T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma "Fairness-Aware Classifier with Prejudice Remover Regularizer" Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), Part II, pp.35-50 (2012)

# Fairness-aware Classifier with Prejudice Remover Regularizer

Toshihiro Kamishima[1], Shotaro Akaho[1], Hideki Asoh[1], and Jun Sakuma[2]

[1] National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki, 305-8568 Japan, mail@kamishima.net, s.akaho@aist.go.jp, and h.asoh@aist.go.jp, WWW home page: http://www.kamishima.net

[2] University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8577 Japan; and, Japan Science and Technology Agency, 4-1-8, Honcho, Kawaguchi, Saitama, 332-0012 Japan 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 individuals' 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 nondiscriminatory and fair 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. 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, logistic regression, classification, social responsibility, information theory

## 1 Introduction

Data mining techniques are being increasingly used for serious determinations such as credit, insurance rates, employment applications, and so on. 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 guarantee fairness in both social and legal viewpoints; that is, they must be unbiased and nondiscriminatory in relation to sensitive features such as gender, religion, race, ethnicity, handicaps, political convictions, and so on. Thus, sensitive features must be carefully treated in the processes and algorithms for data mining.

There are reasons other than the need to avoid discrimination for prohibiting the use of certain kinds of features. Pariser pointed out a problem that friend candidates recommended to him in Facebook were biased in terms of their political convictions without his permission [15]. For this problem, it would be helpful to make recommendations that are neutral in terms of the user's specified feature, i.e., the candidate friends' political convictions. Further, there are features that cannot legally be exploited due to various regulations or contracts. For example, exploiting insider information and customer data are respectively restricted by stock trading regulation and privacy policies.

Several researchers have recently begun to attempt the development of analytic techniques that are aware of social fairness or discrimination [17,3]. 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 [3] or indirect discrimination [17].

In this paper, we formulate causes of unfairness in data mining, develop widely applicable and efficient techniques to enhance fairness, and evaluate the effectiveness and efficiency of our techniques. First, we consider unfairness in terms of its causes more deeply. 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.

Second, we then focus on indirect prejudice and develop a technique to reduce it. This technique is implemented as regularizers that restrict the learner's 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.

Finally, we perform experiments to test the effectiveness and efficiency of our methods. We evaluate the effectiveness of our approach and examine the balance between prediction accuracy and fairness. We demonstrate that our method can learn a classification model by taking into account the difference in influence of different features on sensitive information.

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 "discrimination" / "discriminatory"

for two reasons. First, as described above, these technologies can be used for various purposes other than avoiding discrimination. Second, because the term *discrimination* is frequently used for the meaning of classification in the data mining literature, using this term becomes highly confusing.

We discuss causes of unfairness in section 2 and propose our methods for enhancing fairness in section 3. Our methods are empirically compared with a 2-naïve-Bayes method proposed by Calders and Verwer in section 4. Section 5 shows related work, and section 6 summarizes our conclusions.

## 2 Fairness in Data Analysis

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

### 2.1 Illustration of the Difficulties in Fairness-aware Learning

We here introduce an example from the literature to show the difficulties in fairness-aware learning [3], which is a simple analytical result for the data set described in section 4.2. The researchers performed a classification problem. 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. There were some other non-sensitive features, $X$. The ratio of Female records comprised about 1/3 of the data set; that is, the number of Female records was much smaller than that of Male records. Additionally, while about 30% of Male records were classified into the High class, only 11% of Female records were. Therefore, Female–High records were the minority in this data set.

In this data set, we describe how Female records tend to be classified into the 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 naïve Bayes classifier from data involving a sensitive feature, the CV score on the predicted classes increases to about 0.34. This shows that Female records are more frequently misclassified to the Low class than Male records; and thus, Female–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 naïve 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* [3], 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 another techniques have to be adopted to correct the unfairness in data mining.

### 2.2 Three Causes of Unfairness

In this section, we discuss the social fairness in data analysis. Previous works [17,3] have focused on unfairness in the resultant determinations. To look more carefully at the problem of fairness in data mining, 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*.

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 $S$, $\Pr^*[X,S]$, we define

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

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

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

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

**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* [17]. 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 this direct prejudice and statistical dependence. After eliminating the sensitive variable, equation (1) can be rewritten as

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

This equation states that $S$ and $Y$ are conditionally independent given $X$, i.e., $\mathrm{Y} \perp\!\!\!\perp \mathrm{S} \,|\, \mathrm{X}$. Hence, we can say that when the condition $\mathrm{Y} \not\perp\!\!\!\perp \mathrm{S} \,|\, \mathrm{X}$ 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 $\Pr[Y, X, S]$ is equal to $\Pr[Y|X]\Pr[S|X]\Pr[X]$, these variables satisfy the condition $\mathrm{Y} \perp\!\!\!\perp \mathrm{S} \,|\, \mathrm{X}$, 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 [17] or a red-lining effect [3] as described in section 2.1. 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 equation (1) is unknown, we adopt sample distributions in equation (2) over a given sample set, $\mathcal{D}$:

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

We refer to this index as a (indirect) *prejudice index* (PI for short). For convenience, the application of the normalization technique for mutual information [21] leads to a *normalized prejudice index* (NPI for short):

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

where $\mathrm{H}(\cdot)$ is an entropy function. $\mathrm{PI}/\mathrm{H}(Y)$ is the ratio of information of $S$ used for predicting $Y$, and $\mathrm{PI}/\mathrm{H}(S)$ is the ratio of information that is exposed if a value of $Y$ is known. This NPI can be interpreted as the geometrical mean of these two ratios. 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 $\mathrm{Y} \perp\!\!\!\perp \mathrm{S} \,|\, \mathrm{X}$ 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. Removal of latent 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 $S$.

**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:

$$\mathrm{UEI} = \sqrt{\frac{1}{2}\sum_{y,s\in\mathcal{D}}\left(\sqrt{\hat{\Pr}[y,s]}-\sqrt{\tilde{\Pr}[y,s]}\right)^2} = \sqrt{1-\sum_{y,s\in\mathcal{D}}\sqrt{\hat{\Pr}[Y,S]\tilde{\Pr}[Y,S]}}. \quad (5)$$

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

**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 [24]. 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 [10].

Regulations or laws that demand the removal of latent 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.

## 3 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.

### 3.1 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 $\mathcal{M}[Y|X, S; \boldsymbol{\Theta}]$, where $\boldsymbol{\Theta}$ 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:

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

We adopted two types of regularizers. The first regularizer is a standard one to avoid over-fitting. We used an $L_2$ regularizer $\|\boldsymbol{\Theta}\|_2^2$. The second regularizer, $R(\mathcal{D}, \boldsymbol{\Theta})$, 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:

$$-\mathcal{L}(\mathcal{D}; \boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D}, \boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2, \tag{7}$$

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:

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

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

### 3.2 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

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

$\sum_{X,S} \tilde{\Pr}[X, S]$ can be replaced with $(1/|\mathcal{D}|) \sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}}$, and then the scaling factor, $1/|\mathcal{D}|$, can be omitted. The argument of the logarithm can be rewritten as $\hat{\Pr}[Y|s_i]/\hat{\Pr}[Y]$, by reducing $\hat{\Pr}[S]$. We obtain

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

The straight way to compute $\hat{\Pr}[y|s]$ is to marginalize $\mathcal{M}[y|X, s; \boldsymbol{\Theta}]\Pr^*[X|s]$ over $X$. However, if the domain of $X$ is large, this marginalization is computationally heavy. We hence take an approach by which this marginalization is replaced with a sample mean. More specifically, this marginalization is formulated by

$$\hat{\Pr}[y|s] = \int_{\mathrm{dom}(X)} \Pr^*[X|s]\mathcal{M}[y|X, s; \boldsymbol{\Theta}]dX,$$

where $\mathrm{dom}(X)$ is the domain of $X$. We approximated this formula by the following sample mean:

$$\hat{\Pr}[y|s] \approx \frac{\sum_{(\mathbf{x}_i, s_i)\in\mathcal{D} \text{ s.t. } s_i=s} \mathcal{M}[y|\mathbf{x}_i, s; \boldsymbol{\Theta}]}{|\{(\mathbf{x}_i, s_i) \in \mathcal{D} \text{ s.t. } s_i = s\}|}. \tag{9}$$

Similarly, we approximated $\hat{\Pr}[y]$ by

$$\hat{\Pr}[y] \approx \frac{\sum_{(\mathbf{x}_i, s_i)\in\mathcal{D}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}]}{|\mathcal{D}|}. \tag{10}$$

Note that in our preliminary work [12], we took the approach of replacing $X$ 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 $s$. However, we unfortunately failed to obtain good approximations by this approach.

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

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

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 in the final determination. In the case of logistic regression, the objective function (7) to minimize is rewritten as

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

where $\mathcal{M}[y|\mathbf{x}, s; \boldsymbol{\Theta}]$ is equation (8) and $\mathrm{R}_{\mathsf{PR}}(\mathcal{D}, \boldsymbol{\Theta})$ is equation (11). In our experiment, parameter sets are initialized by applying standard logistic regression to training sets according to the values of a sensitive feature, and this objective function is minimized by a conjugate gradient method. After this optimization, we obtain an optimal parameter set, $\{\mathbf{w}_s^*\}$.

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

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

## 4 Experiments

We compared our method with Calders and Verwer's method on the real data set used in their previous study [3].

### 4.1 Calders-Verwer's 2-naïve-Bayes

We briefly introduce Calders and Verwer's 2-naïve-Bayes method (CV2NB for short), which was found to be the best of three methods in the previous study using the same dataset [3]. 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 in a similar way as in the estimation of parameters of a naïve 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 naïve Bayes, every non-sensitive feature, $X_i$, depends on both $Y$ and $S$ in the case of CV2NB. $\mathcal{M}[Y, S]$ is then modified so that the resultant CV score approaches zero. Note that we slightly changed this algorithm as described in [12], because the original algorithm may fail to stop.

### 4.2 Experimental Conditions

We summarize our experimental conditions. We tested a previously used real data set [3], as shown in section 2.1. This set includes 16281 data in an adult.test file of the Adult / Census Income distributed at the UCI Repository [7]. 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 naïve Bayes, parameters of models, $\mathcal{M}[X_i|Y, S]$, are estimated by a MAP estimator with multinomial distribution and Dirichlet priors. In our case of 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), naïve Bayes with a sensitive feature (NB), naïve Bayes without a sensitive feature (NBns), and Calders and Verwer's 2-naïve-Bayes (CV2NB). We show the means of the statistics obtained by the five-fold cross-validation.

**Table 1.** 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 | 4.91E-02 | 0.039 | 0.184 | 1.99E-01 |
| PR $\eta$=5 | 0.842 | 0.240 | 4.24E-02 | 0.088 | 0.143 | 1.91E-01 |
| PR $\eta$=15 | 0.801 | 0.158 | 2.38E-02 | 0.212 | 0.050 | 1.62E-01 |
| PR $\eta$=30 | 0.769 | 0.046 | 1.68E-02 | 0.191 | 0.010 | 3.94E-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 $n_1 \times 10^{n_2}$. $L_2$ regularizer: $\lambda = 1$.

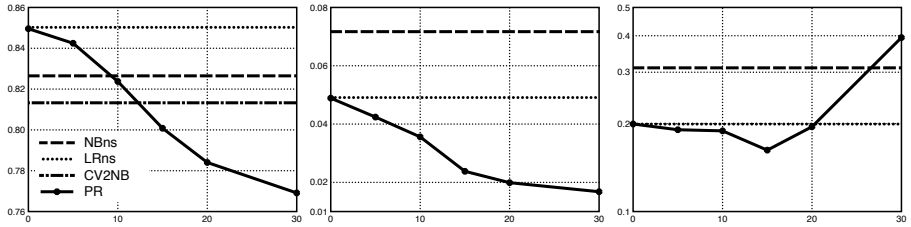
### 4.3 Experimental Results

Table 1 shows accuracies (Acc), NPI and UEI in section 2, 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 was sacrificed by obtaining a unit of information about the correct label. This can be used to measure the efficiency in the trade-off between prediction accuracy and prejudice removal. The 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 1. Compared with NBns, our method was superior both in accuracy and NPI at $\eta = 5$; and hence, ours was superior in the efficiency index, PI/MI. When comparing LRns, the prejudice in decisions was successfully removed by our prejudice remover in exchange for the prediction accuracy. 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 1, as $\eta$ increased, our PR generally become degraded in accuracy.

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 1. We focus on our PR method. The increase of $\eta$ generally damaged accuracy because the prejudice remover regularizer is designed to remove prejudice by sacrificing accuracy in prediction. This effect was observed by the increase in NPI. The peak in trade-off efficiency was observed at $\eta = 15$. More prejudice could be removed by increasing $\eta$, but the accuracy in prediction was fairly damaged.

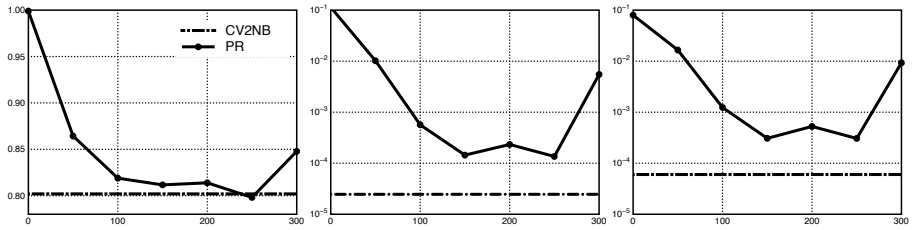
We next compared our PR with other methods. By observing Figure 1(c), our PR demonstrated better performance in trade-offs between accuracy and prejudice removal than the NBns. When compared to the baseline LRns, more prejudice was successfully removed by increasing $\eta$. The Figure 1(a) showed that




(a) Acc (b) NPI (c) PI/MI

**Fig. 1.** The change in performance for the Adult / Census Income data according to $\eta$

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. Larger Acc values indicate better performance, and smaller NPI and PI/MI values indicate better performance. NPI and PI/MI of CV2NB were out of the bounds of these charts and are properly noted in Table 1.




(a) Acc (b) NPI (c) PI/MI

**Fig. 2.** The change in performance for our synthetic data according to $\eta$

NOTE: The meanings of axes and line styles are the same as in Figure 1.

this was achieved by sacrificing the prediction accuracy. The efficiencies in the trade-offs of our PR was better than those of LRns if $\eta$ ranged between 0 and 20. The performance of CV2NB was fairly good, and our PR was inferior to it except for accuracy at the lower range of $\eta$.

To show how the difference in the prejudice removal between CV2NB and PR is brought about by the ability of our method to take into account the difference in influence of different features on sensitive information, we applied our PR to a synthetic data set. To synthesize data, $\epsilon_i$ was sampled from the normal distribution $\mathcal{N}(0,1)$, and $s_i \in \{0,1\}$ was sampled uniformly at random. The first feature $x_{ai} = \epsilon_i$, and the second feature $x_{bi} = 1 + \epsilon_i$ if $s_i = 1$; otherwise $x_{bi} = -1 + \epsilon_i$. The class $y_i$ was set to 0 if $x_{ai} + x_{bi} < 0$; otherwise 1. We generated 20 000 samples and applied CV2NB and our PR by changing $\eta$ from 0 to 300. Because $x_{ai}$ and $x_{bi}$ are equivalent up to bias, these two features are comparable in usefulness for class prediction. The first feature, $x_{ai}$, is independent from $s_i$, while the second feature, $x_{bi}$, depends on $s_i$.

**Table 2.** The learned weight vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ in equation (8)

| | $\mathbf{w}_0$ | $\mathbf{w}_1$ |
|---|---|---|
| $\eta = 0$ | [11.3, 11.3 ,−0.0257] | [11.3, 11.4 ,0.0595] |
| $\eta = 150$ | [55.3,−53.0, −53.6 ] | [56.1,−54.1, 53.6 ] |

NOTE: The first, second, and third elements of $\mathbf{w}_s$ were weights for the first feature, $x_{ai}$, the second feature, $x_{bi}$, and a bias constant, respectively.

We showed the change in three indexes, accuracy, NPI, and PI/MI, on independently generated test data according to the parameter $\eta$ in Figure 1. Unfortunately, results became unstable if $\eta$ is larger than 200 because the objective function (12) has many local minima for large $\eta$. However, when comparing the results in Table 1 with those in this figure, the differences in NPI derived by CV2NB and PR became much smaller.

To exemplify the reason for these differences, we then showed the learned weight vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ in equation (8) in Table 2. By observing the weights more carefully, the weights for $x_{ai}$ and $x_{bi}$ were roughly equal when $\eta = 0$. However, when $\eta = 150$, the absolute values of weights for $x_{bi}$ were smaller than those for $x_{ai}$. This indicates that to remove prejudice, our PR tries to ignore features that depend on a sensitive feature. Therefore, if there are features that are useful for classification and additionally independent from a sensitive feature, our PR can remove prejudice effectively. In other words, our method is designed to learn a classification model by taking into account difference in influence of different features on sensitive information. On the other hand, according to the generative model (13), CV2NB treats all features equally and simply modifies the $\mathcal{M}[Y, S]$ for removing prejudice. Therefore, CV2NB cannot learn a model that reflects such differences.

This difference would cause the following effect in practical use. When considering a case of credit scoring, because CV2NB treats all features equally, scores of all individuals who are in a sensitive state would be raised equally. However, the repayment capacities of these individuals are certainly unequal, and our method can change credit scoring by taking into account individuals' repayment capacity. On the other hand, if the repayment capacities of all individuals in a sensitive state are nearly equal, our method cannot reduce prejudice without degrading prediction accuracy. However, CV2NB can remove prejudice independently of the states of individuals' repayment capacity. Note that fair decision-making that takes into account the differences in effects of features has also been discussed in [13,23].

In summary, our PR could successfully reduce indirect prejudice when compared with baseline methods. Our method is inferior to CV2NB in its efficiency of prejudice removal, but it can learn a classification rule by taking into account the difference in influence of different features on sensitive information. Additionally, our framework has the advantage that it can be applied to any probabilistic discriminative classifier.

## 5 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 [17].They advocated the notion of $\alpha$-*protection*, which is the condition that association rules were fair. Given a rule whose consequent exhibited determination is disadvantageous to individuals, it would be unfair if the confidence of the rule substantially increased by adding a condition associated with 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 [18]. 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 [20].

Calders and Verwer proposed several methods to modify naïve Bayes for enhancing fairness as described in section 4.1 [3]. Kamiran et al. developed algorithms for learning decision trees while taking fairness consideration [11]. When choosing features to divide training examples at non-leaf nodes of decision trees, their algorithms take care of the information gain regarding sensitive information as well as about target decisions. Additionally, the labels at leaf nodes are changed so as to avoid unfair decisions.

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 [13]. 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 [5]. 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.

Žliobaitė et al. discussed handling conditional discrimination [23]. They considered the case where even if the difference between probabilities of receiving advantageous judgment given different values of sensitive features, some extent of the difference can be explained based on the values of non-sensitive features.

For example, even though females are less frequently admitted to a university than males, this decision is considered as fair if this is due to the fact that females tend to try more competitive programs. They proposed a sampling technique to remove the unfair information from training samples while excluding such explainable factors.

In a broad sense, fairness-aware learning is related to causal inference [16], because the final decision becomes unfair if the decision depends on a sensitive status. Fairness in data mining can be interpreted as a sub-notion of legitimacy, which means that models can be deployed in the real world [19]. Gondek and Hofmann devised a method for finding clusters that were not relevant to a given grouping [8]. 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 [9].

The removal of prejudice is closely related to privacy-preserving data mining [1], 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 [22]. 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. In our case, the leaked information is quantified by mutual information, but other criteria for privacy, such as differential privacy [14], might be used for the purpose of maintaining fairness.

Techniques of cost-sensitive learning [6] might be helpful for addressing underestimation problems.

As described in section 2.2, 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" [10]. Among many types of transfer learning, the problem of a sample selection bias [24] 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 [4]. Using these approaches, if given a small amount of fairly labeled data, other data sets that might be unfairly labeled would be correctly processed.

## 6 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 characteristics of our methods.

Research on fairness-aware learning is just beginning, and 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. Because of the lack of convexity of the objective function, our method is occasionally trapped by local minima. To avoid this, we plan to try other types of independence indexes, such as kurtosis, which has been used for independent component analysis. If a sensitive feature is a multivariate variable whose domain is large or is a real variable, our current prejudice remover cannot be applied directly; these limitations must be overcome.

The use of data mining technologies in our society will only become greater and greater. Unfortunately, their results can occasionally damage people's lives [2]. 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.

**Acknowledgments**: We wish to thank Dr. Sicco Verwer for providing detail information about his work and to thank the PADM2011 workshop organizers and anonymous reviewers for their valuable comments. This work is supported by MEXT/JSPS KAKENHI Grant Number 16700157, 21500154, 22500142, 23240043, and 24500194, and JST PRESTO 09152492.

## References

1. Aggarwal, C.C., Yu, P.S. (eds.): Privacy-Preserving Data Mining: Models and Algorithms. Springer (2008)
2. Boyd, D.: Privacy and publicity in the context of big data. In: Keynote Talk of The 19th Int'l Conf. on World Wide Web (2010)
3. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21, 277–292 (2010)
4. Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: Proc. of the 24th Int'l Conf. on Machine Learning. pp. 193–200 (2007)

5. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. arxiv.org:1104.3913 (2011)
6. Elkan, C.: The foundations of cost-sensitive learning. In: Proc. of the 17th Int'l Joint Conf. on Artificial Intelligence. pp. 973–978 (2001)
7. Frank, A., Asuncion, A.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences (2010), ⟨`http://archive.ics.uci.edu/ml`⟩
8. Gondek, D., Hofmann, T.: Non-redundant data clustering. In: Proc. of the 4th IEEE Int'l Conf. on Data Mining. pp. 75–82 (2004)
9. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience (2001)
10. NIPS workshop — inductive transfer: 10 years later (2005), ⟨`http://iitrl.acadiau.ca/itws05/`⟩
11. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Proc. of the 10th IEEE Int'l Conf. on Data Mining. pp. 869–874 (2010)
12. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: Proc. of The 3rd IEEE Int'l Workshop on Privacy Aspects of Data Mining. pp. 643–650 (2011)
13. Luong, B.T., Ruggieri, S., Turini, F.: 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. pp. 502–510 (2011)
14. Nissim, K.: Private data analysis via output perturbation. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, chap. 4. Springer (2008)
15. Pariser, E.: The Filter Bubble: What The Internet Is Hiding From You. Viking (2011)
16. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, 2nd edn. (2009)
17. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proc. of the 14th Int'l Conf. on Knowledge Discovery and Data Mining (2008)
18. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proc. of the SIAM Int'l Conf. on Data Mining. pp. 581–592 (2009)
19. Perlich, C., Kaufman, S., Rosset, S.: Leakage in data mining: Formulation, detection, and avoidance. In: Proc. of the 17th Int'l Conf. on Knowledge Discovery and Data Mining. pp. 556–563 (2011)
20. Ruggieri, S., Pedreschi, D., Turini, F.: DCUBE: Discrimination discovery in databases. In: Proc of The ACM SIGMOD Int'l Conf. on Management of Data. pp. 1127–1130 (2010)
21. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
22. Venkatasubramanian, S.: Measures of anonimity. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, chap. 4. Springer (2008)
23. Žliobaitė, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Proc. of the 11th IEEE Int'l Conf. on Data Mining (2011)
24. Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: Proc. of the 21st Int'l Conf. on Machine Learning. pp. 903–910 (2004)