2011 11th IEEE International Conference on Data Mining Workshops

# Fairness-aware Learning through Regularization Approach

Toshihiro Kamishima*,Shotaro Akaho*, and Jun Sakuma†
*National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Central 2, Umezono 1–1–1, Tsukuba, Ibaraki, 305–8568 Japan,
Email: mail@kamishima.net (http://www.kamishima.net/) and s.akaho@aist.go.jp
†University of Tsukuba; and Japan Science and Technology Agency,
1-1-1 Tennodai, Tsukuba, Japan; and 4-1-8, Honcho, Kawaguchi, Saitama, 332-0012 Japan
Email: jun@cs.tsukuba.ac.jp

***Abstract*—With the spread of data mining technologies and the accumulation of social data, such technologies and data are being used for determinations that seriously affect people's lives. For example, credit scoring is frequently determined based on the records of past credit data together with statistical prediction techniques. Needless to say, such determinations must be socially and legally fair from a viewpoint of social responsibility; namely, it must be unbiased and nondiscriminatory in sensitive features, such as race, gender, religion, and so on. Several researchers have recently begun to attempt the development of analysis techniques that are aware of social fairness or discrimination. They have shown that simply avoiding the use of sensitive features is insufficient for eliminating biases in determinations, due to the indirect influence of sensitive information. From a privacy-preserving viewpoint, this can be interpreted as hiding sensitive information when classification results are observed. In this paper, we first discuss three causes of unfairness in machine learning. We then propose a regularization approach that is applicable to any prediction algorithm with probabilistic discriminative models. We further apply this approach to logistic regression and empirically show its effectiveness and efficiency.**

***Keywords*-fairness, discrimination, privacy, classification, logistic regression, information theory**

## I. Introduction

Data mining techniques are being increasingly used for serious determinations such as credit, insurance rates, employment applications, and so on. Their emergence has been made possible by the accumulation of vast stores of digitized personal data, such as demographic information, financial transactions, communication logs, tax payments, and so on. Additionally, the spread of off-the-shelf mining tools have made it easier to analyze these stored data. Such determinations often affect people's lives seriously. For example, credit scoring is frequently determined based on the records of past credit data together with statistical prediction techniques.

Needless to say, such serious determinations must be socially and legally fair from a viewpoint of social responsibility; that is, they must be unbiased and nondiscriminatory in relation to sensitive features such as race, gender, religion, and so on. Blindness to such factors must be ensured in determinations that affect people's lives directly. Thus, sensitive features must be carefully treated in the processes and algorithms for machine learning.

In some cases, some features must be carefully processed for reasons other than avoiding discrimination. One such reason would be contracts between service providers and customers. Consider the case in which personal information about customer demographics is collected to recommend items at an e-commerce site. If the site collects these data under a privacy policy that restricts the use of the data for the purpose of recommendation, personal information must not be used for the selection of customers to be provided personalized discount coupons. In this case, the use of unrestricted data would be problematic. Because purchasing logs are influenced by recommendations based on personal information, careful consideration would be required for the use of such data.

Several researchers have recently begun to attempt the development of analytic techniques that are aware of social fairness or discrimination [1], [2]. They have shown that the simple elimination of sensitive features from calculations is insufficient for avoiding inappropriate determination processes, due to the indirect influence of sensitive information. For example, when determining credit scoring, the feature of race is not used. However, if people of a specific race live in a specific area and address is used as a feature for training a prediction model, the trained model might make unfair determinations even though the race feature is not explicitly used. Such a phenomenon is called a red-lining effect [2] or indirect discrimination [1], and we describe it in detail in section II-A. New analytic techniques have been devised to deal with fairness. For example, Calders and Verwer proposed a naive Bayes that is modified so as to be less discriminatory [2], and Pedreschi et al. discussed discriminatory association rules [1].

In this paper, we formulate causes of unfairness in machine learning, develop widely applicable and efficient techniques to enhance fairness, and evaluate the effectiveness and efficiency of our techniques. First, we discuss the causes of unfairness in machine learning. In previous works, several notions of fairness have been proposed and successfully exploited. Though these works focused on

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.83

resultant unfairness, we consider unfairness in terms of its causes. We describe three types of cause: *prejudice*, *underestimation*, and *negative legacy*. Prejudice involves a statistical dependence between sensitive features and other information; underestimation is the state in which a classifier has not yet converged; and negative legacy refers to the problems of unfair sampling or labeling in the training data. We also propose measures to quantify the degrees of these causes using mutual information and the Hellinger distance.

Second, we then focus on indirect prejudice and develop a technique to reduce it. This technique is implemented as regularizers that restrict the learners' behaviors. This approach can be applied to any prediction algorithm with discriminative probabilistic models, such as logistic regression. In solving classification problems that pay attention to sensitive information, we have to consider the trade-off between the classification accuracy and the degree of resultant fairness. Our method provides a way to control this trade-off by adjusting the regularization parameter. We propose a *prejudice remover* regularizer, which enforces a determination's independence from sensitive information. As we demonstrate, such a regularizer can be built into a logistic regression model.

Finally, we perform experiments to test the effectiveness and efficiency of our methods. We compare our methods with the two-naive-Bayes on a real data set used in a previous study [2]. We evaluate the effectiveness of our approach and examine the balance between prediction accuracy and fairness.

Note that in the previous work, a learning algorithm that is aware of social discrimination is called *discrimination-aware mining*. However, we hereafter use the terms, 'unfairness' / 'unfair', instead of the 'discrimination' / 'discriminatory' for two reasons. First, as described above, these technologies can be used for complying with laws, regulations, or contracts that are irrelevant to discrimination. Second, because the term *discrimination* is frequently used for the meaning of classification in the machine learning literature, using this term becomes highly confusing. Worse yet, in this paper, we target a *discriminative* model, i.e., logistic regression.

We discuss causes of unfairness in section II and propose our methods for enhancing fairness in section III. Our methods are empirically compared with two-naive-Bayes in section IV. Section V shows related work, and section VI summarizes our conclusions.

## II. Fairness in Data Analysis

After introducing an example of the difficulty in fairness-aware learning, we show three causes of unfairness and quantitative measures for the degrees of these causes.

### A. *Illustration of the Difficulties in Fairness-aware Learning*

We here introduce an example from the literature to show the difficulties in fairness-aware learning [2], which is a simple analytical result for the data set described in section IV-B. The researchers performed a classification problem to predict whether the income of an individual would be high or low.

The sensitive feature, S, was gender, which took a value, Male or Female, and the target class, Y, indicated whether his/her income is High or Low. The sensitive feature, $S$, was gender, which took a value, $\mathsf{Male}$ or $\mathsf{Female}$, and the target class, $Y$, indicated whether his/her income is $\mathsf{High}$ or $\mathsf{Low}$. There were some other non-sensitive features, $X$. The ratio of $\mathsf{Female}$ records comprised about 1/3 of the data set; that is, the number of $\mathsf{Female}$ records was much smaller than that of $\mathsf{Male}$ records. Additionally, while about 30% of $\mathsf{Male}$ records were classified into the $\mathsf{High}$ class, only 11% of $\mathsf{Female}$ records were. Therefore, $\mathsf{Female}$–$\mathsf{High}$ records were the minority in this data set.

In this data set, we describe how $\mathsf{Female}$ records tend to be classified into the $\mathsf{Low}$ class unfairly. Calders and Verwer defined a *discrimination score* (hereafter referred to as the Calders-Verwer score (CV score) by subtracting the conditional probability of the positive class given a sensitive value from that given a non-sensitive value. In this example, a CV score is defined as

$$\Pr[Y{=}\mathsf{High}|S{=}\mathsf{Male}] - \Pr[Y{=}\mathsf{High}|S{=}\mathsf{Female}].$$

The CV score calculated directly from the original data is $0.19$. After training a naive Bayes classifier from data involving a sensitive feature, the CV score on the predicted classes increases to about $0.34$. This shows that $\mathsf{Female}$ records are more frequently misclassified to the $\mathsf{Low}$ class than $\mathsf{Male}$ records; and thus, $\mathsf{Female}$–$\mathsf{High}$ individuals are considered to be unfairly treated. This phenomenon is mainly caused by an Occam's razor principle, which is commonly adopted in classifiers. Because infrequent and specific patterns tend to be discarded to generalize observations in data, minority records can be unfairly neglected. Even if the sensitive feature is removed from the training data for a naive Bayes classifier, the resultant CV score is $0.28$, which still shows an unfair treatment for minorities. This is caused by the indirect influence of sensitive features. This event is called by a *red-lining effect* [2], a term that originates from the historical practice of drawing red lines on a map around neighborhoods in which large numbers of minorities are known to dwell. Consequently, simply removing sensitive features is insufficient, and affirmative actions have to be adopted to correct the unfairness in machine learning.

### B. *Three Causes of Unfairness*

In this section, we discuss the social fairness in data analysis. Previous works [1], [2] have focused on unfairness in the resultant determinations. To look more carefully at the problem of fairness in machine learning, we shall examine the underlying causes or sources of unfairness. We suppose that there are at least three possible causes:

*prejudice*, *underestimation*, and *negative legacy*. Note that these are not mutually exclusive, and two or more causes may compositely lead to unfair treatments.

Before presenting these three causes of unfairness, we must introduce several notations. Here, we discuss supervised learning, such as classification and regression, which is aware of unfairness. $Y$ is a target random variable to be predicted based on the instance values of features. The sensitive variable, $S$, and non-sensitive variable, $X$, correspond to sensitive and non-sensitive features, respectively. We further introduce a prediction model $\mathcal{M}[Y|X,S]$, which models a conditional distribution of $Y$ given $X$ and $S$. With this model and a true distribution over $X$ and $S$, $\mathrm{Pr}^*[X,S]$, we define

$$\mathrm{Pr}[Y,X,S] = \mathcal{M}[Y|X,S]\mathrm{Pr}^*[X,S]. \tag{1}$$

Applying marginalization and/or Bayes' rule to this equation, we can calculate other distributions, such as $\mathrm{Pr}[Y,S]$ or $\mathrm{Pr}[Y|X]$. We use $\tilde{\mathrm{Pr}}[\cdot]$ to denote sample distributions. $\hat{\mathrm{Pr}}[Y,X,S]$ is defined by replacing a true distribution in (1) with its corresponding sample distribution:

$$\hat{\mathrm{Pr}}[Y,X,S] = \mathcal{M}[Y|X,S]\tilde{\mathrm{Pr}}[X,S], \tag{2}$$

and induced distributions from $\hat{\mathrm{Pr}}[Y,X,S]$ are denoted by using $\hat{\mathrm{Pr}}[\cdot]$.

*1) Prejudice:* Prejudice means a statistical dependence between a sensitive variable, $S$, and the target variable, $Y$, or a non-sensitive variable, $X$. There are three types of prejudices: direct prejudice, indirect prejudice, and latent prejudice.

The first type is *direct prejudice*, which is the use of a sensitive variable in a prediction model. If a model with a direct prejudice is used in classification, the classification results clearly depend on sensitive features, thereby generating a database containing *direct discrimination* [1]. To remove this type of prejudice, all that we have to do is simply eliminate the sensitive variable from the prediction model. We then show a relation between such this direct prejudice and statistical dependence. After eliminating the sensitive variable, equation (1) can be rewritten as

$$\mathrm{Pr}[Y,X,S] = \mathcal{M}[Y|X]\mathrm{Pr}^*[S|X]\mathrm{Pr}^*[X].$$

This equation states that $S$ and $Y$ are conditionally independent given $X$, i.e., $Y \perp\!\!\!\perp S \mid X$. Hence, we can say that when the condition $Y \not\perp\!\!\!\perp S \mid X$ is not satisfied, the prediction model has a direct prejudice.

The second type is an *indirect prejudice*, which is statistical dependence between a sensitive variable and a target variable. Even if a prediction model lacks a direct prejudice, the model can have an indirect prejudice and can make an unfair determination. We give a simple example. Consider the case that all $Y$, $X$, and $S$ are real scalar variables, and these variables satisfy the equations:

$$Y = X + \varepsilon_Y \quad \text{and} \quad S = X + \varepsilon_S,$$

where $\varepsilon_Y$ and $\varepsilon_S$ are mutually independent random variables. Because $\mathrm{Pr}[Y,X,S]$ is equal to $\mathrm{Pr}[Y|X]\,\mathrm{Pr}[S|X]\,\mathrm{Pr}[X]$, these variables satisfy the condition $Y \perp\!\!\!\perp S \mid X$, but do not satisfy the condition $Y \perp\!\!\!\perp S$. Hence, the adopted prediction model does not have a direct prejudice, but may have an indirect prejudice. If the variances of $\varepsilon_Y$ and $\varepsilon_S$ are small, $Y$ and $S$ become highly correlated. In this case, even if a model does not have a direct prejudice, the determination clearly depends on sensitive information. Such resultant determinations are called indirect discrimination [1] or a red-lining effect [2] as described in section II-A. To remove this indirect prejudice, we must use a prediction model that satisfies the condition $Y \perp\!\!\!\perp S$.

We next show an index to quantify the degree of indirect prejudice, which is straightforwardly defined as the mutual information between $Y$ and $S$. However, because a true distribution in (1) is unknown, we adopt sample distributions in equation (2) over a given sample set, $\mathcal{D}$:

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\mathrm{Pr}}[y,s] \ln \frac{\hat{\mathrm{Pr}}[y,s]}{\hat{\mathrm{Pr}}[s]\hat{\mathrm{Pr}}[s]}. \tag{3}$$

We refer to this index as a (indirect) *prejudice index* (PI for short). For convenience, the application of the normalization technique for mutual information [3] leads to a *normalized prejudice index* (NPI for short):

$$\mathrm{NPI} = \mathrm{PI}/(\sqrt{\mathrm{H}(Y)\mathrm{H}(S)}), \tag{4}$$

where an entropy function $\mathrm{H}(X)$ is defined as $-\sum_{x\in\mathcal{D}} \hat{\mathrm{Pr}}[x] \ln \hat{\mathrm{Pr}}[x]$. The range of this NPI is $[0,1]$.

The third type of prejudice is latent prejudice, which is a statistical dependence between a sensitive variable, $S$, and a non-sensitive variable, $X$. Consider an example that satisfies the equations:

$$Y = X_1 + \varepsilon_Y, \quad X = X_1 + X_2, \quad \text{and} \quad S = X_2 + \varepsilon_S,$$

where $\varepsilon_Y \perp\!\!\!\perp \varepsilon_S$ and $X_1 \perp\!\!\!\perp X_2$. Clearly, the conditions $Y \perp\!\!\!\perp S \mid X$ and $Y \perp\!\!\!\perp S$ are satisfied, but $X$ and $S$ are not mutually independent. This dependence doesn't cause a sensitive information to influence the final determination, but it would be exploited for training learners; thus, this might violate some regulations or laws. Recall our example about personal information in section I. The use of raw purchasing logs may violate contracts with customers, because the logs are influenced by recommendations based on personal information, even if it is irrelevant to the final selection of customers. Removal of potential prejudice is achieved by making $X$ and $Y$ independent from $S$ simultaneously. Similar to a PI, the degree of a latent prejudice can be quantified by the mutual information between $X$ and $S$.

*2) Underestimation:* Underestimation is the state in which a learned model is not fully converged due to the finiteness of the size of a training data set. Given a learning algorithm that can acquire a prediction model without indirect prejudice, it will make a fair determination if infinite training examples are available. However, if the size of the training data set is finite, the learned classifier may lead to more unfair determinations than that observed in the training sample distribution. Though such determinations are not intentional, they might awake suspicions of unfair treatment. In other words, though the notion of convergence at infinity is appropriate in a mathematical sense, it might not be in a social sense. We can quantify the degree of underestimation by assessing the resultant difference between the training sample distribution over $\mathcal{D}$, $\tilde{\Pr}[\cdot]$, and the distribution induced by a model, $\hat{\Pr}[\cdot]$. Along this line, we define the *underestimation index* (UEI) using the Hellinger distance:

$$\begin{aligned}\mathrm{UEI} &= \Big(\frac{1}{2}\sum_{y,s\in\mathcal{D}}\Big(\sqrt{\hat{\Pr}[y,s]}-\sqrt{\tilde{\Pr}[y,s]}\Big)^2\Big)^{1/2}\\ &= \Big(1-\sum_{y,s\in\mathcal{D}}\sqrt{\hat{\Pr}[Y,S]\tilde{\Pr}[Y,S]}\Big)^{1/2}. \qquad (5)\end{aligned}$$

Note that we did not adopt the KL-divergence because it can be infinite and this property is inconvenient for an index.

*3) Negative Legacy:* Negative legacy is unfair sampling or labeling in the training data. For example, if a bank has been refusing credit to minority people without assessing them, the records of minority people are less sampled in a training data set. A sample selection bias is caused by such biased sampling depending on the features of samples. It is known that the problem of a sample selection bias can be avoided by adopting specific types of classification algorithms [4]. However, it is not easy to detect the existence of a sample selection bias only by observing training data. On the other hand, if a bank has been unfairly rejecting the loans of the people who should have been approved, the labels in the training data would become unfair. This problem is serious because it is hard to detect and correct. However, if other information, e.g., a small-sized fairly labeled data set, can be exploited, this problem can be corrected by techniques such as transfer learning [5].

Regulations or laws that demand the removal of potential prejudices are rare. We investigate UEIs in the experimental sections of this paper, but we don't especially focus on underestimation. As described above, avoiding a negative legacy can be difficult if no additional information is available. We therefore focus on the development of a method to remove indirect prejudice.

## III. Prejudice Removal Techniques

We here propose a technique to reduce indirect prejudice. Because this technique is implemented as a regularizer, which we call a prejudice remover, it can be applied to wide variety of prediction algorithms with probabilistic discriminative models.

### A. General Framework

We focused on classification and built our regularizers into logistic regression models. $Y$, $X$, and $S$ are random variables corresponding to a class, non-sensitive features, and a sensitive feature, respectively. A training data set consists of the instances of these random variables, i.e., $\mathcal{D}=\{(y,\mathbf{x},s)\}$. The conditional probability of a class given non-sensitive and sensitive features is modeled by $\mathcal{M}[Y|X,S;\mathbf{\Theta}]$, where $\mathbf{\Theta}$ is the set of model parameters. These parameters are estimated based on the maximum likelihood principle; that is, the parameters are tuned so as to maximize the log-likelihood:

$$\ell(\mathcal{D};\mathbf{\Theta})=\sum_{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}}\ln\mathcal{M}[y_i|\mathbf{x}_i,s_i;\mathbf{\Theta}]. \qquad (6)$$

We adopted two types of regularizers. The first regularizer is a standard one to avoid over-fitting. We used an $L_2$ regularizer $\|\mathbf{\Theta}\|_2^2$. The second regularizer, $R(\mathcal{D},\mathbf{\Theta})$, is introduced to enforce fair classification. We designed this regularizer to be easy to implement and to require only modest computational resources. By adding these two regularizers to equation (6), the objective function to minimize is obtained:

$$-\ell(\mathcal{D};\mathbf{\Theta})+\eta\mathrm{R}(\mathcal{D},\mathbf{\Theta})+\frac{\lambda}{2}\|\mathbf{\Theta}\|_2^2, \qquad (7)$$

where $\lambda$ and $\eta$ are positive regularization parameters.

We dealt with a classification problem in which the target value $Y$ is binary $\{0,1\}$, $X$ takes a real vectors, $\mathbf{x}$, and $S$ takes a discrete value, $s$, in a domain $\mathcal{S}$. We used a logistic regression model as a prediction model:

$$\mathcal{M}[y|\mathbf{x},s;\mathbf{\Theta}]=y\sigma(\mathbf{x}^\top\mathbf{w}_s)+(1-y)(1-\sigma(\mathbf{x}^\top\mathbf{w}_s)), \quad (8)$$

where $\sigma(\cdot)$ is a sigmoid function, and the parameters are weight vectors for $\mathbf{x}$, $\mathbf{\Theta}=\{\mathbf{w}_s\}_{s\in\mathcal{S}}$. Note that a constant term is included in $\mathbf{x}$ without loss of generality. We next introduce a regularizer to reduce the indirect prejudice.

### B. Prejudice Remover

A *prejudice remover* regularizer directly tries to reduce the prejudice index and is denoted by $\mathrm{R}_{\mathsf{PR}}$. Recall that the prejudice index is defined as

$$\begin{aligned}\mathrm{PI} &= \sum_{Y,S}\hat{\Pr}[Y,S]\ln\frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}\\ &= \sum_{Y,X,S}\mathcal{M}[Y|X,S;\mathbf{\Theta}]\tilde{\Pr}[X,S]\ln\frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}.\end{aligned}$$

$\sum_{X,S}\tilde{\Pr}[X,S]$ can be replaced with $\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}}$, and the argument of logarithm can be rewritten as $\hat{\Pr}[Y|s_i]/\hat{\Pr}[Y]$, by reducing $\hat{\Pr}[S]$. We obtain

$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}}\sum_{y\in\{0,1\}}\mathcal{M}[y|\mathbf{x}_i,s_i;\mathbf{\Theta}]\ln\frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}.$$

The straightforward way to compute $\hat{\Pr}[y|s]$ is to marginalize $\mathcal{M}[y|X,s;\mathbf{\Theta}]\hat{\Pr}[X,s]$ over $X$. However, if the domain of $X$ is large, this marginalization is computationally heavy. We hence take a drastically simple approach. We replace $X$ with $\bar{x}_s$, which is a sample mean vector of $\mathbf{x}$ over a set of training samples whose corresponding sensitive feature is equal to $s$, $\{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}\text{ s.t. }s_i=s\}$, and we get

$$\hat{\Pr}[y|s]=\mathcal{M}[y|\bar{\mathbf{x}}_s,s;\mathbf{\Theta}], \tag{9}$$

$$\hat{\Pr}[y]=\sum_{s\in\mathcal{S}}\hat{\Pr}[s]\mathcal{M}[y|\bar{\mathbf{x}}_s,s;\mathbf{\Theta}]. \tag{10}$$

Finally, the prejudice remover regularizer $\mathrm{R}_{\mathsf{PR}}(\mathcal{D},\mathbf{\Theta})$ is

$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}}\sum_{y\in\{0,1\}}\mathcal{M}[y|\mathbf{x}_i,s_i;\mathbf{\Theta}]\ln\frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}, \tag{11}$$

where $\hat{\Pr}[y|s]$ and $\hat{\Pr}[y]$ are equations (9) and (10), respectively. This regularizer becomes large when a class is determined mainly based on sensitive features; thus, sensitive features become less influential to the final determination. In the case of logistic regression, the objective function (7) to minimize is rewritten as

$$-\sum_{(y_i,\mathbf{x}_i,s_i)}\ln\mathcal{M}[y_i|\mathbf{x}_i,s_i;\mathbf{\Theta}]+\mathrm{R}_{\mathsf{PR}}(\mathcal{D},\mathbf{\Theta})+\frac{\lambda}{2}\sum_{s\in\mathcal{S}}\|\mathbf{w}_s\|_2^2, \tag{12}$$

where $\mathcal{M}[y|\mathbf{x},s;\mathbf{\Theta}]$ is equation (8) and $\mathrm{R}_{\mathsf{PR}}(\mathcal{D},\mathbf{\Theta})$ is equation (11). In our experiment, this objective function is minimized by a conjugate gradient method starting from the initial condition $\mathbf{w}_s=\mathbf{0}$, $\forall s\in\mathcal{S}$, and we obtain an optimal parameter set, $\{\mathbf{w}_s^*\}$.

The probability of $Y=1$ given a sample without a class label, $(\mathbf{x}_{\text{new}},s_{\text{new}})$ can be predicted by

$$\Pr[Y{=}1|\mathbf{x}_{\text{new}},s_{\text{new}};\{\mathbf{w}_s^*\}]=\sigma(\mathbf{x}_{\text{new}}^\top\mathbf{w}_{s_{\text{new}}}^*).$$

## IV. Experiments

We compared our method with Calders and Verwer's method on the real data set used in a previous study [2].

### A. Calders-Verwer's 2-Naive-Bayes

We briefly introduce Calders and Verwer's 2-naive-Bayes method (CV2NB), which was found to be the best method in the previous study using the same dataset [2]. This method targets a binary classification problem. The number of sensitive features is one and the feature is binary. The generative model of this method is

$$\Pr[Y,\mathbf{X},S]=\mathcal{M}[Y,S]\prod_i\mathcal{M}[X_i|Y,S]. \tag{13}$$

$\mathcal{M}[X_i|Y,S]$ models a conditional distribution of $X_i$ given $Y$ and $S$, and the parameters of these models are estimated by the similar way in the estimation of parameters of a naive Bayes model. $\mathcal{M}[Y,S]$ models a joint distribution $Y$ and $S$. Because $Y$ and $S$ are not mutually independent, the final determination might be unfair. While each feature depends only on a class in the case of the original naive Bayes, every non-sensitive feature, $X_i$, depends on both $Y$ and $S$ in the case of CV2NB. It is as if two naive Bayes classifiers are learned depending on each value of the sensitive feature; that is why this method was named by the *2-naive-Bayes*. To make the classifier fair, $\mathcal{M}[Y,S]$ is initialized by the sample distribution $\tilde{\Pr}[Y,S]$, and this model is modified by the algorithms in Figure 1. A model parameter $\mathcal{M}(y,s)$ is derived by $N(y,s)/\sum_{Y,S}N(y',s')$. This algorithm is designed so as to update $\Pr[Y,S]$ gradually until a CV score becomes positive. Note that we slightly modified the original algorithm by adding line 10 in Figure 1, which guarantees the parameters, $N(Y,S)$, to be non-negative, because the original algorithm may fail to stop.

```
1  Calculate a CV score, $disc$, of the predicted classes by the current model.
2  while $disc > 0$
3      $numpos$ is the number of positive samples classified by the current model.
4      if $numpos <$ the number of positive samples in $\mathcal{D}$ then
5          $N(Y{=}1,S{=}0) \leftarrow N(Y{=}1,S{=}0) + \Delta N(Y{=}0,S{=}1)$
6          $N(Y{=}0,S{=}0) \leftarrow N(Y{=}0,S{=}0) - \Delta N(Y{=}0,S{=}1)$
7      else
8          $N(Y{=}0,S{=}1) \leftarrow N(Y{=}0,S{=}1) + \Delta N(Y{=}1,S{=}0)$
9          $N(Y{=}1,S{=}1) \leftarrow N(Y{=}1,S{=}1) - \Delta N(Y{=}1,S{=}0)$
10     if any of $N(Y,S)$ is negative then
           cancel the previous update of $N(Y,S)$ and abort
11     Recalculate $\Pr[Y|S]$ and a CV score, $disc$ based on updated $N(Y,S)$
```

Figure 1. naive Bayes modification algorithm

NOTE: $N(Y{=}y,S{=}s)$ denotes the number of samples in $\mathcal{D}$, whose class and sensitive feature are $y$ and $s$, respectively. In our experiment, $\Delta$ was set to 0.01 as in the original paper.

### B. Experimental Conditions

We summarize our experimental conditions. We tested a previously used real data set [2], as shown in section II-A. This set includes 16281 data in an adult.test file of the Adult/Census Income distributed at the UCI Repository [6]. The target variable indicates whether or not income is larger than 50M dollars, and the sensitive feature is gender. Thirteen non-sensitive features were discretized by the procedure in the original paper. In the case of the naive Bayes, parameters of models, $\mathcal{M}[X_i|Y,S]$, are estimated by a MAP estimator with multinomial distribution and Dirichlet priors. In the case of our logistic regression, discrete variables are represented by 0/1 dummy variables coded by a so-called 1-of-$K$ scheme. The regularization parameter for the $L_2$ regularizer, $\lambda$, is fixed to 1, because the performance of pure logistic regression was less affected by this parameter in our preliminary experiments. We tested six methods: logistic regression with a sensitive feature (LR), logistic regression without a sensitive feature (LRns), logistic regression with a prejudice remover regularizer (PR), naive Bayes with a sensitive feature (NB), naive Bayes without a sensitive feature (NBns), and Caldars and Verwer's 2-naive-Bayes (CV2NB). We show the means of the statistics obtained by the five-fold cross-validation.

Table I
A SUMMARY OF EXPERIMENTAL RESULTS

| method | Acc | NMI | NPI | UEI | CVS | PI / MI |
|---|---|---|---|---|---|---|
| LR | 0.851 | 0.267 | 5.21E-02 | 0.040 | 0.189 | 2.10E-01 |
| LRns | 0.850 | 0.266 | 4.99E-04 | 0.036 | -0.033 | 1.06E-03 |
| PR $\eta$=0 | 0.850 | 0.265 | 4.94E-02 | 0.038 | 0.185 | 2.01E-01 |
| PR $\eta$=0.1 | 0.850 | 0.264 | 4.11E-02 | 0.036 | 0.170 | 1.68E-01 |
| PR $\eta$=0.3 | 0.774 | 0.149 | 7.53E-03 | 0.127 | -0.095 | 5.47E-02 |
| PR $\eta$=1 | 0.720 | 0.124 | 1.29E-05 | 0.148 | -0.004 | 1.12E-04 |
| PR $\eta$=10 | 0.676 | 0.013 | 2.13E-01 | 0.259 | -0.472 | 1.84E+01 |
| NB | 0.822 | 0.246 | 1.12E-01 | 0.068 | 0.332 | 4.90E-01 |
| NBns | 0.826 | 0.249 | 7.17E-02 | 0.043 | 0.267 | 3.11E-01 |
| CV2NB | 0.813 | 0.191 | 3.64E-06 | 0.082 | -0.002 | 2.05E-05 |

NOTE: $\langle n_1 \rangle \mathrm{E} \langle n_2 \rangle$ denotes $n_1 \times 10^{n_2}$.

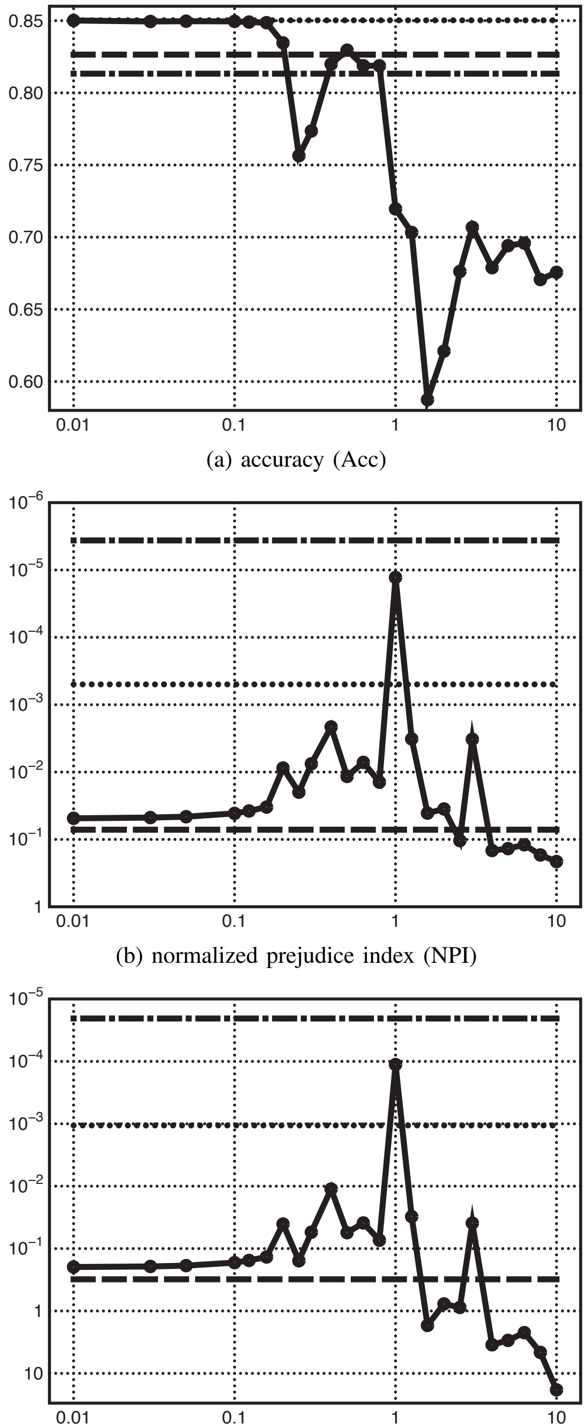
### C. Experimental Results

Table I shows accuracies (Acc), NPI and UEI in section II, and CV scores (CVS). MI denotes mutual information between sample labels and predicted labels, NMI was obtained by normalizing this MI in a process similar to NPI. PI / MI quantifies a prejudice index that is sacrificed by obtaining a unit of information about the correct label. This can be used to measure the efficiency of the trade-off between prediction accuracy and prejudice removal. A smaller PI / MI value indicates higher efficiency in this trade-off.

We first compare the performance of our method with that of baselines in Table I. Compared with NBns, our method was superior both in accuracy and NPI at $\eta = 0.1$. Because LRns successfully removed prejudice without sacrificing accuracy unlike NBns, our PR at $\eta = 1$ was better in PI / MI, but accuracy was fairly degraded. Note that two methods, PR at $\eta = 0$ and LR, behaved similarly, because our PR is almost equivalent to LR if the prejudice remover is eliminated by setting $\eta = 0$.

We next moved on to the influence of the parameter, $\eta$, which controls the degree of prejudice removal. We expected that the larger the $\eta$, the more prejudice would be removed, whereas accuracy might be sacrificed. According to Table I, as $\eta$ increased, our PR generally become degraded in accuracy, but was also not fully improved in prejudice removal.

To further investigate the change of performance depending on this parameter $\eta$, we demonstrated the variations in accuracy (Acc), normalized prejudice index (NPI), and the trade-off efficiency between accuracy and prejudice removal (PI / MI) in Figure 2. We focus on our PR method. Overall, the changes were rather unstable in all statistics. The reasons for this instability would be the sub-optimality in solutions stemming from the lack of convexity of the objective function (12) and the approximation by replacing the marginal values of $X$ with their sample means. The increase of $\eta$ generally damaged accuracy because a prejudice remover regularizer is designed to remove prejudice by sacrificing correctness in prediction. NPI peaked at $\eta = 1$, though



(a) accuracy (Acc)

(b) normalized prejudice index (NPI)

(c) trade-off efficiency between accuracy and prejudice removal (PI / MI)

Figure 2. The change in performance according to the parameter $\eta$

NOTE: Horizontal axes represent the parameter $\eta$, and vertical axes represent statistics in each subtitle. Solid, chain, dotted, and broken lines indicate the statistics of PR, CV2NB, LRns, and NBns, respectively.

we expected that more prejudice would be removed as $\eta$ increased. We postulate that this would be due to the approximation in the marginalization of $X$; further investigation is required for this point. The peak in trade-off efficiency was observed at $\eta = 1$, but accuracy was fairly damaged at this point.

We next compared our PR with other methods. The performance of CV2NB was fairly good, and our PR was inferior except for accuracy at the lower range of $\eta$. When compared to the baseline LRns, by tuning the parameter $\eta$, our PR could exceed in all statistics. However, it failed to exceed in both accuracy and prejudice removal at the same $\eta$.

In summary, our PR could successfully reduce indirect prejudice, but accuracy was sacrificed for this reduction. We must further improve the efficiency in the trade-off between accuracy and prejudice removal.

## V. Related Work

Several analytic techniques that are aware of fairness or discrimination have recently received attention. Pedreschi et al. emphasized the unfairness in association rules whose consequents include serious determinations [1], [7]. They advocated the notion of *$\alpha$-protection*, which is the condition that association rules were fair. Given a rule whose consequent exhibited negative determination, it would be unfair if the confidence of the rule substantially increased by adding a condition related to a sensitive feature to the antecedent part of the rule. The $\alpha$-protection constrains the rule so that the ratio of this increase is at most $\alpha$. They also suggested the notions of *direct discrimination* and *indirect discrimination*. A direct discriminatory rule directly contains a sensitive condition in its antecedent, and while an indirect discriminatory rule doesn't directly contain a sensitive condition, the rule is considered to be unfair in the context of background knowledge that includes sensitive information. Their work has since been extended [8]. Various kinds of indexes for evaluating discriminatory determinations were proposed and their statistical significance has been discussed. A system for finding such unfair rules has been proposed [9]. Calders and Verwer proposed several methods to modify naive Bayes for enhancing fairness as described in section IV-A [2]. Luong et al. proposed a notion of situation testing, wherein a determination is considered unfair if different determinations are made for two individuals all of whose features are equal except for sensitive ones [10]. Such unfairness was detected by comparing the determinations for records whose sensitive features are different, but are neighbors in non-sensitive feature space. If a target determination differs, but non-sensitive features are completely equal, then a target variable depends on a sensitive variable. Therefore, this situation testing has connection to our indirect prejudice. Dwork et al. argued a data transformation for the purpose of exporting data while keeping aware of fairness [11]. A data set held by a data owner is transformed and passed to a vendor who classifies the transformed data. The transformation preserves the neighborhood relations of data and the equivalence between the expectations of data mapped from sensitive individuals and from non-sensitive ones. In a sense that considering the neighborhood relations, this approach is related to the above notion of situation testing. Because their proposition 2.2 implies that the classification results are roughly independent from the membership in a sensitive group, their approach has relation to our idea of prejudice.

In a broad sense, fairness-aware learning is a kind of cost-sensitive learning [12]. That is to say, the cost of enhancing fairness is taken into account. Fairness in machine learning can be interpreted as a sub-notion of legitimacy, which means that models can be deployed in the real world [13]. Gondek and Hofmann devised a method for finding clusters that were not relevant to a given grouping [14]. If a given grouping contains sensitive information, this method can be used for clustering data into fair clusters. Independent component analysis might be used to maintain the independence between features [15].

The removal of prejudice is closely related to privacy-preserving data mining [16], which is a technology for mining useful information without exposing individual private records. The privacy protection level is quantified by mutual information between the public and private realms [17]. In our case, the degree of indirect prejudice is quantified by mutual information between classification results and sensitive features. Due to the similarity of these two uses of mutual information, the design goal of fairness-aware learning can be considered the protection of sensitive information when exposing classification results.

Regarding underestimation, the concepts of anytime algorithms in planning or decision making [18] might be useful.

As described in section II-B, the problem of negative legacy is closely related to transfer learning. Transfer learning is "the problem of retaining and applying the knowledge learned in one or more tasks to efficiently develop an effective hypothesis for a new task" [5]. Among many types of transfer learning, the problem of a sample selection bias [4] would be related to the negative legacy problem. Sample selection bias means that the sampling is not at random, but biased depending on some feature values of data. Another related approach to transfer learning is weighting samples according the degree of usefulness for the target task [19]. Using these approaches, if given a small amount of fairly labeled data, other data sets that might be unfairly labeled would be correctly processed.

## VI. Conclusions and Future Work

The contributions of this paper are as follows. First, we proposed three causes of unfairness: prejudice, underestimation, and negative legacy. Prejudice refers to the dependence between sensitive information and the other information, either directly or indirectly. We further classified prejudice into three types and developed a way to quantify them by mutual information. Underestimation is the state in which a classifier has not yet converged, thereby producing more unfair determinations than those observed in a sample distribution. Negative legacy is the problem of unfair sampling or labeling in the training data. Second, we developed techniques to reduce indirect prejudice. We proposed a prejudice remover regularizer, which enforces a classifier's independence from sensitive information. Our methods can be applied to any algorithms with probabilistic discriminative models and are simple to implement. Third, we showed experimental results of logistic regressions with our prejudice remover regularizer. The experimental results showed the effectiveness and efficiency of our methods. We further proposed a method to evaluate the trade-offs between the prediction accuracy and fairness.

Research on fairness-aware learning is just beginning; thus, there are many problems yet to be solved; for example, the definition of fairness in data analysis, measures for fairness, and maintaining other types of laws or regulations. The types of analytic methods are severely limited at present. Our method can be easily applied to regression, but fairness-aware clustering and ranking methods are also needed.

The use of data mining technologies in our society will only become greater with time. Unfortunately, their results can occasionally damage people's lives [20]. On the other hand, data analysis is crucial for enhancing public welfare. For example, exploiting personal information has proved to be effective for reducing energy consumption, improving the efficiency of traffic control, preventing infectious diseases, and so on. Consequently, methods of data exploitation that do not damage people's lives, such as fairness/discrimination-aware learning, privacy-preserving data mining, or adversarial learning, together comprise the notion of *socially responsible mining*, which it should become an important concept in the near future.

## Acknowledgment

We wish to thank Dr. Sicco Verwer for providing detail information about his work. This work is supported by the grants-in-aid 14658106, 16700157, and 21500154 of the Japan society for the promotion of science.

## References

[1] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. of The 14th Int'l Conf. on Knowledge Discovery and Data Mining*, 2008.

[2] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, pp. 277–292, 2010.

[3] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[4] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. of The 21st Int'l Conf. on Machine Learning*, 2004, pp. 903–910.

[5] "NIPS workshop — inductive transfer: 10 years later," 2005, http://iitrl.acadiau.ca/itws05/.

[6] A. Frank and A. Asuncion, "UCI machine learning repository," University of California, Irvine, School of Information and Computer Sciences, 2010, http://archive.ics.uci.edu/ml.

[7] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 2, 2010.

[8] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *Proc. of the SIAM Int'l Conf. on Data Mining*, 2009, pp. 581–592.

[9] S. Ruggieri, D. Pedreschi, and F. Turini, "Dcube: Discrimination discovery in databases," in *Proc of The ACM SIGMOD Int'l Conf. on Management of Data*, 2010, pp. 1127–1130.

[10] 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 Int'l Conf. on Knowledge Discovery and Data Mining*, 2011, pp. 502–510.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," arxiv.org:1104.3913, 2011.

[12] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. of the 17th Int'l Joint Conf. on Artificial Intelligence*, 2001, pp. 973–978.

[13] C. Perlich, S. Kaufman, and S. Rosset, "Leakage in data mining: Formulation, detection, and avoidance," in *Proc. of The 17th Int'l Conf. on Knowledge Discovery and Data Mining*, 2011, pp. 556–563.

[14] D. Gondek and T. Hofmann, "Non-redundant data clustering," in *Proc. of The 4th IEEE Int'l Conf. on Data Mining*, 2004, pp. 75–82.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[16] C. C. Aggarwal and P. S. Yu, Eds., *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[17] S. Venkatasubramanian, "Measures of anonimity," in *Privacy-Preserving Data Mining: Models and Algorithms*, C. C. Aggarwal and P. S. Yu, Eds. Springer, 2008, ch. 4.

[18] S. Zilberstein, "Using anytime algorithms in intelligent systems," *AI Magazine*, vol. 17, no. 3, pp. 73–86, 1996.

[19] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. of The 24th Int'l Conf. on Machine Learning*, 2007, pp. 193–200.

[20] D. Boyd, "Privacy and publicity in the context of big data," in *Keynote Talk of The 19th Int'l Conf. on World Wide Web*, 2010.