2012 IEEE 12th International Conference on Data Mining

Decision Theory for Discrimination-aware Classification

Faisal Kamiran*, Asim Karim†, and Xiangliang Zhang*
*King Abdullah University of Science and Technology (KAUST), The Kingdom of Saudi Arabia
Email: faisal.kamiran, xiangliang.zhang@kaust.edu.sa
†Lahore University of Management Sciences, Pakistan
Email: akarim@lums.edu.pk

***Abstract*—Social discrimination (e.g., against females) arising from data mining techniques is a growing concern worldwide. In recent years, several methods have been proposed for making classifiers learned over discriminatory data discrimination-aware. However, these methods suffer from two major shortcomings: (1) They require either modifying the discriminatory data or tweaking a specific classification algorithm and (2) They are not flexible w.r.t. discrimination control and multiple sensitive attribute handling. In this paper, we present two solutions for discrimination-aware classification that neither require data modification nor classifier tweaking. Our first and second solutions exploit, respectively, the reject option of probabilistic classifier(s) and the disagreement region of general classifier ensembles to reduce discrimination. We relate both solutions with decision theory for better understanding of the process. Our experiments using real-world datasets demonstrate that our solutions outperform existing state-of-the-art methods, especially at low discrimination which is a significant advantage. The superior performance coupled with flexible control over discrimination and easy applicability to multiple sensitive attributes makes our solutions an important step forward in practical discrimination-aware classification.**

I. Introduction

Social discrimination is said to occur when a decision in favor of or against a person or thing is made based on the group, class, or category to which that person or thing belongs to rather than on merit.The discrimination-aware classification problem studies the construction and use of classifiers learned from discriminatory or biased data. The do-nothing approach of simply using a classifier built 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. This do-nothing approach can lead to legal violations and penalties as well.

Although several methods have been proposed in recent years for discrimination-aware classification, they have two key shortcomings. First, they require that either the discriminatory data is 'cleansed' of discriminatory patterns before learning a classifier or a specific classifier's learning algorithm is modified to make it discrimination-aware. Second, they do not provide flexible control over discrimination. A direct consequence of the first shortcoming is that whenever discrimination w.r.t. a different sensitive attribute (or set of attributes) needs to be addressed, the historical data or classifier needs to be processed again. Being restricted to a specific discrimination-aware classifier (e.g., naive Bayes [1], decision tree [2]) is also an issue because that classifier may not be the best performing classifier for a given dataset.

In this paper, we propose two flexible and easy-to-use solutions for discrimination-aware classification based on an intuitive hypothesis: discriminatory decisions are often made close to the decision boundary because of decision maker's bias. We implement this hypothesis via decision theoretic concepts of prediction confidence and ensemble disagreement. Our first solution, called Reject Option based Classification (ROC), exploits the low confidence region of a single or an ensemble of probabilistic classifiers for discrimination reduction. More specifically, ROC invokes the reject option and labels instances belonging to deprived and favored groups in a manner that reduces discrimination. Our second solution, called Discrimination-Aware Ensemble (DAE), exploits the disagreement region of a classifier ensemble to relabel deprived and favored group instances for reduced discrimination. Our proposed solutions have following advantages over existing discrimination-aware classification methods:

1) Our solutions are not restricted to a particular classifier: our first solution works with any probabilistic classifier, while our second solution works with general classifier ensembles.
2) Our solutions require neither modification of learning algorithm nor preprocessing of historical data – pre-trained classifiers can be made discrimination-aware at *prediction time*. Thus, the change in the sensitive attribute can be handled easily by decision makers.
3) Our solutions give better control and interpretability of discrimination-aware classification to decision makers.

We perform extensive experimental evaluation of our solutions on real-world datasets. The results demonstrate better control of discrimination and superior accuracy-discrimination trade-off, when compared to existing sate-of-the-art discrimination-aware classification methods.

II. Related Work

The topic of social discrimination-aware data mining was introduced by Pedreschi et al. [3], [4], focusing on discovery of discriminatory classification rules from biased datasets

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.45

following a frequent itemset mining approach coupled with a measure of discrimination. Proposed methods for discrimination prevention are either based on data preprocessing or algorithm tweaking. Data preprocessing methods modify the biased data to remove discriminatory patterns from it before learning a prediction model from it. In [5], data transformations is proposed for making discovered discriminatory classification rules discrimination-free according to a discrimination measure. The key limitation of the methods of [5] is their applicability to the rule based classifiers only which may not be the best classifier for a given problem. In [6], [7], data sampling and massaging techniques are presented for removing discrimination w.r.t. a single sensitive attribute. Although these techniques can support the learning of any classifier, they are restricted to a single sensitive attribute at a time.

Proposed methods for discrimination prevention using model adaptation include the tweaking of decision trees [2], naive Bayes classifiers [1], and logistic regression [8]. All these methods require that the learning model or algorithm is tweaked, and the first two methods are specific to their respective classifiers. For example in [2], the authors propose a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free.

III. Preliminaries

We consider a two-class problem with label $C \in \{C^+, C^-\}$ defined over instances $X \in \mathcal{X}$ described by a fixed number of attributes. A discriminatory dataset $\mathcal{D} = \{X_i, C_i\}_{i=1}^N$ is available in which the labels C_i may be biased w.r.t. sensitive or discriminatory attributes, e.g., Gender or Race. We assume that C^+ is the desirable label. The instances in $\mathcal{X}$ can be distinguished between those belonging to a given deprived group $\mathcal{X}^d$ and favored group $\mathcal{X}^f$, where $\mathcal{X}^d \cap \mathcal{X}^f = \emptyset$ and $\mathcal{X}^f = \mathcal{X} \setminus \mathcal{X}^d$. All instances in the deprived group $\mathcal{X}^d$ have specific value(s) assigned to specific attribute(s). These attributes are called the sensitive or discriminatory attributes (e.g., Gender, Race) of the problem, which together with their values (e.g., {Gender=Female, Race=Black}), define the deprived group of instances $\mathcal{X}^d$.

The task is to learn a classifier $\mathcal{F} : \mathcal{X} \rightarrow \{C^+, C^-\}$ over the given discriminatory data $\mathcal{D}$ that does not make discriminatory decisions based on the sensitive attribute(s) due to legal constraints. 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 minimally as discrimination is reduced to zero.

We use the definition of [1], [2], [6] to measure discrimination where the discrimination is equivalent to $p(C^+|X \in \mathcal{X}^f) - p(C^+|X \in \mathcal{X}^d)$.

IV. Our Solutions: ROC and DAE

We propose two solutions for the discrimination-aware classification problem. These solutions are obtained by relating decision theoretic concepts with the discrimination model (proposed in [7]) for better interpretation and flexible control of decisions. Our first solution, called Reject Option based Classification (ROC), utilizes posterior probabilities produced by one or more probabilistic classifiers to identify instances for labeling in a manner that neutralizes the effect of discrimination. Our second solution, called Discrimination-Aware Ensemble (DAE), utilizes an ensemble of classifiers to identify instances on which it disagrees for labeling in a manner that reduces discrimination. Both solutions provide an excellent control over the accuracy-discrimination trade-off for the future classifications.

A. Reject Option based Classification (ROC)

Traditionally, a learned classifier assigns an instance to the class with the highest posterior probability. Our first solution 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 ROC for single and multiple classifiers and then relate it with decision theory for interpretation and control.

1) Single Classifier: 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. However, when $p(C^+|X)$ is closer to 0.5 then the label for instance X is more uncertain. A *reject option* can be adopted in classification whereby all instances for which $\max[p(C^+|X), 1 - p(C^+|X)] \leq \theta$ $(where\ 0.5 < \theta < 1)$ are not assigned labels (or are labeled as 'reject'). We refer to this region as *critical region*. The instances in the critical region (rejected instances) are considered to be ambiguous and influenced by biases.

To reduce discrimination, these rejected instances are labeled as follows: if the instance is an example of a deprived group ($\mathcal{X}^d$) 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.

2) Multiple Classifiers: Classifier ensembles are known to be more robust 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 $\mathcal{F}_k$ $(k = 1, \ldots, K)$ denote the kth classifier in an ensemble of $K > 1$ classifiers, and $p(C, \mathcal{F}_k|X)$ be the posterior probability of classification of instance X produced by classifier $\mathcal{F}_k$. The posterior probability of classification

of the ensemble $p(C|X) = \sum_{k=1}^{K} p(C|X, \mathcal{F}_k)p(\mathcal{F}_k)$. The prior probability of a classifier, $p(\mathcal{F}_k)$, can be taken to be proportional to the accuracy of classifier $\mathcal{F}_k$ 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 $p(C|X)$ (as given in the previous paragraph), ROC proceeds in the manner as discussed for a single classifier in the previous subsection. ROC algorithm is shown in Algorithm 1.

Algorithm 1: Reject Option based Classification (ROC)

Input: $\{\mathcal{F}_k\}_{k=1}^{K}$ ($K \geq 1$ probabilistic classifiers trained on $\mathcal{D}$), $\mathcal{X}$ (test set), $\mathcal{X}^d$ (deprived group), θ
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathcal{X}$)
** Critical region **
$\forall X_i \in \{Z|Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] < \theta\}$
  If $X \in \mathcal{X}^d$ **then** $C_i = C^+$
  If $X \notin \mathcal{X}^d$ **then** $C_i = C^-$
** Standard decision rule **
$\forall X_i \in \{Z|Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] \geq \theta\}$
  $C_i = \text{argmax}_{\{C^+, C^-\}}[p(C^+|X_i), p(C^-|X_i)]$

3) Interpreting and Controlling ROC: In this section, we develop a decision theoretic understanding of ROC. Given the posterior probability $p(C^+|X)$ produced by a single or ensemble of probabilistic classifiers, the best label for instance X, that minimizes the expected loss of classification, is given by the $j \in \{+, -\}$ that minimizes:

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

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 I). If all classification errors incur a constant loss (e.g., $L_{+,-} = L_{-,+}$ and $L_{+,+} = L_{-,-} = 0$), then the above decision rule, which is the standard decision rule, ensures the lowest loss in the accuracy of classification.

The trade-off between accuracy and discrimination is controlled by θ; in general the resultant discrimination decreases by increasing the value of θ, as more deprived and favored group instances are likely to be labeled with C^+ and C^-, respectively. Note that, for any given value of θ, the expected reduction in accuracy is the minimum possible because only instances with small posterior probabilities (close to decision boundary) might be potentially misclassified in the relabeling process. To achieve a specified discrimination level, the value of θ can be determined by using a validation dataset or input by the domain expert.

Table I
LOSS MATRIX

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

Table II
ROC LOSS MATRICES.

	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

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 I), and conveniently this loss can be taken to be 1. 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 is.

The entire rejection and relabeling procedure of ROC can also be modeled via loss matrices. Consider a separate 2×2 (no C^r label) loss matrix for deprived and favored group instances (Table II). Our 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 I).

Thus, ROC 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.

B. Discrimination-Aware Ensemble (DAE)

Reject option based classification fits well to probabilistic classifiers. However, not all classification models produce probability estimates, and probabilistic classifiers may not perform well over some given dataset. Our second solution is not restricted to probabilistic classifiers only. It makes an ensemble of (probabilistic, non-probabilistic, or mixed) classifiers discrimination-aware by exploiting the disagreement region among the classifiers.

A standard classifier ensemble classifies new instances by assigning the majority class label. Our solution deviates from this standard procedure to neutralize the effect of discrimination. Specifically, if all member classifiers predict the same label, the agreed class label is assigned; otherwise, we compensate the instances belonging to the deprived group by assigning them the C^+ label and penalize the instances belonging to the favored group by giving the C^- label.

Algorithm 2: Discrimination-Aware Ensemble (DAE)

Input: $\{\mathcal{F}_k\}_{k=1}^{K}$ ($K > 1$ classifiers trained on $\mathcal{D}$), $\mathcal{X}$ (test set), $\mathcal{X}^d$ (deprived group)
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathcal{X}$)
** Disagreement **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \exists (j,k)\ \mathcal{F}_j(Z) \neq \mathcal{F}_k(Z)\}$
  If $X \in \mathcal{X}^d$ **then** $C_i = C^+$
  If $X \notin \mathcal{X}^d$ **then** $C_i = C^-$
** Agreement **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \forall (j,k)\ \mathcal{F}_j(Z) = \mathcal{F}_k(Z)\}$
  $C_i = \text{argmax}_{\{C^+, C^-\}}[p(C^+|X_i), p(C^-|X_i)]$

This strategy is based on the discrimination model of [7] that discrimination impacts instances close to the decision boundary. We use the same intuition in this solution that member classifiers disagree more on the instances that are close to the decision boundary. In other words, disagreement allows us to identify instances that may be misclassified due to discrimination. We can draw a parallel between an ensemble and an admission committee: let us assume that some members of the committee are biased against female applicants and try to reject their applications, 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 may neutralize the discriminatory effect of ensemble due to the fair classifiers. Thus, using ensembles is very useful by nature towards the solution of discrimination-aware classification problem.

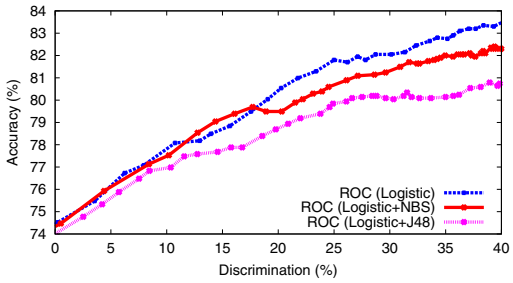
Selecting and Controlling a DAE: We next discuss an important question that which DAE should we choose and how does it impact the discrimination? This is an important question for practitioners of discrimination-aware classification. It is possible that one DAE performs very well w.r.t. accuracy but also produces high discrimination. In this section, we develop an understanding of DAE's performance. We start by defining a measure of DAE disagreement:

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

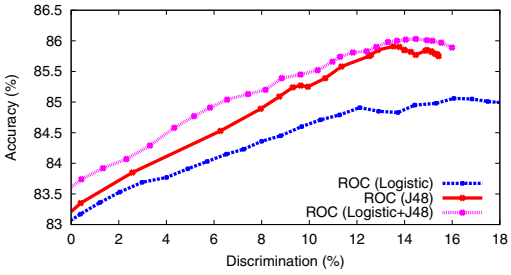
$$disagr_{\mathcal{D}} = \frac{|\{X_i | \exists (j,k)\ \mathcal{F}_j(X_i) \neq \mathcal{F}_k(X_i)\}|}{|\{X_i\}|}$$

In general, lager disagreement of a DAE leads to lower discrimination, because the DAE 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 as a measure of ensemble diversity as well. Ensemble diversity has been shown to be positively correlated with ensemble accuracy determined via majority vote [9]. Although we do not follow majority vote strategy in DAE, classification accuracy of DAE is preserved as only ambiguous (disagreed upon) instances are relabeled.

The trade-off between accuracy and discrimination will depend upon both disagreement and the number of instances that are incorrectly classified. As a DAE with more diverse classifiers tends to have larger disagreement, it will cause less discrimination. Therefore, the discrimination of a DAE can be controlled by changing the diversity of its member classifiers. To select a DAE with a specific discrimination level, a validation dataset can be used. The Discrimination-Aware Ensemble (DAE) algorithm is given in Algorithm 2.



(a) Communities and Crimes



(b) Adult

Figure 1. Discrimination-accuracy trade-off of ROC on two datasets. For each dataset, θ is increased from 0.5 (top right points representing standard decision boundaries) to 0.95 (bottom left points).

V. Experimental Evaluation

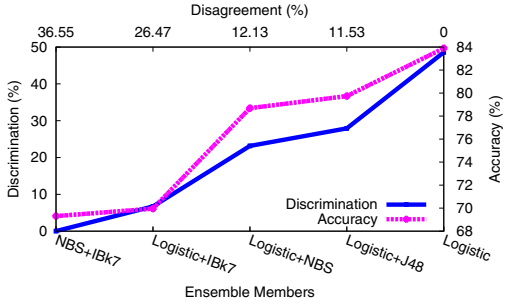
We present and discuss results of the following experiments: (1) Reject Option based Classification (ROC) using single and multiple probabilistic classifiers, identified as **ROC (classifier)** and **ROC (1st classifier+2nd classifier+...)**, respectively. (2) Discrimination-Aware Ensemble (DAE) of two or more classifiers, identified as **DAE (1st classifier+2nd classifier+...)**. (3) Comparison of our solutions' results with those of current state-of-the-art discrimination-aware classification methods, identified

as **Prev Methods**; (4) Evaluation of ROC and DAE w.r.t. different and multiple sensitive attributes.

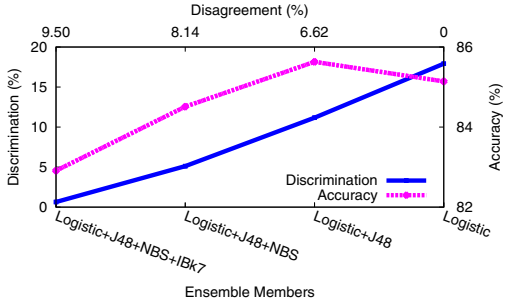
Since our solutions are not restricted to any specific classifier, we consider several standard classifiers for discrimination-aware classification (identifying label of each classifier is given in parenthesis): naive Bayes (**NBS**), logistic regression (**Logistic**), k-nearest neighbor (**IBK**), and decision tree (**J48**).

Datasets: We conduct our experiments on two real-world datasets: Adult, Communities and Crimes [10]. The Adult dataset has 16,281 instances of demographic information of people. Each instance is described by 8 categorical and 6 numerical attributes. We use the income attribute as the class attribute. We consider sex to be the sensitive attribute and sex=female as the deprived group (X^d). The Communities and Crimes dataset contains information about the criminal involvement of 1,994 individuals in the United States. Each individual is described by 122 attributes which are used to predict the total number of violent crimes per 100K population. We consider black individuals to form deprived group, and define this group as all individuals with the numerical attribute racepctblack > 0.06. To define the desirable and undesirable classes we discretize the prediction attribute into minor and major violent communities.

All results reported in the paper are obtained using *10-fold cross-validation* and each point in the figures represents the result of an independent experiment.



(a) Communities and Crimes



(b) Adult

Figure 2. Discrimination-accuracy trade-off of DAE on two datasets. For each dataset, several classifier ensembles are shown with their accuracy and discrimination.

A. Results of ROC

Figure 1 shows the results of our experiments with ROC on two datasets ((a), (b)). The X- and Y-axis of these plots represents classifier(s) discrimination and accuracy, respectively, and each point is for a specific value of θ which is varied from 0.5 to 0.95. It is observed that as the value of θ is increased from 0.5 to 0.95, the discrimination usually reduces to zero around $\theta = 0.9$. Furthermore, the reduction in discrimination with increase in θ is generally smooth and consistent across datasets and classifier(s). Thus, ROC's discrimination level can be controlled easily by varying the value of θ. The minimum drop in accuracy for any given value of θ makes ROC a robust solution for practical discrimination-aware classification.

Figure 1 shows the results for selected single and multiple classifiers. The flexibility in choice of classifier(s) makes ROC widely applicable to different domains and datasets. In general, the classifier(s) that produces the highest accuracy at $\theta = 0.5$ for a given dataset also has a good accuracy-discrimination trade-off curve, making the choice of classifier(s) easier for decision makers.

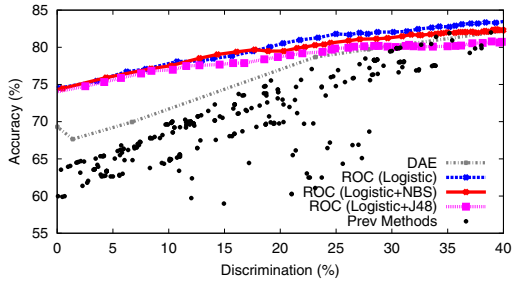
B. Results of DAE

Figure 2 shows the results of our experiments with DAE over two real world datasets ((a), (b)). In these plots, member classifiers of different DAEs are listed on the lower X-axis, DAE disagreement is given on the upper X-axis, DAE discrimination is shown on left y-axis, and DAE accuracy is given on right Y-axis. These results demonstrate that discrimination can be controlled by varying the disagreement of the DAE. For all datasets, higher disagreement results in lower discrimination. The disagreement of a DAE, 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 a DAE.

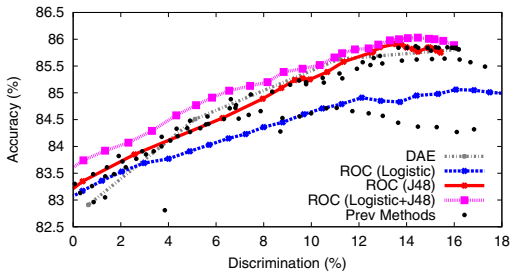
Accuracy generally decreases with increase in disagreement and reduction in discrimination. Nonetheless, accuracy remains robust since it is based on the agreement of member classifiers of an ensemble.

C. Comparison with Previous Methods

We compare the performance of ROC and DAE with that of previous methods of discrimination-aware classification. Figure 3 provides a detailed comparison of results on the two real-word datasets ((a), (b)). It is clear from this figure that our solutions outperform the previously proposed discrimination-aware classification methods of [1], [2], [6] w.r.t. accuracy discrimination trade-off. For each dataset, the accuracy-discrimination curve of a ROC or DAE lies above all previously reported results, confirming the performance superiority of our solutions. More importantly, our solutions significantly outperform previous methods on the left end of the plots where discrimination is low but accuracy is high.



(a) Communities and Crimes



(b) Adult

Figure 3. Comparison of ROC and DAE with existing state-of-the-art methods on three datasets.

These results, coupled with ease-of-use and flexible control, of our solutions make them a major step forward in practical discrimination-aware classification.

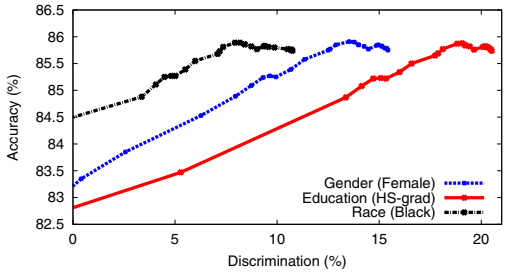


Figure 4. Performance of ROC with different sensitive attributes on the Adult dataset.

D. Multiple Sensitive Attributes

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 4, which shows the accuracy-discrimination trade-off of ROC 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.

VI. Conclusion

In this paper, we present and evaluate two new solutions for the discrimination-aware classification problem. These easy-to-use and flexible solutions utilize decision theory to make standard probabilistic classifiers (ROC) and classifier ensembles (DAE) discrimination-aware. Both ROC and DAE ensure discrimination-aware classifications at run-time without data modification or algorithm tweaking. Moreover, both solutions provide the decision maker with easy control over the resulting discrimination. ROC can also be interpreted as a cost-based classification method in which the cost of misclassifying a deprived group instance as negative is much higher than that of misclassifying a favored group instance as negative. Our experimental evaluations on two real-word datasets confirm the advances of our solutions and their superior performance when compared to existing state-of-the-art methods. As such, our solutions appear to be a major step forward in practical discrimination-aware classification.

In future, we plan to further investigate the critical region, e.g., instead of applying a uniform strategy to all rejected instances we can handle the rejected instances w.r.t. their individual characteristics.

References

[1] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *DMKD*, vol. 21, no. 2, pp. 277–292, 2010.

[2] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *ICDM*, 2010, pp. 869–874.

[3] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *KDD*, 2008.

[4] B. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *KDD*, 2011, pp. 502–510.

[5] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *TKDE*, vol. accepted, 2012.

[6] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *KAIS*, pp. 1–33, 2012.

[7] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *ICDM)*, 2011, pp. 992–1001.

[8] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *ICDMW)*, 2011.

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

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