatory decisions can still be made due to correlation of sensitive attributes with other attributes (indirect discrimination or *redlining*[1]. This issue has been studied in greater detail in [48]. The authors of [48] also present the concept of explainable and illegal discrimination and propose a variant of data preprocessing approaches of [1] to prevent the illegal discrimination only. However, their method is unable to handle multiple explanatory attributes and both explainable and illegal discrimination simultaneously. More recently, propensity score modeling has been introduced by [65] to filter out illegal discrimination from data. Subsequently, they develop analytical solutions for discrimination-aware linear regression that controls the illegal effect of an attribute on the outcome.

A technical approach that tackles both privacy in disclosing data mining models and discrimination in applying such models is discussed by Hajian et al. (2012). The work considers classification rule models and measures privacy by k-anonymity and discrimination by the number of PD rules. [66] propose a model of fairness of classifiers and relate it to differential privacy in databases. The model imposes that the predictions over two similar cases are also similar. The similarity of cases is formalized by a distance measure between tuples. The similarity of predictions is formalized by the distance between the distributions of probability assigned to class values

[67] presented two strategies for making standard classifiers and classifier ensembles discrimination-aware at run-time. Based on decision theory, these strategies provided stronger control and interpretability of the decisions. A similar approach of shifting the decision boundary has been shown by [60] to produce good accuracy-discrimination trade-off performance. In Chapter 3, we generalize our strategies to a model of discrimination based on reject option in classification. This model leads to a methodology for discrimination control in predictions. Following this methodology, we present three solutions for discrimination control, including a new solution incorporating situation testing, and evaluate them extensively for both illegal and overall discrimination prevention. These solutions require neither data preprocessing nor algorithm tweaking, and can be utilized with a variety of classifiers with ease.

[1] `http://en.wikipedia.org/wiki/Redlining`, October. 12, 2019

Chapter 2

Interesting Itemset Mining

2.1 Introduction and Motivation

In this chapter, we propose a convolutional sparse coding-based approach for interesting itemset mining that is essentially different from the tasks in image processing domain with real values. We propose a matching pursuit greedy approach which performs dictionary learning from transaction data to reduce data loss compression under sparsity constraint. To further enhance its performance, we embed our sparse coding algorithm into a convolutional neural network based architecture such that each layer learns a complex discrete representation from the transformed database. This resembles state-of-the-art convolutional sparse coding in the image processing domain [40, 68]. Adding sparse representation of images and signals into training instances helps to improve the classification accuracy [69]. Nevertheless, leveraging the sparse representation of itemsets to enhance the performance of classifiers (e.g., Naive Bayes, Decision Trees, Random Forest etc) is still an open question. To summarize, we make the following contributions:

- Sparse coding of itemsets is first time addressed and formulated as an optimization problem. We prove it NP-hard by reducing it to set cover problem. We propose approximation based sparse coding algorithm, *Dictionary Learning for Sparse Coding of Itemsets (DSI)* to efficiently learn nonredundant dictionary elements for lossless compression. It provides a bottom-up mapping from transaction to dictionary items, efficiently giving a reconstruction close to the original transactions.

- We propose a new approach *Layered Convolutional Dictionary Learning for Sparse Coding of Itemsets (CDSI)* to deploy sparse coding within a convolutional resembling model to learn grouping representation at each level. The dictionary itemsets are interfused in the database to learn a meaningful representation.

- An extensive empirical validation on thirteen datasets shows the superiority of our proposed methods as compared to the recent works. A text dataset (JMLR) is used to evaluate the pattern meaningfulness just by eyeballing. Transactions of nine UCI [70] and three SIPO [18] datasets

(Sec. 2.5.3) are sparse coded, to determine its impact on the prediction accuracy of different classifiers.

Our targeted problem is formally defined and proved NP-hard in Section 2.2. Greedy approach for sparse representing itemsets is presented in Section 2.3. In Section 2.4, we explain a layered convolutional process for transforming database and dictionary learning. Section 2.5 describes our extensive empirical validation in detail. We conclude our work with future directions in Section 2.6.

2.2 Problem Definition and Proof of NP-Hardness

For ease of presentation, we first introduce some preliminary concepts and notations. Let $D = \{T_1, T_2, \cdots, T_n\}$ be a database of n transactions where each transaction belongs to a set of items I=$\{i_p|1 = 1, ..., p\}$. The cardinality of a transaction is the number of items in it. When a set of items called itemset, contains p items, it is referred as p-itemset. We aim to learn a dictionary $B = \{I_1, I_2, \cdots, I_m\}$ of m basis (itemsets), from which discrete sparse code of the database can be inferred. A sparse code of transaction T is the union of k itemsets $U(b)$: $\cup_{i=1}^{k} B_i$ from B such that $U(b) \subseteq T$ and k is less than the cardinality of T. With these notations, we formulate the following research problems:

Problem 1. *[Finding sparse representation of transaction T] Given a dictionary B and sparsity constraint (k: the maximum number of basis to choose from B), a sparse code of T is denoted as $B(T)$:*

$$B(T) = \arg min_{b \subset B, |b| \leq k} |T - U(b)| \tag{2.1}$$

where $U(b)$ represents a set of items in T that are covered by b.

Example 1. *Given $T = qrvwx$, $B = \{qr, vw, vy, yz\}$ when k is set to 1. The basis are $B(T) = \{qr\}$, and when $k = 2$, $U(B) = \{qrvw\}$.*

Sparse coding over the whole database D with the basis B incurs a loss function defined as $L_B(D) = \Sigma_{j=1}^{n} |T_j - U(B(T_j))|$. In Example 1, the loss for $B(T) = \{qr\}$ to encode $T = qrvwx$ is 3 while the loss for $B(T) = \{qr, vw\}$ is 1. Since vy is not a subset of $qrvwx$, it cannot be added into B. To better preserve the original information contained in a transaction database, a beneficial dictionary with less encoding loss is expected to be learned.

Problem 2. *[Dictionary learning from candidates] Given a database of transactions $D = \{T_1, T_2, \cdots, T_n\}$, the maximum number of basis allowed in a sparse code (sparsity constraint) k, and a set of candidate itemsets C, find a dictionary $B^* \subset C$ with maximum m basis, such that $B^* = \arg min_{B \subset C} L_B(D)$.*

To solve Problem 2, people may solve the following problem first:

Problem 3. *[Candidate set Construction from database] The encoding loss function in Problem 2 requires a candidate set C for inclusion in the dictionary. How to construct a high-quality candidate*

Algorithm 1 Dictionary Learning for Sparse Coding of Itemsets (DSI)

1: **Input**: A database $D = \{T_1, T_2, \cdots, T_n\}$, a candidate itemset C, parameters m, k
2: **Output**: Learned Dictionary B
3: $B = \emptyset$
4: **for** $itr = 1$ to m **do**
5: $\quad minloss = \infty$
6: $\quad$ **for** I in C **do**
7: $\quad\quad B^+ = B \cup \{I\}$
8: $\quad\quad O_{B^+} = |B \cap \{I\}$—
9: $\quad\quad L_{B^+} = \sum_{j=1}^{n} \text{MaxSetCover}(T_j, B^+, k)$
10: $\quad\quad$ **if** $minloss > L_{B^+}$ **then**
11: $\quad\quad\quad B^* = B^+$
12: $\quad\quad\quad minloss = L_{B^+}$
13: $\quad\quad$ **else if** $(minloss = L_{B^+}) AND (O_{B^+} < O_{B^*})$ **then**
14: $\quad\quad\quad B^* = B^+$
15: $\quad$ Subtract B^* from C: $C = C \setminus \{B^*\}$
16: $\quad B = B^*$
17: return B

set C from the database D is another challenging and important problem, as the C contents determine the quality of the learned dictionary and the encoding loss of the database to some extent.

Theorem 1. *Problem 1 is NP-hard.*

Proof We prove the NP-hard nature of problem by reduction to the *set-cover* problem. Let $S = \{1, 2, \cdots, n\}$ and $H = \{s_1, s_2, \cdots, s_m\}$, where $s_i \subset S$. **Set cover** problem asks whether we can construct a set $x \subset H$ such that $|x| = k$ and $\cup_{i=1}^{k} s_i = S$. Let $T = S$ and $B = H$, then solving Eq. 2.1 will result in sparse representation of T, that is $b^* \subset B$ such that $|T - U(b)|$ is minimized. Let b^* be the solution to Problem 1. If $|T - U(b)| = 0$ then it is easy to see that b^* is the set cover of S otherwise if the size is more than zero then no set-cover of size k exists. Hence solving Eq. 2.1 will solve the set-cover problem. Problem 1 has been reduced to the set cover problem and this reduction is polynomial in the problem input size. Hence, the theorem is proved.

2.3 Dictionary Learning for Sparse Coding Itemsets (DSI)

In this section, we present our proposed algorithmic framework (DSI) to learn sparse code dictionary in detail, and the pseudocode is given in Algorithm 1. It iteratively selects m basis from a set of candidate itemsets C. In each iteration, a single itemset I from C is chosen to form a transitory dictionary B^+ with already selected itemsets, and then the encoding loss for the database based on B^+ is computed(lines 7-9). In addition, it also calculates the number of overlapping items between the newly selected itemset I and learned dictionary B. The new itemset I is added to the dictionary if the loss and overlaps with selected basis are less than other candidates so far (lines 10-14). We present Example 2 for the better understanding of DSI:

Example 2. *Assume that we have a database* $D = \{T_1 = qrvwx, T_2 = qrvyz, T_3 = qrvwyz\}$, $C = \{qr, vw, vy, yz\}$, $m = 3$ *and* $k = 2$. *We explain Table 2.1 to show how Algorithm 1 works:*

- ***Step 1:*** *Initially B is empty. In each iteration (lines 6 -14), we look for an itemset I such that when we add it to B, the database D can be encoded with minimum loss and overlaps. Step 1 shows loss of encoding each transaction in D using candidates I from C. As observed, the overall loss is minimum when I = qr with the loss equal to 10. Therefore qr is added to* $B = \{qr\}$ *and deleted from C.*

- ***Step 2:*** *The next itemset that works together with B to minimize the overall loss is* $I = vw$ *with the overall loss equal to 6. We update B to* $\{qr, vw\}$ *and remove vw from C accordingly.*

- ***Step 3:*** *We calculate the encoding loss for each remaining candidate in C considering the learned dictionary B. We can see that* $\{vy\}$ *and* $\{yz\}$ *lead to the same loss value of 4. Nonetheless, item v in vy intersects with the dictionary element vw, making the overlap* O_{vy} *to be 1. On the other hand, yz has no overlap with itemsets in dictionary B, i.e.,* O_{yz}*=0. Ultimately, we update B to* $\{qr, vw, yz\}$ *and stop the algorithm after selecting* $m = 3$ *basis.*

Table 2.1: The illustration of running Algorithm 1 in Example 2. Selected items are emphasized in bold.

Step 1: *qr* is added into *B*.

	Transactions				
I	qrvwx	qrvyz	qrvwyz	L_{BI^+}	O_{BI^+}
q r	**3**	**3**	**4**	**10**	**0**
v w	3	5	4	12	0
v y	5	3	4	12	0
y z	5	3	4	12	0

Step 2: *vw* is added into *B*.

	Transactions				
I	qrvwx	qrvyz	qrvwyz	L_{BI^+}	O_{BI^+}
v w	**1**	**3**	**2**	**6**	**0**
v y	3	1	2	6	0
y z	3	1	2	6	0

Step 3: *yz* is added into *B*.

	Transactions				
I	qrvwx	qrvyz	qrvwyz	L_{BI^+}	O_{BI^+}
v y	1	1	2	4	1
y z	**1**	**1**	**2**	**4**	**0**

DSI uses a greedy method (MaxSetCover) to calculate the encoding loss, pseudocode is given in Algorithm 2. Our loss calculation method greedily encodes every transaction $T_i \in D$ with the basis of B^+. Algorithm 2 follows the standard procedure to solve the max set cover problem which guarantees

Algorithm 2 MaxSetCover(T, C, k)

1: **Input**: A transaction T, a set of potential basis itemsets C, parameter k
2: **Output**: The encoding loss
3: $G = \emptyset$
4: **for** $i = 1$ to k **do**
5: $I^* = \arg MAX_{I \in C, I \subset T} |U(G \cup \{I\})|$
6: Remove I^* from C: $C = C \setminus \{I^*\}$
7: Add I^* to G: $G = G \cup \{I^*\}$
8: Return $|T - U(G)|$

an approximation factor of $1 - \frac{1}{e}$ to the optimal solution [22]. The algorithm inputs a transaction T, a set of potential candidates C and a sparsity parameter k. It performs a greedy strategy to solve the max set cover problem by selecting up to k basis that curtail the encoding loss simultaneously. It returns the encoding loss, i.e., the number of items in T that have not been covered by the selected basis from B^+. Example 3 explains the working of matching pursuit greedy approach given in Algorithm 2.

Example 3. *Assume that $T = qrvwyz$, $C = \{qr, vw, vy, yz\}$ and $k = 2$, Algorithm 2 performs following steps to calculate encoding loss used in Table 2.1:*

- ***Step 1**: Initially G is empty.*
- ***Step 2**: The itemset $I \in C$ that maximizes the coverage of T is qr, so qr is added to G and deleted from C. ($G = \{qr\}$, $C = \{vw, vy, yz\}$).*
- ***Step 3**: The next itemset $I \in C$ that together with selected itemsets in G maximizes the overall coverage of T is vw so $R = \{qr, vw\}$. The algorithm stops when sparsity limit approaches, i.e., $k = 2$. The encoding loss is two ($|T - U(G)| = 6 - 4 = 2$), as two items in T are not covered by G.*

2.4 Layered Convolutional Dictionary Learning for Sparse Coding of Itemsets (CDSI)

In this section, we introduce a novel convolutional sparse coding mechanism (CDSI) to learn statistically dependent sparse dictionary in a hierarchical fashion. Dictionary items are convolved in each layer to transform the database; allowing next layer to learn more complicated patterns. This is similar to the idea of the deep learning technique: Convolutional Neural Networks (CNNs) [68], where learned filters are convolved with the input image and next layer of convolutional filters work on the output of the previous layer, allowing CNN to capture features at different levels of abstractness [71]. The convolution process has an advantage that the itemsets are learned in a hierarchical way, and various dictionaries with different-granularity abstractions can be achieved for different applications. We provide an overview of our layered convolutional dictionary learning algorithm below, and outline how it works:

Algorithm 3 Transform Database(T, B)

Input: Database $D = \{T_1, T_2, \cdots, T_n\}$, learned dictionary B.
Output: Transformed database $D' = \{T_1, T_2, \cdots, T_n\}$
for I in B **do**
    Generate a new symbol for I
    for $j = 1$ to n **do**
        if I in T_j **then**
            Replace I in T_j with the corresponding new symbol
Return transformed database D'

1. Construct a candidate set C using chi-square (see Section 2.4.1 for a discussion of how to construct a meaningful candidate set).

2. Run Algorithm 1 to learn a dictionary from C that sparse-codes the database D well.

3. Run Algorithm 3 to transform the database D using the learned dictionary in the second step (see Section 2.4.2).

4. To learn patterns in the next layer, return to step 1.

2.4.1 Candidate Set Construction

Quality of sparse dictionary learning (Algorithm 1) is highly dependent upon the contents of candidate set C. To build up C, a possible solution is to use frequent pattern mining algorithm such as the Apriori algorithm [9] which is subject to explosion (see Chapter 2 of [16]). In this section, we propose a refined approach to find statistically dependent itemsets. Intuitively, a pattern is only admissible if there is a strong dependency and correlation. Therefore, in order to compose the candidate set C, we use chi-square test [72]. Let q and r be two items and we define:

- $F_{qr} = |\{T_i \in D | qr \in T_i\}|$, i.e., the frequency of the itemset qr.
- $F_{q\bar{r}} = |\{T_i \in D | q \in T_i, r \notin T_i\}|$, i.e., the number of transactions that contain q but not r.
- $F_{\bar{q}r} = |\{T_i \in D | q \notin T_i, r \in T_i\}|$, i.e., the number of transactions that contain r but not q.
- $F_{\bar{q}\bar{r}} = |\{T_i \in D | q \notin T_i, r \notin T_i\}|$, i.e., the number of transactions that neither contain q nor r.
- $E_{qr} = \frac{F_{qr}^2}{N}$, i.e., the expected frequency of qr given the assumption that q is independent from r.
- $E_{q\bar{r}} = \frac{F_{q\bar{r}}^2}{N}$, i.e., the expected number of transactions that contain q but not r.
- $E_{\bar{q}r} = \frac{F_{\bar{q}r}^2}{N}$, i.e., the expected number of transactions that contain r but not q.
- $E_{\bar{q}\bar{r}} = \frac{F_{\bar{q}\bar{r}}^2}{N}$, i.e., the expected number of transactions that neither contain q nor r.

The chi-square statistics is defined as follows:

$$chi-square = \frac{(F_{qr}-E_{qr})^2}{E_{qr}} + \frac{(F_{\bar{q}r}-E_{\bar{q}r})^2}{E_{\bar{q}r}} + \frac{(F_{q\bar{r}}-E_{q\bar{r}})^2}{E_{q\bar{r}}} + \frac{(F_{\bar{q}\bar{r}}-E_{\bar{q}\bar{r}})^2}{E_{\bar{q}\bar{r}}} \tag{2.2}$$

If q and r are statistically independent then it follows a chi-square distribution with one degree of freedom. Based on this observation, we can test for our null hypothesis: *q and r are statistically independent*. The test can be performed for any pair of items in the database and only pair that passes the test (when null hypothesis is rejected at a significant level of 0.05) will be scrutinized as potential itemsets in the candidate set C. Adding statistically dependent item pairs into the candidate set, ultimately leads to the dictionary learning by running Algorithm 1.

Table 2.2: CDSI: Dictionary learning from convolved and transformed database at second layer. Items placed in $B = \{\ \alpha\beta, \gamma x, \alpha v\}$ are highlighted in bold.

Step 1:$\alpha\beta$ is added into B.

	Transactions				
I	$\alpha\beta x$	$\alpha v\gamma$	$\alpha\beta\gamma$	L_{B^+}	O_{B^+}
$\alpha\beta$	**1**	**2**	**1**	**4**	**0**
αv	2	1	2	5	0
βx	1	3	2	6	0
γx	2	2	2	6	0

Step 2: γx is added into B.

	Transactions				
I	$\alpha\beta x$	$\alpha v\gamma$	$\alpha\beta\gamma$	L_{B^+}	O_{B^+}
αv	1	1	1	3	1
βx	0	2	1	3	1
γx	**0**	**1**	**0**	**1**	**0**

Step 3: αv is added into B.

	Transactions				
I	$\alpha\beta x$	$\alpha v\gamma$	$\alpha\beta\gamma$	L_{B^+}	O_{B^+}
αv	**0**	**0**	**0**	**0**	**1**
βx	0	1	0	1	2

2.4.2 Database Transformation and Convolution

We elucidate database transformation process with the toy database described in Example 2. Given a dictionary $B = \{qr, vw, yz\}$, Algorithm 3 transforms the database D into an advanced database with refined items where each item corresponds to an itemset in the dictionary B. Let us re-write the basis itemsets in B as $B = \{\alpha = qr, \beta = vw, \gamma = yz\}$, where each basis itemset in B is now represented by a new item (symbol) that is not present in the current alphabet. Algorithm 3 transforms the database $D = \{T_1 = qrvwx, T_2 = qrvyz, T_3 = qrvwyz\}$ into $D' = \{T_1 = \alpha\beta x, T_2 = \alpha v\gamma, T_3 = \alpha\beta\gamma\}$. The new database D' contains transactions with dependent itemsets $\{\alpha, \beta, \gamma, v, x\}$, while the original itemsets

Table 2.3: Summary of datasets

Dataset	Transactions	Items	Labels
asl-gt-thad	3464	47	40
breast	699	18	2
congress	435	34	2
context	240	56	5
ecoli	336	26	8
glass	214	41	7
hepatitis	155	54	2
iris	150	16	3
jmlr	788	4976	NA
mushroom	8124	88	2
skating	530	41	6
soybean	683	99	19
zoo	101	35	7

were $\{q,r,v,w,x,y,z\}$. Table 2.2 shows the process of dictionary learning for the transformed database D' at second layer. Note that the candidate set C in this example is constructed by randomly selecting item pairs from the transaction database, as there are only three transactions, making it impossible for chi-square to find any dependent patterns.

2.5 Experiments

In interesting itemset mining, a powerful representation of data has higher values of (i) pattern interpretability, and (ii) classification accuracy. Our extensive empirical validation also considers these criteria to evaluate the effectiveness of proposed algorithmic framework. We compare our proposed sparse coding techniques with the IIM [6] and MTV [7], because they represent the state-of-the-art techniques for itemset mining and significantly outperform existing approaches developed in [18,20,21] on similar standard datasets as adopted in our experiment.

2.5.1 Dataset Description

We use discretized version of Semi Interval Partial Order (SIPO) datasets (introduced in [18]) and UCI datasets [70] for classification. Table 2.3 summarizes the characteristics of datasets used. It is always a challenging task to measure the meaningfulness of discovered patterns as a potential solution, thus text datasets are used to informally evaluate the quality by comparing pattern interpretability and relevance. We use the JMLR abstract text dataset from Journal of Machine Learning website [1] which is easy to interpret.

[1] http://jmlr.csail.mit.edu/

Table 2.4: Top 10 non-singleton patterns selected from the JMLR abstracts dataset to compare pattern interpretability for CDSI (Sec. 2.4), IIM [6] and MTV [7].

CDSI	IIM	MTV
select featur	associ rule	experiment result
machin learn	support vector machin svm	synthetic real
exact approxim	parameter parameters	real datasets
graphic variabl	anomali detect	pattern discov
data set	synthetic real life	associ rule mine
problem solv	sequenc sequential	frequent pattern mine algorithm
error bound	background knowledg	train classifi
probabl distribut	semi supervised	address problem
lower bound	local global	classifi class
independ compon analysi	linear discriminant analysi	machin learn

2.5.2 Interpretability of Sparse Representation

Table 2.4 shows MTV returns interrelated and less diverse frequent patterns, e.g., "synthetic real", "real datasets", "train classifi", "classifi class", etc. IIM derives relevant patterns (e.g., "anomali detect" and "semi supervised", etc.), however, a few patterns (e.g., "parameter", "parameters" and "sequenc", "sequential") require stemming to remove redundant patterns. Patterns extracted by CDSI at 4th layer of convolution dictionary with parameters $(m = 10, k = 5)$ are also given. We can observe that CDSI generates more revealing, diverse and comprehensive patterns, e.g., "machine learning", "graphic variable", "probabl distribut", etc. Besides, they do not require stemming. To conclude, CDSI comparatively generates interpretable, heterogeneous and less redundant patterns.

Table 2.5: Data preparation for classification using CDSI (Sec. 2.4), IIM [6] and MTV [7] mined patterns as binary features.

TID	Patterns	Transactions (Singletons)	Extended Transactions	Label
T_1	q,r,v	1,1,1,0,0	1,1,1,0,0,1,0	A
T_2	r,v	0,1,1,0,0	0,1,1,0,0,1,0	A
T_3	q,v,x	1,0,1,0,1	1,0,1,0,1,0,0	L
T_4	r,w,x	0,1,0,1,1	0,1,0,1,1,0,1	L
T_5	q,r,v	1,1,1,0,0	1,1,1,0,0,1,0	L

2.5.3 Classification Accuracy

Classification accuracy inflates conceding that sparse representation techniques or interesting itemset mining algorithms are employed on data [17, 73]. Table 2.5 presents a fictitious scenario to explain our experimental setup with a database D containing 5 transactions: $D = \{T_1, T_2, \cdots, T_5\}$ and two

class labels $\{A, L\}$. These transactions are illustrating the purchase of items $\{q, r, v, w, x\}$ with the proportionate input vector presentations, e.g., (T_1) ={1,1,1,0,0}. Since 0 and 1 exhibit if any specific item has been purchased, third element of T_1 is set 1 to depict purchase of v. These labeled transactions are feed to various classifiers. To evaluate if mined patterns are boosting the classification accuracy, we wrap them as binary values within transactions. To do so, we increase the length of transactions to append discovered patterns. Let us say if CDSI discovers two patterns (r, v) and (r, x) then 6^{th} and 7^{th} elements are added in each transaction to demonstrate the presence of distinct pattern (given in the extended transaction column of Table 2.5). Now the vector representations of T_1 will become $\{1,1,1,0,0,1,0\}$ while preserving record of purchase of remaining items q, r, v.

Table 2.6 presents the accuracy of different classifiers (e.g., Naive Bayes, J48, Random Forest, and IBk) for SIPO and UCI datasets described in Table 2.3. To be unbiased, the number of mined patterns is set to minimum patterns returned by any of the algorithms. These patterns are incorporated in the transactions (singletons) following the way extended input vectors are created in Table 2.5. We run our experiments using WEKA [74] over 5 fold cross-validation with parameters set to default values. Patterns are extracted using CDSI (with parameters $layers = 10, k = 10$), IIM [6] and MTV [7] (default parameter values adjusted in online available codes are used for existing approaches). Each cell of Table 2.6 shows the accuracies of different methods for respective classifiers. The highest prediction accuracy for any input vector type is emphasized in bold. The last column (Best) shows the highest accuracy for all types of input data and highlights the topmost value in bold. The prediction accuracy of all datasets increases when extended transactions are fed in comparison to when the classifier is only trained on the transactions themselves (singletons). Generally, CDSI significantly improves the prediction accuracy certifying our assumption about convolutional sparse coded dictionary carrying influential objective information.

2.6 Conclusion

Convolutional sparse model dictionary learning has been used before in the image processing domain [40, 41], it is still not studied for the itemset mining so far. In this chapter, we present approximation based algorithms to find the sparse representation of itemsets, which is discrete in nature. We propose an optimization technique to learn dictionary under the sparsity constraint from the transaction dataset. Based on this mechanism, a convolutional dictionary learning method is presented that allows extracting dictionaries at different levels of abstractness. Chi-square test is performed to extract statistically dependent patterns from the transaction data and input it to the layered dictionary learning algorithm; generating increasingly complex and statistically dependent patterns in each layer. We conduct extensive experiments on various datasets showing that sparse representation forms a succinct input representation and when combined with different classifiers, their efficacy is increased.

Table 2.6: CDSI (Sec. 2.4) improves the prediction accuracy of IIM [6] and MTV [7]. For a fair comparison, identical number of patterns returned from each method are used.

		Naive Bayes	J48	Random Forest	IBk	Best
asl-gt-thad	Singletons	5.79	8.73	**12.94**	9.00	12.94
	CDSI	4.79	9.35	**13.26**	9.05	**13.26**
	IIM	4.82	8.52	**12.97**	9.02	12.97
	MTV	5.14	9.00	**13.08**	8.82	13.08
breast	Singletons	94.13	93.56	**94.42**	93.99	**94.42**
	CDSI	93.07	94.13	**94.42**	94.13	**94.42**
	IIM	94.13	93.56	**94.42**	93.99	**94.42**
	MTV	92.56	94.13	**94.42**	94.13	**94.42**
congres	Singletons	91.49	94.71	**95.86**	92.64	95.86
	CDSI	91.49	94.94	**96.09**	92.18	**96.09**
	IIM	91.95	94.71	**95.86**	93.56	95.86
	MTV	91.49	94.71	**95.40**	93.79	95.40
context	Singletons	**77.89**	70.00	74.21	72.10	77.89
	CDSI	**78.42**	71.05	73.15	72.63	**78.42**
	IIM	**77.89**	70.00	73.15	69.47	77.89
	MTV	**76.84**	68.42	72.10	69.47	76.84
ecoli	Singletons	80.95	82.14	81.54	**83.63**	83.63
	CDSI	81.85	81.25	82.44	**83.92**	83.92
	IIM	79.76	81.54	81.54	**83.63**	83.63
	MTV	81.25	82.14	82.14	**84.22**	**84.22**
glass	Singletons	72.42	69.15	**72.89**	68.69	72.89
	CDSI	**72.90**	70.56	72.89	72.42	72.90
	IIM	71.96	69.15	**74.29**	69.15	**74.29**
	MTV	71.49	69.15	**73.36**	68.69	73.36
hepatitis	Singletons	**83.71**	78.06	80.64	81.93	83.71
	CDSI	83.87	76.77	82.58	**85.16**	**85.16**
	IIM	**83.87**	78.06	78.70	81.93	83.87
	MTV	**83.22**	80.00	81.93	81.93	83.22
iris	Singletons	**94.00**	94.00	94.00	94.00	94.00
	CDSI	**94.66**	94.00	94.66	94.00	**94.66**
	IIM	**94.00**	94.00	94.00	94.00	94.00
	MTV	**94.00**	94.00	94.00	94.00	94.00
mushroom	Singletons	97.84	**100.00**	100.00	100.00	**100.00**
	CDSI	98.65	**100.00**	100.00	100.00	**100.00**
	IIM	97.80	**100.00**	100.00	100.00	**100.00**
	MTV	97.83	**100.00**	100.00	100.00	**100.00**
skating	Singletons	**67.45**	58.82	65.09	52.74	**67.45**
	CDSI	**64.31**	61.76	62.74	53.52	64.31
	IIM	**67.25**	58.82	63.72	52.74	67.25
	MTV	**63.52**	58.62	63.33	51.17	63.52
soybean	Singletons	92.97	**93.70**	92.97	91.80	93.70
	CDSI	92.82	93.55	**93.55**	91.80	**93.55**
	IIM	92.53	**93.41**	93.11	91.65	93.41
	MTV	93.41	93.11	**93.55**	91.80	93.55
zoo	Singletons	**96.03**	93.06	96.03	96.03	96.03
	CDSI	94.05	93.06	97.02	**97.02**	97.02
	IIM	96.03	93.06	96.03	**97.02**	97.02
	MTV	96.03	93.06	**98.01**	96.03	**98.01**

Chapter 3

Reject Option in Classification for Social Discrimination Control

In this chapter, we present three rejection strategies and corresponding rules for discrimination control in predictions. The first solution called Probabilistic Rejection (PR), rejects instances with uncertain posterior probabilities, thus enabling it to be used with any probabilistic classifier or ensemble of classifiers. Our second rejection strategy, called Ensemble Rejection (ER), identifies instances that are not unanimously labeled by an ensemble of classifiers, thus emulating the natural decision making process by a group of experts. Our third rejection strategy, called Situational Rejection (SR), combines probabilistic rejection or ensemble rejection with situation testing to identify discriminated instances. Situation testing is a legally admissible procedure for verifying discrimination cases by comparing them with other similar cases. All strategies/solutions include relabeling rules with parametric control over the resulting discrimination. We perform extensive experiments to verify the superior performance of our methodology. In particular, we also demonstrate that our methodology prefers removing illegal discrimination over explainable discrimination while reducing overall discrimination. Thus, it addresses a common criticism that discrimination prevention methods disregard explainable discrimination while removing overall discrimination.

We use this third approach to show that our proposed solutions are the most appropriate ones for discrimination prevention. The rest of the chapter is organized as follows. Section 3.1 defines the problem setting and measures for overall and illegal discrimination. We present our reject option based methodology and specific solutions in Section 3.2. Section 3.3 presents experimental evaluations and discussions of our solutions. We summarize and conclude our contribution in Section 3.4.

3.1 Background and Notation

This section defines the problem setting and introduces the measures used in this work.

3.1.1 Problem Definition

We consider a two-class classification problem with label $C \in \{C^+, C^-\}$ defined over instances $X \in \mathscr{X}$ described by a fixed number of attributes. A discriminatory dataset $\mathscr{D} = \{X_i, C_i\}_{i=1}^N$ is available in which the labels C_i are biased w.r.t. one or more sensitive or discriminatory attributes S, e.g., Gender or Race. We assume that C^+ is the desirable label. The instances in $\mathscr{X}$ can be distinguished between those belonging to a deprived group $\mathscr{X}^d$ or a favored group $\mathscr{X}^f$, where $\mathscr{X}^d \cap \mathscr{X}^f = \emptyset$ and $\mathscr{X}^f = \mathscr{X} \setminus \mathscr{X}^d$. This dichotomous grouping of the instances is based on the values of the sensitive attributes. Besides the sensitive attributes there are some attributes that represent the plausible reasons for preferential treatment on the basis of sensitive attributes. We refer to these attributes as explanatory attributes and denote them by E.

To illustrate the notations, consider a university where women have been denied admission in comparison to men. Here gender is a sensitive attribute (S), males belong to the favored group ($\mathscr{X}^f$), females are the deprived group ($\mathscr{X}^d$), and the acceptance or rejection decision of the selection committee defines the class label (C). Every applicant (X) who has ever applied for admission is taken as an instance of database ($\mathscr{D}$). Part of the discriminatory behavior towards women can be explained by attributes like program preference that are correlated with both the sensitive attribute and the decision. Thus, program preference is an *explanatory attribute* ($e \in E$) that is correlated with the sensitive attribute (S), and gives some objective information about the class label C. While selection of explanatory attributes is often debatable, we assume that they are nominated by the domain experts externally. We restrict this work to nominal explanatory attributes only.

The task is to learn a classifier $\mathscr{F} : \mathscr{X} \to \{C^+, C^-\}$ from the given discriminatory data $\mathscr{D}$ that does not make discriminatory decisions w.r.t. sensitive attribute(s) while predicting future instances. As the convention for this problem setting, the performance of the discrimination-aware classification methods is determined by reporting their accuracy and discrimination. Ideally, accuracy should suffer the least as discrimination is reduced to zero.

3.1.2 Measuring Discrimination

Several measures of discrimination have been proposed in the discrimination-aware classification research. In this work, we distinguish between two types of discrimination: overall and illegal discrimination. We use the definitions of [1–3, 48] for overall discrimination. Overall discrimination quantifies the difference in treatment (i.e., labelings) between deprived and favored groups on the basis of sensitive attributes only, ignoring all other explanations for the differential treatment.

Definition 1. ***(Overall Discrimination, D_{all}):*** *Given a labeled dataset $\mathscr{D} = \{X_i, C_i\}_{i=1}^N$, sensitive attributes S and their respective domains describing instances in deprived and favored groups ($\mathscr{X}^d$ and $\mathscr{X}^f$), the discrimination in dataset $\mathscr{D}$ w.r.t. sensitive attributes S, denoted by $D_{all}(\mathscr{D}, S)$, is defined*

as:

$$D_{all}(\mathscr{D},S) := \frac{|\{X \in \mathscr{X}^f, C = C^+\}|}{|\{X \in \mathscr{X}^f\}|} - \frac{|\{X \in \mathscr{X}^d, C = C^+\}|}{|\{X \in \mathscr{X}^d\}|} .$$

In probabilities, this is equivalent to $p_{\mathscr{D}}(C^+|\mathscr{X}^f) - p_{\mathscr{D}}(C^+|\mathscr{X}^d)$.

When clear from the context, we will omit the subscript and parameters in the notation, and more often, refer to this measure as overall discrimination. It is equal to the difference of the probability of acceptance for the favored community $p_{\mathscr{D}}(C^+|\mathscr{X}^f)$ and the deprived community $p_{\mathscr{D}}(C^+|\mathscr{X}^d)$.

Overall discrimination disregards other plausible reasons for the differential treatment between the two groups. As such, this measure is appropriate when discrimination w.r.t. sensitive attribute alone needs to be controlled (e.g., when stipulated by law). For instance, recently a ruling of European Court of Justice declared that varied insurance premiums on the basis of gender of drivers would be considered discrimination and violation of law [75]. Thus, despite knowing from historical records that male drives have riskier driving habits and are more likely to be involved in accidents [1], insurance companies are not allowed to use this information and are bound by law to treat both male and female drivers equally. The measure of overall discrimination applies to such scenarios.

In other scenarios, part of the differential treatment between deprived and favored groups can be explained by other attributes. For instance, low acceptance rate of female applicants to a university can be explained by their preference for more competitive disciplines (e.g., medicine). In such a scenario, discrimination that cannot be explained is called illegal discrimination. It quantifies preferential treatment on the basis of sensitive attributes without any plausible reason. We use the definition of [48] to measure illegal discrimination.

Definition 2. ***(Illegal Discrimination,*** $D_{illegal}$***):*** *Given a discriminatory labeled dataset* $\mathscr{D}$*, sensitive attributes S distinguishing between instances in deprived and favored groups (*$\mathscr{X}^d$ *and* $\mathscr{X}^f$*), and explanatory attributes E. Let* $dom(E) = \{1,\ldots,k\}$ *be the domain of E. The explainable discrimination* $D_{expl}(\mathscr{D},S,E)$ *in dataset* $\mathscr{D}$ *w.r.t. the sensitive attributes S and the explanatory attributes E is calculated as follows:*

$$D_{expl}(\mathscr{D},S,E) := \sum_{i=1}^{k} \Big(p(E_i|\mathscr{X}^f) - p(E_i|\mathscr{X}^d)\Big)\, p^\star(C^+|E_i)$$

where

$$p^\star(C^+|E_i) := \frac{P(C^+|E_i,\mathscr{X}^f) + p(C^+|E_i,\mathscr{X}^d)}{2}.$$

Then, the illegal discrimination $D_{illegal}(\mathscr{D},S,E)$ *in dataset* $\mathscr{D}$ *w.r.t. the sensitive attributes S and the explanatory attributes E is given by:*

$$D_{illegal}(\mathscr{D},S,E) := D_{all}(\mathscr{D},S) - D_{expl}(\mathscr{D},S,E)$$

Here, $D_{all}(\cdot)$ *is the overall discrimination in* $\mathscr{D}$ *as defined in Definition 1. [48].*

[1] `http://www.insurance.com/auto-insurance/safety/are-men-better-drivers-than-women.aspx`

When clear from the context, we will omit the subscript and the parameters in the notation, and more often, refer to this measure as illegal discrimination.

The above measures calculate the discrimination in any given labeled dataset. We can use the same discrimination measures to calculate the discrimination of a classifier by assuming the given dataset to be a test dataset labeled by the classifier. In many practical applications, the number of instances in $\mathscr{X}^d$ is less than the number of instances in $\mathscr{X}^f$ in the biased training dataset, i.e., $|\mathscr{X}^d| < |\mathscr{X}^f|$. Consequently, due to sample imbalance and classifier over-fitting it is often the case that $D_{all}(\mathscr{F},S) > D_{all}(\mathscr{D},S)$ where $D_{all}(\mathscr{F},S)$ represents the discrimination in the predictions of a classifier $\mathscr{F}$ learnt over biased data. This fact highlights the inadequacy of discrimination prevention by just modifying the training data as proposed by some earlier discrimination-aware methods.

3.2 Methodology for Discrimination Control

In this section, we present a methodology for social discrimination control that exploits the reject option in classification. The reject option in classification discards a predicted label when it is found to be highly uncertain or ambiguous. This rejection provides an opportunity for relabeling the instance in a manner that reduces discrimination while maintaining prediction accuracy over the biased dataset. We present three reject option based solutions for discrimination control: Probabilistic Rejection (PR), Ensemble Rejection (ER), and Situatinal Rejection (SR). We start by defining our discrimination model underlying the methodology.

3.2.1 Discrimination Model: Reject Option in Classification

Recently, a discrimination model has been presented that describes the process leading to biased labeling of instances during classification [48]. According to this model, a decision maker obtains a preliminary score m quantifying the worthiness of an individual X without relying upon the sensitive attributes describing X. Thus, this score is evaluated objectively and on merit. Then, the discrimination bias $b \geq 0$ is introduced by looking at the sensitive attributes and their values for the individual. A uniform bias is either added (positive bias) or subtracted (negative bias) from the merit-based score m, to yield the overall score $m^* = m \pm b$. In general, the bias can vary for different individuals, however, in this study we assume a uniform bias b is added/subtracted to favor/discriminate the unprotected/protected group instances. In the social sciences, this bias is referred to as an unconscious bias [76]. The final decision of individual X is made by using score m^*.

This discriminatory decision making process impacts the decision of instances that are close to the decision boundary according to their score m. It is quite intuitive that the addition or subtraction of the bias b will not affect the decision of instances with very high or low merit-based scores m.

In our setting, we already have a discriminatory dataset $\mathscr{D}$ that captures information about the decision making process. We know key attributes of the classification problem including the sensitive attributes S, the explanatory attributes E, and the class label C. However, we do not have a clear

distinction between objective or merit-based and biased contributions in the labeling process. As is required by law, the sensitive attributes cannot be used in learning and prediction. Nonetheless, because of correlation between sensitive and explanatory attributes the classifier learns the bias through the explanatory attributes. This phenomenon has been demonstrated in previous works [77].

Given the above observations, we propose the following discrimination model. Let $\mathscr{F}$ be a classifier (or a classifier ensemble) learned over the discriminatory dataset $\mathscr{D}$ without considering the sensitive attributes S, and let $0 \leq \mathscr{F}(X,C^+) \leq 1$ be the score (e.g., posterior probability or confidence) for label C^+ of instance X produced by $\mathscr{F}$ and $\mathscr{F}(X,C^-) = 1 - \mathscr{F}(X,C^+)$. Then, instance $X \in \mathscr{X}^d$ with label C^- is likely to be discriminated when $\mathscr{F}(X,C^+) \geq 0.5 - \eta$ where $0 < \eta \leq 0.5$ is a parameter that specifies the bias in the dataset. Similarly, instance $X \in \mathscr{X}^f$ with label C^+ is likely to be favored when $\mathscr{F}(X,C^+) \leq 0.5 + \eta$. Otherwise, instance X is neither discriminated nor favored according to this model.

The classifier's score $\mathscr{F}(X,C^+)$ and the parameter η correspond roughly to m^* and b, respectively, in the basic discrimination model outlined earlier. The value of η controls the region on both sides of the classifier's decision boundary within which classification scores are considered ambiguous; instances whose scores lie in this region are not assigned a label by the classifier (i.e., their labels are rejected) and are considered likely to be the result of discriminatory practices captured in the dataset.

The parameter η can be estimated automatically when a non-discriminatory dataset is available. Alternatively, a domain expert can analyze potentially discriminated/favored instances close to the decision boundary to fix an appropriate value for η.

Definition 3. ***(Discrimination and Favoritism Potential):*** *The Discrimination Potential of an instance $X \in \mathscr{X}^d$ with label C^- in a discriminatory dataset $\mathscr{D}$ is defined as*

$$DP(X \in \mathscr{X}^d) = \mathscr{F}(X,C^+) - (0.5 - \eta) \geq 0$$

Similarly, the Favoritism Potential of an instance $X \in \mathscr{X}^f$ with label C^+ in a discriminatory dataset $\mathscr{D}$ is defined as

$$FP(X \in \mathscr{X}^f) = (0.5 + \eta) - \mathscr{F}(X,C^+) \geq 0$$

Here, $\mathscr{F}(X,C^+)$ is the score for label C^+ for instance X produced by classifier $\mathscr{F}$ learned over the discriminatory dataset $\mathscr{D}$.

$DP(\cdot)$ and $FP(\cdot)$ range from 0 to 0.5 with higher values signifying greater potential of being discriminated or favored in the dataset. The expressions for computing DP and FP can return a negative value which implies that no discrimination or favoritism exists.

This discrimination model can be used for both discrimination discovery and discrimination prevention. The Discrimination and Favoritism Potentials described above allow easy identification and ranking of instances that have potentially biased decisions in a dataset. In the following sections, we present our discrimination control solutions based on our discrimination model.

Decision theory tells us that when we utilize the region of low prediction confidence to relabel instances for reduced discrimination, the impact on accuracy will be minimum. This idea is adopted in our solution for discrimination-aware classification.

Similarly, we know from decision theory that disagreement among an ensemble of classifiers identifies a region of low prediction confidence. This idea is exploited in our second solution for discrimination-aware classification.

3.2.2 Probabilistic Rejection (PR)

Our first reject option based solution for discrimination control, called Probabilistic Rejection(PR), utilizes posterior probabilities produced by one or more probabilistic classifiers to identify instances with high label uncertainty. These instances are then labeled in a manner that neutralizes the effect of discrimination. Based on the discrimination model introduced in the previous section, PR embodies strong theoretical concepts to provide excellent control over the accuracy-discrimination trade-off for future classifications.

Before proceeding further, it is worth re-emphasizing that effective discrimination control in our setting (only discriminatory dataset available) is possible only when group membership of individuals is known. Knowledge of this information is also necessary for litigation processing and affirmative action.

Labeling Strategy

Traditionally, a learned classifier assigns an instance to the class with the highest posterior probability. PR deviates from this traditional decision rule and gives the idea of a critical region in which instances belonging to deprived and favored groups are labeled with desirable and undesirable labels, respectively. We first present PR for single and multiple classifiers and then relate PR with decision theory for interpretation and control.

Consider a single classifier, and let $p(C^+|X)$ be the posterior probability for instance X produced by this classifier. When $p(C^+|X)$ is close to 1 or 0 then the label for instance X is specified with a high degree of certainty. On the other hand, when $p(C^+|X)$ is close to 0.5 then the label for instance X is more uncertain. Probabilistic rejection is adopted for all instances for which $\max[p(C^+|X), 1 - p(C^+|X)] \leq \theta$ where $(0.5 < \theta < 1)$. These instances, which lie within the *critical region*, are not assigned labels (or are labeled as 'reject'). The labels for instances in the critical region (rejected instances) are considered to be ambiguous and influenced by biases. Note that $\eta = \theta - 0.5$ relates the parameter θ with the parameter η introduced in the discrimination model.

To reduce discrimination, these rejected instances are labeled as follows; if the instance is from the deprived group ($\mathscr{X}^d$) then label it as C^+ otherwise label it as C^-.

The instances outside the critical region are classified according to the standard decision rule, i.e., if $p(C^+|X) > p(C^-|X)$ then C^+ will be assigned to instance X; otherwise, C^- will be assigned to instance X.

Probabilistic rejection is not restricted to work with a single classifier; it can also be used for an ensemble of probabilistic classifiers. In our problem setting of discrimination-aware classification,

a classifier ensemble can be thought of as a pool of experts with varying characteristics and biases – their combined output is expected to be more reliable w.r.t. both accuracy and discrimination.

Let $\mathscr{F}_k$ $(k = 1,\ldots,K)$ denote the kth classifier in an ensemble of $K > 1$ classifiers, and $p(C,\mathscr{F}_k|X)$ be the posterior probability of classification of instance X produced by classifier $\mathscr{F}_k$. The posterior probability of classification of the ensemble $p(C|X)$ is given by

$$p(C|X) = \sum_{k=1}^{K} p(C|X,\mathscr{F}_k)p(\mathscr{F}_k) \tag{3.1}$$

The prior probability of a classifier, $p(\mathscr{F}_k)$, can be taken to be proportional to the accuracy of that classifier on the data. Or, if such information is considered uninformative, the prior probability distribution can be taken to be uniform, in which case, the posterior probability of the ensemble is simply the average of the posterior probabilities of each classifier in the ensemble.

Given the posterior probability of an ensemble $p(C|X)$, PR proceeds in the manner as discussed for a single classifier above. This labeling strategy will ensure that only higher risk instances are rejected and thus its impact on accuracy of the classifier is a minimum. PR's methodology is illustrated in Figure 3.1. PR algorithm is shown in Algorithm 4. The inputs required for PR are one or more probabilistic classifiers trained on discriminatory dataset, information for identifying deprived group instances, and parameter θ. It outputs discrimination-aware labels for new instances. Instance labeling is distinguished between two regions. In the critical region, instances are labeled in a manner to neutralize discrimination (lines 6 to 10), while instances outside the critical region are labeled using the standard decision rule (lines 11 to 16).

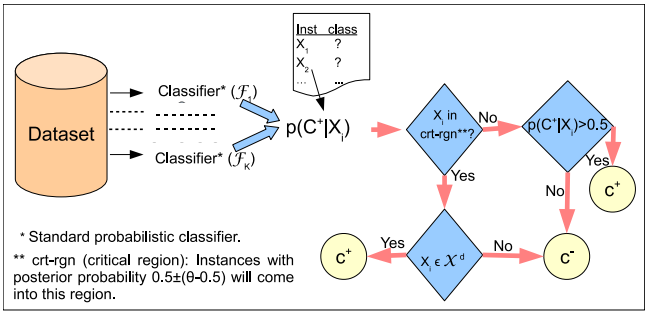


Figure 3.1: Framework of Probabilistic Rejections (PR)

Decision Theoretic Interpretation

In this section, we develop a decision theoretic understanding of PR. The expected loss of a single classifier or an ensemble of classifiers ($\mathscr{F}$) that produces posterior probabilities $p(C^+|X)$ and $p(C^-|X) = 1 - p(C^+|X)$ for instance X is given by

$$\mathscr{E}[L] = \sum_{\{X\in\mathscr{X}|\mathscr{F}(X)=C^+\}} L_{-,+}p(C^-|X)p(X) + \sum_{\{X\in\mathscr{X}|\mathscr{F}(X)=C^-\}} L_{+,-}p(C^+|X)p(X). \tag{3.2}$$

Algorithm 4 Probabilistic Rejections (PR)

Input: $\{\mathscr{F}_k\}_{k=1}^{K}$ ($K \geq 1$ probabilistic classifiers trained on $\mathscr{D}$), $\mathscr{X}$ (test set), $\mathscr{X}^d$ (deprived group), θ
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathscr{X}$)
for $i = 1 \rightarrow M$ **do**
    $p(C^+|X_i) \leftarrow$ posterior probability for C^+ produced by classifier(s)
    if $\max(p(C^+|X_i), 1 - p(C^+|X_i)) \leq \theta$ **then**
        ** Critical region **
        if $X_i \in \mathscr{X}^d$ **then**
            $C_i \leftarrow C^+$
        else
            $C_i \leftarrow C^-$
    else
        ** Standard decision rule **
        if $p(C^+|X_i) \geq p(C^-|X_i)$ **then**
            $C_i \leftarrow C^+$
        else
            $C_i \leftarrow C^-$

Here, $L_{+,-}$ quantifies the loss incurred in classifying a positive instance as negative. These quantities are typically given in a loss matrix, with rows representing actual labels and columns giving predicted labels (Table 3.1). There is no loss when the predicted and actual labels match; hence, $L_{+,+} = L_{-,-} = 0$ while $L_{+,-}, L_{-,+} > 0$.

The best label for each instance X, that ensures the minimum expected loss of classification (Equation 3.2), is given by the $j \in \{+,-\}$ that minimizes [78]:

$$L_{+,j}p(C^+|X) + L_{-,j}(1 - p(C^+|X)). \tag{3.3}$$

When all classification errors incur a constant loss (e.g., $L_{+,-} = L_{-,+}$, then the above decision rule assigns each instance X to the label whose posterior probability is the largest. This is the standard decision rule that ensures the lowest loss in the accuracy of classification.

Table 3.1: Loss matrix

Actual↓, Predicted→	C^+	C^-	C^r
C^+	$L_{+,+}$	$L_{+,-}$	$L_{+,r}$
C^-	$L_{-,+}$	$L_{-,-}$	$L_{-,r}$

The reject option in classification is invoked when $max[p(C^+|X), 1 - p(C^+|X)] < \theta$. From Equation 3.2, it is clear that even when all rejected instances (say R instances) are misclassified the increase in expected loss is a minimum as compared to any other set of R misclassified instances from a given dataset. This is because the rejected instances have a low maximum posterior probability. The labeling strategy of Probabilistic Rejection (PR), however, only relabels deprived group instances with negative labels and favored group instances with positive labels. This strategy reduces discrimination by decreasing the dependence of the sensitive attributes on the class attribute without impacting the

dependence of other attributes on the class attributes. Thus, PR reduces illegal discrimination first while maintaining the explainable discrimination.

In PR, the trade-off between accuracy and discrimination is controlled by θ; in general the larger the value of θ the greater the reduction in classifier discrimination, as more deprived and favored group instances are likely to be labeled with C^+ and C^-, respectively. For any given value of θ, the expected reduction in accuracy is the minimum possible as pointed out in the preceding paragraph. To achieve a specified discrimination level, the value of θ can be determined by using a validation dataset.

Typically in classification, a uniform cost or loss is associated with all errors, irrespective of them being false positives or false negatives. That is, $L_{+,-} = L_{-,+}$ (see Table 3.1), and conveniently this loss can be taken to be 1 unit. The reject option can be invoked by considering a third prediction label (C^r for reject) and taking $L_{+,r} = L_{-,r} = 1 - \theta$. Thus, the loss for rejecting an instance depends upon the value of θ – the larger its value is, the smaller the loss for rejection.

The PR labeling strategy can be interpreted via loss matrices. Consider a separate 2×2 (no C^r label) loss matrix for deprived and favored group instances (Table 3.2). The discrimination reducing and accuracy preserving classification is achieved when $L^d_{+,-} = L^f_{-,+} = \theta/(1-\theta)$, with the other values remaining unchanged from the usual loss matrix (Table 3.1).

Table 3.2: Loss matrices for probabilistic rejection (PR). The left matrix is for deprived instances and the right is for favored instances.

	Deprived Insts		Favored Insts	
Actual↓, Predicted→	C^+	C^-	C^+	C^-
C^+	0	$\frac{\theta}{1-\theta}$	0	1
C^-	1	0	$\frac{\theta}{1-\theta}$	0

Thus, PR can be interpreted as a cost-based prediction method in which the cost or loss of misclassifying a deprived group instance as negative is $\theta/(1-\theta)$ times that of misclassifying it as positive. A similar statement can be made for favored group instances. For example, when $\theta = 0.6$ then a 50% higher loss is associated with one type of error as compared to the other.

3.2.3 Ensemble Rejection (ER)

Our second reject option based solution for discrimination-aware classification, called Ensemble Rejection (ER), relabels instances on which an ensemble of classifiers disagrees significantly. Unlike PR, ER is not restricted to probabilistic classifiers only; an ensemble comprising of any type of classifier can be used in this solution. As pointed out earlier, classifier ensembles often produce robust classifications by taking advantage of the diversity of member classifiers. Furthermore, a classifier ensemble mimics practical decision making where a panel of experts converge on an outcome (e.g., acceptance or rejection) for an individual. For discrimination prevention and control, ER provides additional flexibility in the choice of a classification system.

Labeling Strategy

Typically, a classifier ensemble labels a new instance with the majority class label (majority-vote rule). Ensemble Rejection (ER) deviates from this standard rule to neutralize the effect of discrimination. Specifically, it labels instances on which member classifiers disagree significantly in a manner that reduces discrimination.

Formally, let $K \geq 2$ be the number of classifiers in an ensemble $\mathscr{F}$, and $0 \leq K^+ \leq K$ be the number of classifiers in the ensemble predicting label C^+ for an instance X. Then, the confidence of the C^+ label produced by the classifier ensemble $\mathscr{F}$ is defined as

$$conf(\mathscr{F}, X, C^+) = K^+/K.$$

Likewise, the confidence of the C^- label is given by $conf(\mathscr{F}, X, C^-) = 1 - conf(\mathscr{F}, X, C^+)$. Given these confidence values, ER labels instance X using the following decision rule: if $\max[conf(\mathscr{F}, X, C^+), conf(\mathscr{F}, X, C^-)) \leq \theta$ then instance X) is assigned the desired label (C^+) if it belongs to the deprived group and the undesired label (C^-) if it belongs to the favored group. Otherwise (i.e., when $\max[conf(\mathscr{F}, X, C^+), conf(\mathscr{F}, X, C^-)) > \theta$), the standard majority-vote label is assigned to instance X.

As in PR the parameter θ, which varies from 0.5 to 1, controls the critical region in input space where the standard decision rule (majority-vote) is rejected in favor of the discrimination-aware rule to reduce discrimination. A value of $\theta = 0.5$ means that the standard majority-vote rule is utilized for all instances, while a value of $\theta = 1$ means that the majority-vote label is rejected for all instances. Thus, θ controls the trade-off between discrimination and accuracy of a specific classifier ensemble.

A special case of the ER labeling strategy is when θ is just less than one (e.g., $\theta = 0.99$). In this case, when all member classifiers predict the same label for a given instance, the agreed class label is assigned to it; otherwise, if the instance belongs to the deprived group it is assigned the C^+ label and if the instance belongs to the favored group it is given the C^- label. In other words, all instances for which the member classifiers disagree are rejected and labeled to reduce discrimination.

Based on our discrimination model, the ER labeling strategy considers that instances on which more member classifiers disagree are closer to the decision boundary and are more likely to be discriminated. We can draw a parallel between an ensemble and an admission committee: assume that some members of the committee are biased against female applicants and try to reject their applications. Hence, it is very likely that these members will only be able to affect the applicants close to the decision boundary because the highly qualified female applicants cannot be rejected due to their overall high score. If we consider member classifiers of an ensemble as admission committee members, then having more classifiers in the ensemble or increasing the acceptance confidence may neutralize the discriminatory effect of ensemble due to the fair classifiers. Thus, using classifier ensembles is a natural fit to the solution of discrimination-aware classification problem.

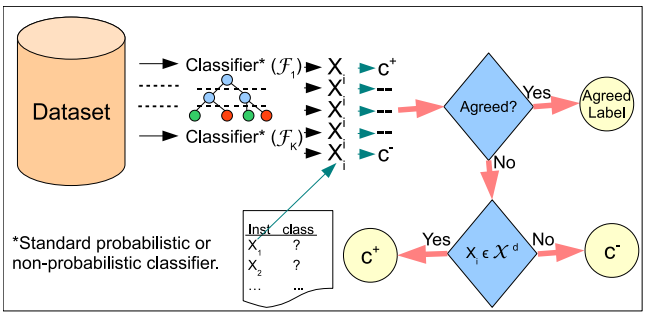


Figure 3.2: Framework of Discrimination-Aware Ensemble (DAE).

Controlling Discrimination

There are two approaches towards controlling discrimination with ER. The first approach assumes a fixed classifier ensemble. In this approach, the trade-off between discrimination and accuracy is controlled by varying the value of θ. This approach and the corresponding discrimination-accuracy behavior is similar to that for PR.

The second approach assumes that an instance is rejected for discrimination-aware labeling whenever a given classifier ensemble disagrees on its label. In this approach, the trade-off between accuracy and discrimination is controlled by varying the composition of the ensemble. The question now is: which members should we choose and how does this impact discrimination? The accuracy-discrimination performance of a given ensemble with ER depends upon the disagreement among the member classifiers, which is defined as:

Definition 4. ***(Disagreement of a Classifier Ensemble):*** *Given a classifier ensemble* $\{\mathscr{F}_k\}_{k=1}^{K}$ $(K > 1)$ *trained on discriminatory dataset* $\mathscr{D} = \{X_i, C_i\}_{i=1}^{N}$*, the disagreement of the ensemble w.r.t. dataset* $\mathscr{D}$*, denoted as* $disagr_{\mathscr{D}}$*, is defined as:*

$$disagr_{\mathscr{D}} = \frac{|\{X_i | \exists j,k \; \mathscr{F}_j(X_i) \neq \mathscr{F}_k(X_i)\}|}{|\{X_i\}|}$$

When clear from the context, we will drop the subscript or simply use disagreement while referring to this measure.

Equivalently, $disagr_{\mathscr{D}} = d/N$, where d is the number of instances on which the ensemble disagrees. If a is the number of instances on which the ensemble agrees, then $a + d = N$. However, it is worth noticing that not all instances in a are correctly classified; the ensemble can agree on an incorrect label for an instance. ER's methodology is illustrated in Figure 3.2 and algorithm is shown in Algorithm 5.

In general, the higher the disagreement of an ensemble on a given dataset, the lower will be the discrimination produced by this ensemble with ER on new instances since the ensemble will disagree on more instances and all such instances belonging to the deprived group are labeled with C^+ and the rest are labeled with C^-. Disagreement, as defined above, can be considered to be a measure of ensemble diversity as well. Ensemble diversity has been shown to be positively correlated with ensemble accuracy determined via majority vote [79]. Another measure of ensemble diversity is average pairwise correlation between member classifiers. In [80], error bounds have been developed for classifier ensemble under reject option as a function of correlation. Therefore, a key thumb rule to

remember while selecting member classifiers of an ensemble for ER is: the more diverse the member classifiers are, the higher will be the disagreement (or lower will be the correlation) among them, and the greater will be the reduction in discrimination. This means that we can control the discrimination of an ensemble with ER by changing the diversity of member classifiers. To select an ensemble with ER having a specific discrimination level, a validation dataset can be used.

The trade-off between accuracy and discrimination will depend upon both disagreement and the number of instances in a that are incorrectly classified.

Algorithm 5 Discrimination-Aware Ensemble (DAE)

Input: $\{\mathscr{F}_k\}_{k=1}^{K}$ ($K > 1$ classifiers trained on $\mathscr{D}$), $\mathscr{X}$ (test set), $\mathscr{X}^d$ (deprived group)
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathscr{X}$)
for $i = 1 \rightarrow M$ **do**
    if $\mathscr{F}_j(X_i) = \mathscr{F}_k(X_i)\ \forall j,k$ **then**
        ** Agreement **
        $C_i \leftarrow \mathscr{F}_1(X_i)$
    else
        ** Disagreement **
        if $X_i \in \mathscr{X}^d$ **then**
            $C_i \leftarrow C^+$
        else
            $C_i \leftarrow C^-$

3.2.4 Situational Rejection (SR)

Our third solution for discrimination control, called Situational Rejection (SR), combines PR or ER with a legally-grounded procedure of *situation testing*. SR includes an additional check, based on a local model of classification, for instances that are rejected and relabeled in PR or ER. As such, SR is more careful in relabeling and hence less 'brute force' in its labeling strategy. Furthermore, SR provides additional insights into the prevalence of discrimination and its control in future predictions.

Labeling Strategy

Situational rejection's labeling strategy for discrimination control deviates from that for PR and ER with the addition of situation testing. Situation testing or situational judgement test is a systematic procedure employed in the legal domain for determining the response of a decision maker towards an applicant's suitability for a benefit or service under different settings. In this procedure, a hypothetical situation is assumed where a pair of applicants with similar qualifications (e.g., education, experience) but from different sensitive groups (e.g., race)apply for certain benefits (e.g., job) simultaneously. The different outcomes of such a controlled experiment can assist victims of discrimination to establish the evidence against the discriminatory practices w.r.t. certain sensitive characteristics [52, 81, 82]. Specifically, if it is found that the victim was denied the benefits while his pair was awarded the benefits then this provides evidence for the discriminatory practice.

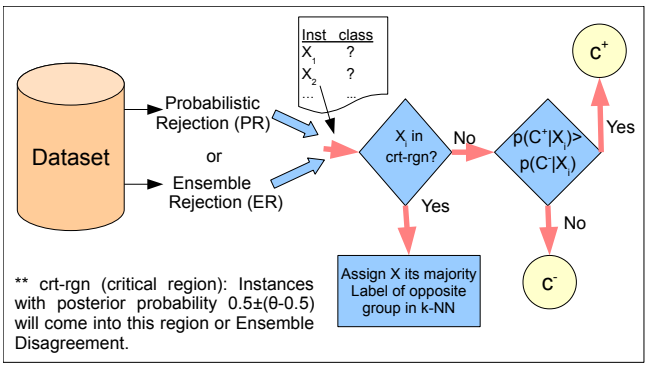


Figure 3.3: Framework of Situation Testing.

We model situation testing via a k-nearest neighbor (k-NN) classifier [83]. This local model of classification is applied to each instance that is rejected by a probabilistic classifier or a classifier ensemble learned on the discriminatory data (i.e., the instances in the critical region produced in PR and ER). A rejected instance is compared with its neighbors and is labeled w.r.t. the majority class of its neighbors from the opposite group of sensitive attribute. For instance, a rejected female will be labeled according to majority class of the k-nearest male neighbors of this rejected female. The intuition of this method is to relabel only those rejected instances that have been treated differently as compared to their peers rather than relabeling all the rejected instances.

Figure 3.3 represents the situation testing framework. SR changes the labels of selected deprived and favored group instances in the critical region it is less 'forceful' in reducing discrimination. As such, in general, to achieve the same level of discrimination a larger critical region may be required. It is also worth noting that SR can be applied to all instances and not just to those in the critical region.

In the legal domain, situation testing is a systematic research procedure for creating controlled experiments analyzing decision maker's candid responses to applicant's personal characteristics. In situation testing, pairs of research assistants undergo the same kind of selection, for example they apply for the same job, they present themselves at the same night club, and so on. Within each pair, applicant characteristics likely to be related to the situation (characteristics related to a worker's productivity on the job in the first case, look, age and the like in the second case) are made equal by selecting, training, and credentialing testers to appear equally qualified for the activity. Simultaneously, membership to a protected group is experimentally manipulated by pairing testers who differ in membership for example, a black and a white, a male and a female, and so on. Situation testing is being experimented worldwide as one of the tools that can assist victims to establish that discrimination may have occurred [52, 81, 82].

3.2.5 Summary of Rejection Option Classifiers

Our discrimination control methodology is outlined in Algorithm 6. The algorithm takes as input a classifier or classifier ensemble ($\mathscr{F}$) trained on a discriminatory dataset ($\mathscr{D}$), test instances to be classified (X_i), knowledge of the sensitive attribute in the training and test datasets, parameter θ, name of the solution to be used (*Solution*), and neighborhood size (k, for SR only). The algorithm outputs

discrimination-aware labels ($C_i \in \{C^+, C^-\}$) to test instances. For ER with disagreement θ is set close to 1. An instance is rejected when its predicted label score (confidence or posterior probability) is low according to the threshold θ. For PR and ER, rejected deprived group instances are given label C^+ and rejected favored group instances are given label C^-. In SR, rejected instances are given the majority label of the opposite group instances within the k neighbors of the instances. Instances that are not rejected are given standard classifier labels.

Our methodology is computationally efficient. Besides training, which is done once and prior to the application of our methodology, the processing time and space complexity is linear in the number of test instances.

Algorithm 6 Summary of Rejection Option Based Classifiers– PR, ER, and SR

Input: $\mathscr{F}$ (classifier or classifier ensemble trained on $\mathscr{D}$), $\mathscr{X}$ (test set), $\mathscr{X}^d, \mathscr{X}^f$ (deprived and favored groups), θ, *Solution* (*PR*, *ER*, or *SR*), k (neighborhood size, for *SR* only)
Output: $\{C_i\}_{i=1}^M$ (labels for instances in $\mathscr{X}$)
for $i = 1 \rightarrow M$ **do**
    $Score \leftarrow \mathscr{F}(X_i, C^+)$
    if $\max(Score, 1 - Score) \leq \theta$ **then**
        if $Solution = PR \vee ER$ **then**
            if $X_i \in \mathscr{X}^d$ **then**
                $C_i \leftarrow C^+$
            else
                $C_i \leftarrow C^-$
        else
            $C_i \leftarrow$ majority label of opposite group in k-NN of X_i
    else
        if $Score \geq 1 - Score$ **then**
            $C_i \leftarrow C^+$
        else
            $C_i \leftarrow C^-$

3.3 Experimental Evaluation

In this section, we discuss the evaluation of our methodology for discrimination control on four real-world datasets. We compare the performance of our solutions with previously proposed discrimination-aware classification methods. Since our solutions are not restricted to any specific classifier, we consider several standard classifiers for discrimination-aware classification (identifying label of each classifier is given in parenthesis): naive Bayes (**NBS**), logistic regression (**Logistic**), k-nearest neighbor (**IBK**), and decision tree (**J48**). The first and second classifiers are generative and discriminative probabilistic classifiers, respectively, while the third is an instance-based classifier with well-defined probabilistic interpretation. We also show results with decision trees, which is an information theoretic classifier, since they have been used popularly in previous discrimination-aware classification research. Besides the above classifiers, we tried many other classifiers as well, including support vector machines (**SVM**), but do not report all results for ease of understanding.

In summary, we present and discuss the results of the following experiments for preventing overall and illegal discrimination:

1. PR: Probabilistic Rejections using single and multiple probabilistic classifiers, identified as **PR (classifier)** and **PR (1st classifier+2nd classifier+...)**, respectively.
2. ER: Ensemble Rejection with two or more classifiers, identified as **ER (1st classifier+2nd classifier+...)**.
3. SR: Situational Rejection using single and multiple probabilistic classifiers, identified as **SR (classifier)** and **SR (1st classifier+2nd classifier+...)**, respectively.
4. Comparison of our solutions' results with those of current state-of-the-art discrimination-aware classification methods, identified as **Prev Methods**.
5. Performance of our solutions (PR, ER, and SR) for illegal discrimination prevention.
6. Evaluation of PR w.r.t. different and multiple sensitive attributes.
7. Evaluation of PR on test dataset with less discrimination.

Datasets: We conduct our experiments on four real-world datasets: *Adult* [84], *Communities and Crime* [84], and *Dutch Census of 1971 and 2001* [85] datasets. Table 3.3 gives the important characteristics of these datasets such as number of instances, number of instances belonging to deprived group ($\mathscr{X}^d$), number of attributes in the dataset, class attribute defining the desirable and undesirable labels, sensitive attribute (SA), and overall discrimination (calculated using Equation 1). For experiments on less discriminatory test sets (reported in Figure 3.10), we change some settings in the *Dutch Census* datasets as follows: use the attribute *economic status* as class attribute rather than *occupation* as class attribute of the *Dutch Census of 2001* dataset and by removing some attributes like *current economic activity* and *occupation* from these experiments to make both datasets (Dutch 1971 and 2001) consistent w.r.t. codings. The discrimination in the *Dutch Census of 2001* dataset w.r.t. *economic status* as class attribute is 28.23%.

Table 3.3: Key characteristics of datasets.

Dataset	Inst.	$\lvert\mathscr{X}^d\rvert$	Attr.	Class	SA	disc%
Adult	16 281	5 421	14	Income	sex	19.45
Communities	1 994	1 024	122	violent criminal	race	43.14
Dutch 71	99 772	51 658	9	economic status	sex	58.66
Dutch 01	15 150	7 603	12	occupation	sex	29.85

All results reported in the chapter (excluding those reported in Figure 3.10) are obtained using *10-fold cross-validation* and each point in the figures represents the result of an independent experiment. The datasets with detailed description and source code of implementations used in this chapter are available at [2].

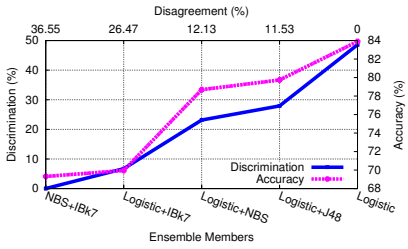
[2] `https://sites.google.com/site/discriminationcode/`

3.3.1 Removing the Sensitive Attribute

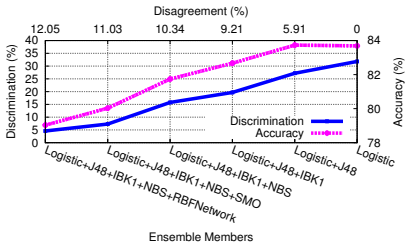
First we report the results of the experiments to show that the straight forward solution of just removing the sensitive attributes does not work as the classifier tends to pick the indirect discrimination from the other correlated attributes of sensitive attributes. Table 3.4 shows the result of experiments to validate this claim. We learn a decision tree classifier over the above mentioned three real world datasets with and without using the sensitive attribute. We can observe from the results given in Table 3.4 that the removal of sensitive attribute has a little impact on the reduction of discrimination. However the Dutch 2001 Census data is one exception where the removal of sensitive attribute has relatively more impact due to the weak correlation of the sensitive attribute with the other attributes. The results shown in this section demonstrate that this straight forward solution does not work and clearly motivate to use more sophisticated discrimination-aware techniques to ensure discrimination-free classification as we do next.

Table 3.4: Removing the sensitive attribute from classification does not ensure discrimination-free classification.

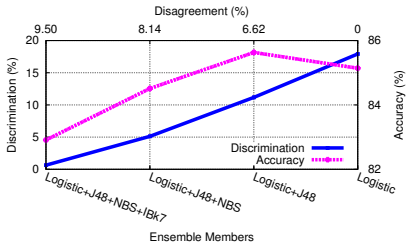
Dataset	**With S**	**Without S**
Adult	16.48%	16.65%
Communities and Crime	40.14%	38.07%
Dutch 2001 Census	34.91%	17.92%



Baseline Disc=43.14 Acc=59.58
(a) Communities and Crime



Baseline Disc=29.85 Acc=52.39
(b) Dutch Census of 2001



Disc=19.3 Acc=76.3
(c) Adult

Figure 3.4: Discrimination-accuracy trade-off of ER (disagreement based) on three datasets. For each dataset, several classifier ensembles are shown with their accuracy and discrimination.

3.3.2 Overall Discrimination Control

In this section, we show that our proposed solutions prevent effectively overall discrimination in future predictions. We also show that our proposed solutions outperform the current state-of-the-art methods over three real-world datasets (the Dutch 71 dataset is only used in Section 3.3.5).

Results of PR and SR

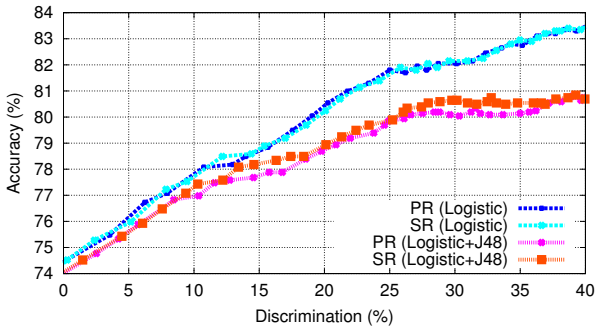
Figure 3.5 shows the results of our experiments with PR and SR (PR combined situation testing) on three datasets (labeled (a), (b), (c)). The x- and y-axis of these plots represent classifiers' discrimination and accuracy respectively, and each point is for a specific value of θ which is varied from 0.5 to a maximum value (usually around 0.9). It is observed that as the value of θ is increased, the discrimination reduces to zero. Furthermore, the reduction in discrimination with the increase in θ is generally smooth and consistent across datasets and classifier(s). Thus, the discrimination level of PR and SR can be controlled easily by varying the value of θ. The generally small decrease in accuracy for specific values of θ makes PR and SR robust solutions appropriate for practical discrimination-aware classification.

We know that the performance of classifiers varies over different datasets; the best performing classifier over one dataset can give poor performance on another one. Figure 3.5 demonstrates this fact and shows that PR and SR can be used with a selected single classifier or classifier ensemble to ensure the best performances. For instance, both PR and SR give better performance with single classifiers over the Communities and Crime dataset (Figure 3.5 (a)). However, PR with an ensemble of logistic regression and J48 outperforms the other tested methods over the Adult dataset (Figure 3.5 (c)). This fact shows that the flexibility in choice of classifier(s) is really important to achieve the best results and it makes our solutions widely applicable to different domains and datasets. We can simply use the best performing classifier (single or an ensemble of multiple classifiers) on any given dataset. In general, it is seen that the classifier(s) that produces the highest accuracy at $\theta = 0.5$ for a given dataset also gives low discrimination scores by maintaining the high accuracy, making the choice of classifier(s) easier for decision makers.

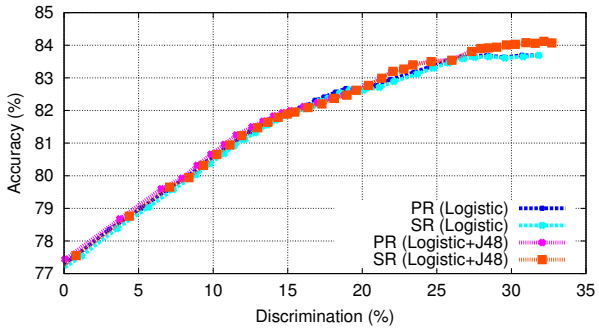
We observe in Figure 3.5 that both PR and SR give comparable performance. However, SR has the advantage that it can be used to establish an evidence of discriminatory practices in the court of law. This advantage of SR makes it a better choice for practitioners.

Results of ER

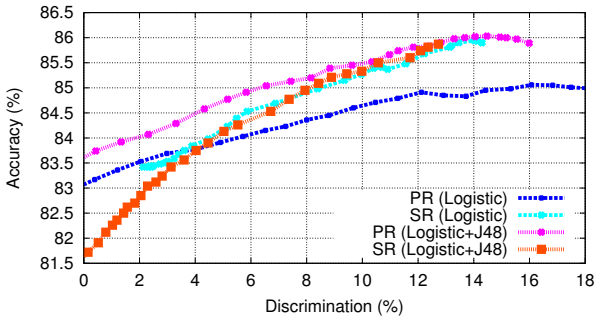
Figure 3.6 shows the results of our experiments with ER over three real world datasets ((a), (b), (c)). In these plots, member classifiers of different ensembles are listed on the lower x-axis, ensemble disagreement is given on the upper x-axis, ER discrimination is shown on left y-axis, and ER accuracy is given on right y-axis. These results demonstrate that discrimination can be controlled by varying the disagreement of the ensemble. For a given dataset, the higher disagreement the ensemble has, the lower is its discrimination with ER. The disagreement of an ensemble, which also measures the diversity of its member classifiers, can be increased by adding more classifiers. Alternatively, the disagreement can be increased by including diverse classifiers in an ensemble. For example, Figure 3.6 (a) shows that it is not always necessary to add more classifiers to reduce discrimination to 0%; just selecting an ensemble with high diversity (e.g., an ensemble comprising of naive Bayes (NBS) and nearest neighbor classifier with $k = 7$ neighbors (IBK7) in this case) is enough to ensure discrimination-free



(a) Communities and Crime



(b) Dutch Census of 2001



(c) Adult

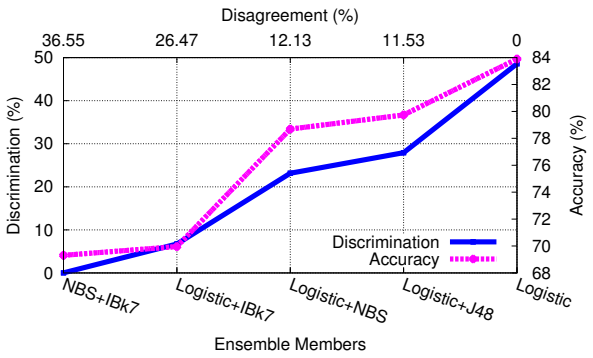
Figure 3.5: Discrimination-accuracy trade-off of PR and SR on three datasets. For each dataset, θ is increased from 0.5 (top right points representing standard decision boundaries) to a maximum value around 0.9 (bottom left points) which reduces the discrimination to 0%.

classification.

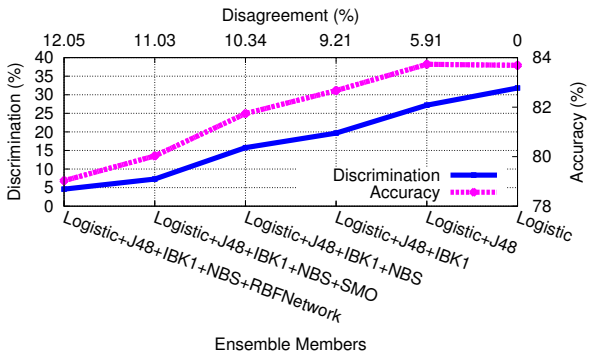
Accuracy and discrimination generally decreases with increase in disagreement. Nonetheless, accuracy remains robust since it is based on agreement of member classifiers of an ensemble. ER has an advantage that it can be used in collaboration with non-probabilistic classifiers; however, its execution time can be higher than that for PR since multiple classifiers need to be learned and applied. Similarly, SR provides a better solution for legal purposes but its execution time is the highest due to the neighborhood search step. The execution times of sample PR, ER, and SR solutions on all datasets are given in Table 3.5. In practice, however, execution time is not a critical deciding factor as real-world predictions do not involve stringent time constraints.

Table 3.5: Average execution time of PR, ER, and SR (in seconds)

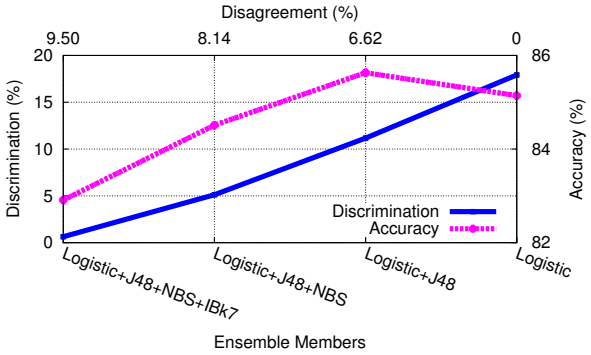
Method↓, Dataset→	Crime	Dutch	Adult
PR (Logistic)	0.58	7.86	14.23
ER (Logistic + J48)	0.76	9.33	18.54
SR (Logistic)	3.2	78	54.55



(a) Communities and Crime



(b) Dutch Census of 2001

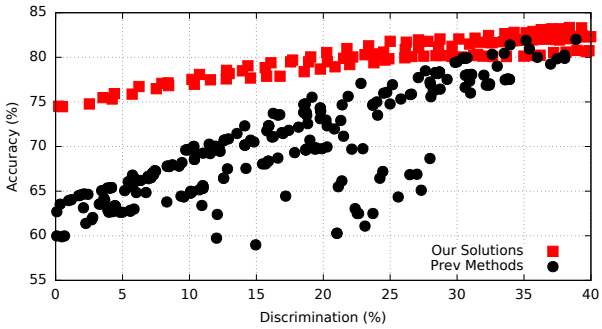


(c) Adult

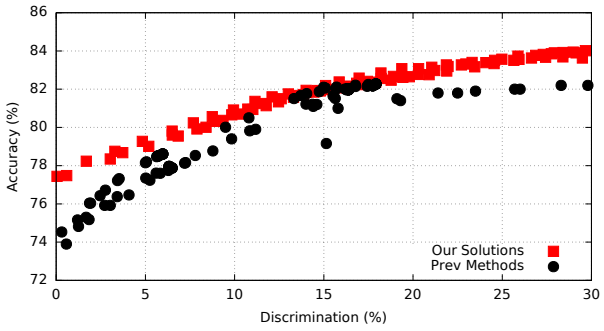
Figure 3.6: Discrimination-accuracy trade-off of ER (disagreement based) on three datasets. For each dataset, several classifier ensembles are shown with their accuracy and discrimination.

Comparison with Previous Methods

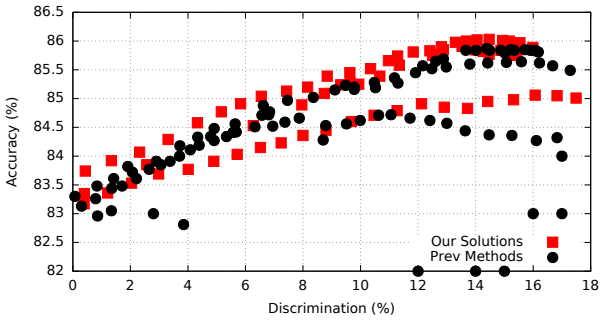
We compare the performance of our solutions (PR, ER, and SR) with that of previous methods of discrimination-aware classification. Figure 3.7 provides a detailed comparison of results on three real-world datasets. It is clear from the figure that our solutions outperform the previously proposed discrimination-aware classification methods of [1–5] w.r.t. accuracy-discrimination trade-off. For each dataset, the accuracy-discrimination curve of our methods lies above all previously reported results, confirming the performance superiority of our solutions. More importantly, our solutions significantly outperform previous methods on the left side of the plots where discrimination is low but accuracy is high. To further discuss the less discriminatory results, we report highest accuracies of our proposed and previous solutions when discrimination is kept only 5%. For *communities and crime* dataset, our solutions find the highest value of accuracy (77%), while the highest accuracy of previous methods is 67% only (Figure 3.7(a)). A similar trend is observed for *Dutch Census of 2001* dataset, where the highest reported accuracy of our solutions is 79.2% and of previous solutions is 78.1 % (Figure 3.7(b)). However, the minimum difference in highest reported accuracies is discovered for the *Adult* dataset, i.e., the previous methods return 84.5% and our solutions return 84.8% (Figure



(a) Communities and Crime



(b) Dutch Census of 2001



(c) Adult

Figure 3.7: Comparison of our solutions with the existing state-of-the-art methods [1–5] on three datasets.

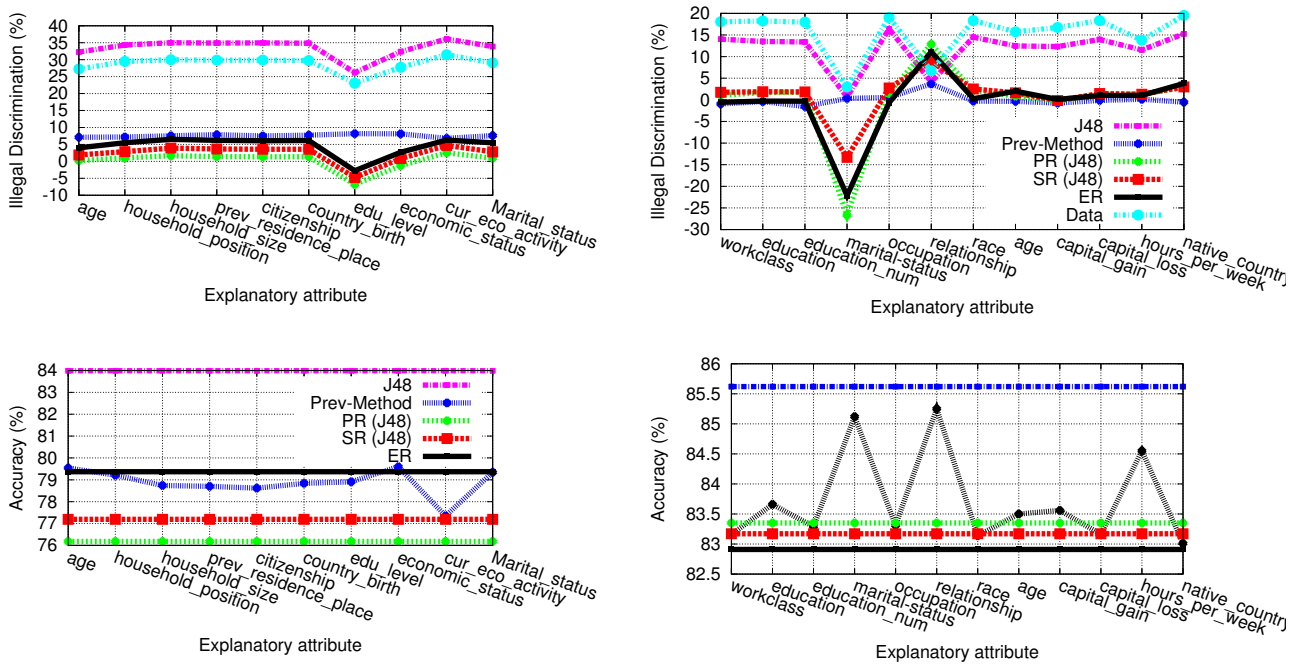
3.7(c)). With the increase in discrimination, the difference in the highest accuracies of our solutions and other state-of-the-arts keep decreasing, which is not justified as eventually discrimination is not prevented. These results, coupled with ease-of-use and flexible control, of our solutions make them a major step forward in practical discrimination-aware classification.

3.3.3 Illegal Discrimination Prevention

In this section, we empirically show that our solutions not only prevent overall discrimination but also ensure illegal discrimination prevention w.r.t. given explanatory attributes. For this purpose we present results of our experiments on two real world datasets: *Adult* and *Dutch Census*. The Communities and Crime dataset is not very appropriate for these experiments because of its small size and all numerical attributes. Although we discretize the numerical attributes in Adult and Dutch Census datasets as well but discretization of numerical attributes in Communities and Crime dataset produces very small data bins that can generate misleading results for overall and illegal discrimination.

The selection of reasonable explanatory attributes is an important step for illegal discrimination calculation and prevention. In the Adult dataset a number of attributes are very weak candidates for being explanatory attributes and thus cannot be presented as an explanation for the low income of females. For instance, we know from biology that race and gender are independent. Thus, race cannot explain the discrimination w.r.t. gender; any such discrimination is either illegal or due to some other

attributes. Similarly, the relationship attribute with values *wife* and *husband* clearly captures the gender information (i.e., is a proxy for gender) and thus cannot be used as an explanation for the low income of females. On the other hand, the attributes age and working hours per week can be considered reasonable for explaining different incomes of males and females. Therefore, it is appropriate to treat them as explanatory attributes. For Dutch Census dataset, attributes education level, age and economic activity are good candidates for explanatory attribute.



(a) Dutch Census of 2001 (b) Adult

Figure 3.8: Performance comparison of our solutions (PR, ER and SR) with the state-of-the-art methods of illegal discrimination prevention.

Selection of explanatory attributes is often difficult and may lead to controversies. Our solutions assume that the explanatory attributes are externally nominated (e.g., by domain experts) and in our experiments we present results by considering each attribute in the dataset as explanatory attribute.

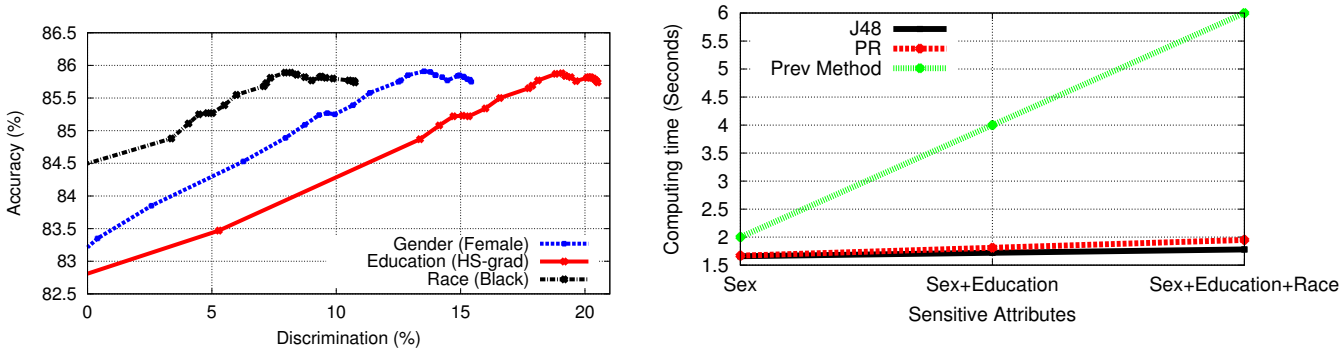
Figure 3.8 shows the performance of our proposed solutions w.r.t. illegal discrimination. In the plots, the x-axis shows different explanatory attributes and the y-axis shows the resultant illegal discrimination (plots on the top) and accuracy (plot in the bottom). Plots on the top of Figure 3.8 present the comparison of illegal discrimination in the actual data (**Data**), in the predictions of a discrimination ignorant classifier, e.g., decision tree in this figure (**J48**), and results of previously proposed methods of [48] (**Prev-Method**) with the illegal discrimination in the predictions of our proposed solutions (PR, ER, SR). We observe that our solutions reduce the illegal discrimination to almost 0% for all reasonable explanatory attributes. In general, our reject option based solutions remove the illegal discrimination with similar magnitude for all explanatory attributes as shown in Figure 3.8. The strange performance observed for the relationship and marital status attributes in the

Adult dataset is due to the fact that these attributes are almost duplicates of the sensitive attribute (gender) and thus are not reasonable explanatory attributes, respectively.

The top two plots of Figure 3.8 also compare the performance of our proposed solutions with the best performing results of [48] where one specialized and independent classifier was learnt for each explanatory attribute separately. It is also very important to mention that our solutions do not require this laborious work of learning a different model for each explanatory attribute. We just learn one model to remove the illegal discrimination w.r.t. all explanatory attributes. We observe that our solutions give comparable performance with the specialized models of [48]. Our solutions are capable of reducing the discrimination to any desired level by changing the value of parameter θ. We observe even the best performing results of previous methods are not able to reduce the illegal discrimination to 0% in the Dutch Census dataset while our solutions reduce the discrimination very close to 0%.

The bottom plots of Figure 3.8 also give the accuracy comparison of our proposed solutions with the best performing and specialized methods of [48]. We observe that our proposed solutions give a comparable accuracy to the previous methods over the Adult dataset. However, in the Dutch Census dataset, PR and SR are a little less accurate as they reduce the illegal discrimination to 0% as compared to the 10% range of specialized methods of [48].

3.3.4 Multiple Sensitive Attributes



(a) Discrimination prevention w.r.t. multiple sensitive attributes (b) Computing time comparison with previous method [48]

Figure 3.9: PR's flexibility to handle discrimination w.r.t. multiple sensitive attributes without training of classification model again.

A key shortcoming of previous methods is the difficulty of handling multiple sensitive attributes which typically requires processing the data or classifier again. On the other hand, our solutions make standard classifier(s) discrimination-aware w.r.t. sensitive attribute(s) at run-time. Thus, our solutions are easy to apply to multiple sensitive attributes or different definitions of deprived groups. We demonstrate this in Figure 3.9(a), which shows the accuracy-discrimination trade-off of PR w.r.t. three sensitive attributes (gender, education, race) on *Adult* dataset. We observe that discrimination decreases

towards zero for all sensitive attributes without repeating the learning procedure by simply increasing the value of θ from 0.5. This flexibility of PR makes it a superior discrimination-aware method as it requires very little computing resources to handle the multiple sensitive attributes as compared to other state-of-the-art methods. Figure 3.9(b) demonstrates this fact by comparing the computing time of PR with a standard decision tree (J48) and a previously proposed discrimination-aware method, i.e., Massaging [1] (Prev Method) on the Adult dataset. We can observe that PR's computing time to handle discrimination w.r.t. multiple sensitive attributes is comparable to the computing time of a standard decision tree. However, the computing time of previous method becomes k times that of a single sensitive attribute when k new sensitive attributes are added, as the method has to re-run the learning process for each sensitive attribute separately. Figure 3.9(b) clearly points out that this drawback of previous discrimination-aware methods would become worse over large datasets.

3.3.5 Performance on Less Discriminatory Test Set

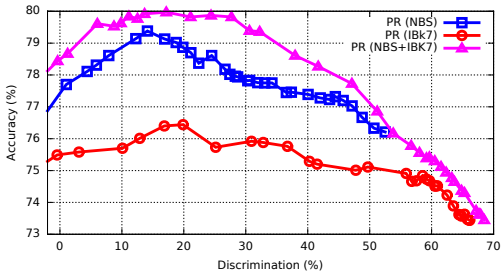


Figure 3.10: Performance of PR on less discriminatory test data.

Ideally, discrimination-aware classification methods trained on discriminatory data should be evaluated on discrimination-free or less discriminatory test sets. However, such evaluation scenarios are not currently available, and in state-of-the-art discrimination-aware classification research, performance is measured via accuracy-discrimination trade-off on discriminatory test sets, as reported in the previous subsections. It is expected that a discrimination-aware classifier that produces high accuracy and low discrimination on discriminatory data will perform with a higher accuracy on less discriminatory test sets. To validate this hypothesis, we construct an experiment in which PR is trained on *Dutch Census of 1971* and tested on *Dutch Census of 2001* datasets. The former dataset has a discrimination of 58.66% while the latter has a discrimination of 24.23%. As discussed while describing the datasets (Section 3.3), the *Dutch Census of 2001* dataset is modified to make it compatible with the *Dutch Census of 1971* dataset for this experiment, and hence, the *Dutch Census of 2001* dataset used in previous subsections is not identical to the one used in this section.

Figure 3.10 shows the performance of PR using single and multiple classifiers when tested on the 2001 version after training on the 1971 version of the *Dutch Census* datasets. Unlike the results reported earlier, where both accuracy and discrimination decreases with an increase in the value of θ, here accuracy actually increases with an increase in θ from 0.5. This trend continues until discrimination is reduced to about 20%, and then accuracy starts decreasing due to the fact that the

test set is not entirely discrimination free. We can expect that accuracy will continue to increase as discrimination reduces to zero if the test set is not entirely discrimination-free. This behavior of PR verifies the hypothesis and confirms its applicability to an ideal scenario where test set is less discriminatory or discrimination-free.

3.3.6 Summary and Discussion

Our experimental evaluations have highlighted several benefits of our proposed solutions for discrimination-aware classification. Table 3.6 summarizes the main advantages, relationships, and differences among the reject option based solutions. We compare our proposed solutions w.r.t. execution time, type of classifiers, and authenticity in the court of law. PR is restricted to single or multiple probabilistic classifiers, while ER and SR can use any type of classifiers. Situational Rejection (SR) is considered highly reliable for justification in the court of law, as it compares the decision of a potentially discriminated/favored instance with its neighbors to establish a case of discrimination or favoritism.

Table 3.6: Main features of proposed methods

Solution↓, Feature→	Non-Prob Classifier	Legal Authenticity	Run Time
PR	No	Medium	Low
ER	Yes	Medium	Medium
SR	Yes	High	High

The most significant benefit of our proposed solutions, specifically PR, is prevention of both overall and illegal discrimination simultaneously. Actually when we increase the value of θ for PR and SR (using PR), it first removes the illegal part of discrimination and further increase of θ removes the rest of the difference in labeling between the sensitive groups to reduce the overall discrimination to zero. This benefit of our solution makes it superior to previously proposed discrimination-aware classification methods as they either reduce illegal discrimination or overall discrimination and not both. Moreover in previous illegal discrimination-aware methods, we have to learn a separate classifier for each explanatory attribute; on the other hand, our reject option based solutions prevent the discrimination w.r.t. all explanatory attributes in a single learning.

Another significant advantage of our solutions is the control over discrimination resulting from the strong correlation between θ (in PR and SR with PR) or disagreement (in ER and SR with ER) and discrimination. This kind of control is not available in the existing discrimination-aware classification methods. We have presented results for different values of θ and disagreement to establish its relationship with discrimination. In practice, if a specific discrimination level is desired, then these parameters can be fixed by using a validation dataset.

3.4 Summary

In this chapter, we present three different solutions for the discrimination-aware classification problem. These easy-to-use and flexible solutions exploit the reject option in classification to identify instances

to label in a manner that reduces discrimination without impacting classification accuracy significantly. The reject option in classification provides a theoretical framework for handling instances close to the decision boundary instances that are more likely to be discriminated. Our solutions employ probabilistic rejection (PR) in probabilistic classifiers, ensemble rejection in classifier ensembles (ER), and PR or ER combined with situation testing (SR). A desirable characteristic of these solutions is their interpretability, i.e., stronger justifications for the decisions as evidence against discriminatory practices in the court of law.

Our experimental evaluations on four real-world datasets confirm the benefits of our solutions and demonstrate our solutions' superior performance when compared to existing state-of-the-art methods. The results also show that our solutions prevent both overall and illegal discrimination simultaneously with minimal loss in accuracy. Stronger justifications, flexibility in practical application, ease-of-use, and overall and illegal discrimination control; these signify a major step forward in practical discrimination-aware classification.

Chapter 4

Future Work

Layered convolutional dictionary learning for sparse coding has been successfully used in different domains, however, has never been employed for the discrete datasets. After using it for interesting itemset mining, we plan to design layered convolutional sparse dictionary learning techniques to tackle sequential, streaming and uncertain discrete data mining problems [17, 86–90]. Discrimination-aware classification is an exciting area of research with many directions for future research. Since decisions impact humans, a broader and less abstract notion of risk needs to be considered in discrimination-aware classifiers: decisions should satisfy safety requirements rather than maximizing accuracy or optimizing accuracy-discrimination trade-off [91]. Furthermore, the learned decision boundary can be quite arbitrary in low density regions thus making the use of distance from decision boundary for risk assessment more uncertain and suggesting greater human oversight in decision making [91]. We believe this direction holds much promise for future research with practical benefits. Another aspect that needs attention in discrimination-aware classification is that of causal inference where the effects of observed and unobserved explainable factors can be controlled in a systematic manner while estimating overall and illegal discrimination (e.g., [92]). In future, we would like to investigate the influence of the critical region on discrimination reduction under different distributions of deprived and favored group instances. Layered convolutional dictionary algorithms for summarizing discriminatory or biased data from financial institutions, hiring agencies, and social service providers can also be designed. This study can yield additional interpretability of deep discrimination-aware classification for decision makers. Detecting and removing illegal or overall discrimination from deep learning based approaches remains an open area for further research.

Bibliography

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

[2] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[3] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *Proc. of IEEE 10th International Conference on Data Mining*, pages 869–874, 2010.

[4] S. Friedler, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *CoRR*, 2014.

[5] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 502–510, 2011.

[6] J. Fowkes and C. Sutton. A bayesian network model for interesting itemsets. In *Proceeding of European Conference Machine Learning and Knowledge Discovery in Databases*, pages 410–425. Springer, 2016.

[7] M. Mampaey, N. Tatti, and J. Vreeken. Tell me what i need to know: succinctly summarizing data with itemsets. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, pages 573–581, 2011.

[8] R. Feldman, Ronen, Sanger, and James. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press, 01 2007.

[9] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. Morgan Kaufmann, 1994.

[10] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.

[11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, May 2000.

[12] G. I. Webb. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. *ACM Transactions on Knowledge and Data Discovery*, 4(1):3:1–3:20, Jan. 2010.

[13] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[14] T. Calders and B. Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1):171–206, 2007.

[15] F. Geerts, B. Goethals, and T. Mielikäinen. *Tiling Databases*, pages 278–289. Springer, Berlin, Heidelberg, 2004.

[16] C. C. Aggarwal and J. Han. *Frequent Pattern Mining*. Springer, 2014.

[17] H. T. Lam, F. Mörchen, D. Fradkin, and T. Calders. Mining compressing sequential patterns. *Statistical Analysis and Data Mining*, 7(1):34–52, Feb. 2014.

[18] F. Mörchen and D. Fradkin. Robust mining of time intervals with semi-interval partial order patterns. Proceedings of Society for Industrial and Applied Mathematics, 2010.

[19] J. Fowkes and C. Sutton. A subsequence interleaving model for sequential pattern mining. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 835–844, USA, 2016.

[20] N. Tatti and J. Vreeken. The long and the short of it: Summarising event sequences with serial episodes. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 462–470, 2012.

[21] J. Vreeken, M. Van Leeuwen, and A. Siebes. Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1):169–214, 2011.

[22] V. V. Vazirani. *Approximation algorithms*. Springer, 2013.

[23] V. Chandola and V. Kumar. Summarization – compressing data into an informative representation. *Knowledge and Information Systems*, 12(3):355–378, 2007.

[24] M. Mampaey, J. Vreeken, and N. Tatti. Summarizing data succinctly with the most informative itemsets. *ACM Transactions on Knowledge Discovery Data*, 6(4):16:1–16:42, Dec. 2012.

[25] K. Smets and J. Vreeken. Slim: Directly mining descriptive patterns. In *Proceedings of SIAM International Conference on Data Mining*, pages 236–247, 2012.

[26] G. I. Webb and J. Vreeken. Efficient discovery of the most interesting associations. *ACM Trans. Knowl. Discov. Data*, 8(3):15:1–15:31, June 2013.

[27] S. Mansha, F. Kamiran, A. Karim, and A. Anwar. A self-organizing map for identifying influentialcommunities in speech-based networks. In *Proceedings of ACM International on Conference on Information and Knowledge Management*, CIKM, pages 1965–1968, 2016.

[28] S. Shang, L. Chen, C. S. Jensen, J.-R. Wen, and P. Kalnis. Searching trajectories by regions of interest. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1549–1562, 2017.

[29] K. Zheng, S. Shang, N. J. Yuan, and Y. Yang. Towards efficient search for activity trajectories. In *Proceedings of IEEE International Conference on Data Engineering*, pages 230–241, April 2013.

[30] S. Shang, R. Ding, B. Yuan, K. Xie, K. Zheng, and P. Kalnis. User oriented trajectory search for trip recommendation. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 156–167. ACM, 2012.

[31] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen. Lcars: A location-content-aware recommender system. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 221–229, 2013.

[32] M. Xie, H. Yin, H. Wang, F. Xu, W. Chen, and S. Wang. Learning graph-based poi embedding for location-based recommendation. In *Proceedings of ACM International on Conference on Information and Knowledge Management*, pages 15–24, 2016.

[33] K. Zheng, H. Su, B. Zheng, S. Shang, J. Xu, J. Liu, and X. Zhou. Interactive top-k spatial keyword queries. In *Proceedings of IEEE International Conference on Data Engineering*, pages 423–434, 2015.

[34] K. Xie, K. Deng, S. Shang, X. Zhou, and K. Zheng. Finding alternative shortest paths in spatial networks. *ACM Trans. Database Syst.*, 37(4):29:1–29:31, 2012.

[35] M. L. Yiu, N. Mamoulis, and D. Papadias. Aggregate nearest neighbor queries in road networks. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):820–833, June 2005.

[36] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang. On discovery of gathering patterns from trajectories. In *Proceedings of IEEE International Conference on Data Engineering*, pages 242–253, 2013.

[37] S. Zhu, Y. Wang, S. Shang, G. Zhao, and J. Wang. Probabilistic routing using multimodal data. *Neurocomputing*, 253(C):49–55, Aug. 2017.

[38] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of International Conference on Machine Learning*, pages 689–696, 2009.

[39] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Proceedings of Advances in Neural Information Processing Systems*, pages 801–808, 2006.

[40] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition)*, pages 2528–2535, 2010.

[41] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun. Learning convolutional feature hierarchies for visual recognition. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1090–1098, 2010.

[42] C. Attorney-General's Dept. Australian sex discimination act 1984., 1984. via: `http://www.comlaw.gov.au/Details/C2010C00056`.

[43] European Union Legislation, 2012. via: `http://europa.eu/legislation_summaries/index_en.htm`.

[44] United Kingdom Legislation, 2012. via: `http://www.legislation.gov.uk/`.

[45] The US Federal Legislation, 2011. via: `http://www.justice.gov/crt`.

[46] European Network Against Racism, 1998. via: `http://www.enar-eu.org/`.

[47] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. In *Proc. of IEEE 12th International Conference on Data Mining Workshops*, pages 370–377, 2012.

[48] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *Proc. of IEEE 11th International Conference on Data Mining*, pages 992–1001, 2011.

[49] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.

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

[51] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data*, 4(2):9, 2010.

[52] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of ACM SIGKDD Conference On Knowledge Discovery And Data Mining*, pages 502–510, 2011.

[53] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté. Rule protection for indirect discrimination prevention in data mining. *Modeling Decision for Artificial Intelligence*, pages 211–222, 2011.

[54] S. Hajian, J. Domingo-Ferrer, and A. Martinez-Balleste. Discrimination prevention in data mining for intrusion and crime detection. In *Proc. of Symposium on Computational Intelligence in Cyber Security*, pages 47–54, 2011.

[55] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, 2013.

[56] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *Proc. of IEEE 11th International Conference on Data Mining Workshops*, page 643â"650, 2011.

[57] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proc. of 30th International Conference on Machine Learning*, pages 325–333, 2013.

[58] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. The independence of fairness-aware classifiers. In *Proc. of IEEE 13th International Conference on Data Mining Workshops*, pages 849–858, 2013.

[59] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: A mechanism for fair classification. 2015.

[60] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. 2015.

[61] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, pages 1–57.

[62] B. Custers, T. Calders, T. Zarsky, and B. Schermer. In *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 341–357. Springer Berlin Heidelberg, 2013.

[63] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Proc. of IEEE 9th International Conference on Data Mining Workshops*, pages 13–18, 2009.

[64] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. of the 19th Ann. Machine Learning Conf. of Belgium and the Netherlands*, pages 1–6, 2010.

[65] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *Proc. of IEEE 13th International Conference on Data Mining*, pages 71–80, 2013.

[66] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proc. of the 3rd ACM Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[67] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *Proc. of IEEE 12th International Conference on Data Mining*, 2012.

[68] Y. LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 2015.

[69] Y.-l. Boureau, Y. L. Cun, et al. Sparse feature learning for deep belief networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1185–1192, 2008.

[70] F. Coenen. The lucs-kdd discretised/normalised arm and carm data library. *URL: http://www.csc.liv.ac.uk/ frans/KDD/Software/ LUCS-KDD-DN/DataSets/dataSets.html*.

[71] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[72] E. W. Weisstein. Chi-squared test. *From MathWorld–A Wolfram Web Resource*, 1999.

[73] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[74] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[75] ECJ. The european court of justice ruling., 2011. via: `http://ec.europa.eu/ireland/press_office/news_of_the_day/ecj-ruling-sex-discrimination-in-insurance-contracts_en.htm`.

[76] M. Hart. Subjective decisionmaking and unconscious discrimination. *Alabama Law Review*, 56:741, 2005.

[77] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. Technical report, Eindhoven University of Technology, 2011.

[78] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[79] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.

[80] K. R. Varshney, R. J. Prenger, T. L. Marlatt, B. Y. Chen, and W. G. Hanley. Practical ensemble classification error bounds for different operating points. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2590–2601, 2013.

[81] M. Bendick. Situation testing for employment discrimination in the united states of america. *Horizons stratégiques*, (3):17–39, 2007.

[82] I. Rorive. Proving discrimination cases: The role of situation testing. 2009.

[83] A. W. David, D. Kibler, and K. M. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[84] A. Asuncion and D. J. Newman. UCI machine learning repository. Online http://archive.ics.uci.edu/ml/, 2007.

[85] Dutch Central Bureau for Statistics. Volkstelling, 2001.

[86] W. Wang, H. Yin, S. Sadiq, L. Chen, M. Xie, and X. Zhou. Spore: A sequential personalized spatial item recommender system. In *Proceedings of International Conference on Data Engineering*, pages 954–965, May 2016.

[87] S. Mansha, Z. Babar, F. Kamiran, and A. Karim. Neural network based association rule mining from uncertain data. In *Proceedings of Neural Information Processing*, pages 129–136. Springer, 2016.

[88] B. Yang, C. Guo, C. S. Jensen, M. Kaul, and S. Shang. Stochastic skyline route planning under time-varying uncertainty. In *Proceedings of IEEE International Conference on Data Engineering*, pages 136–147, March 2014.

[89] A. Zhang, W. Shi, and G. I. Webb. Mining significant association rules from uncertain data. *Data Mining and Knowledge Discovery*, 30(4):928–963, 2016.

[90] Q. Xie, S. Shang, B. Yuan, C. Pang, and X. Zhang. Local correlation detection with linearity enhancement in streaming data. In *Proceedings of ACM International on Conference on Information and Knowledge Management*, CIKM, pages 309–318, 2013.

[91] K. R. Varshney. Engineering safety in machine learning. *arXiv preprint arXiv:1601.04126*, 2016.

[92] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. *arXiv preprint arXiv:1605.03661*, 2016.