Classification with No Discrimination by Preferential Sampling

Faisal Kamiran F.KAMIRAN@TUE.NL
Toon Calders T.CALDERS@TUE.NL
Eindhoven University of Technology, The Netherlands

Abstract

The concept of classification without discrimination is a new area of research. (Kamiran & Calders, 2009) introduced the idea of *Classification with No Discrimination (CND)* and proposed a solution based on "massaging" the data to remove the discrimination from it with the least possible changes. In this paper, we propose a new solution to the *CND* problem by introducing a sampling scheme for making the data discrimination free instead of relabeling the dataset. On the resulting non-discriminatory dataset we then learn a classifier. This new method is not only less intrusive as compared to the "massaging" but also outperforms the "reweighing" approach of (Calders et al., 2009). The proposed method has been implemented and experimental results on the Census Income dataset show promising results: in all experiments our method performs on-par with the state-of-the art non-discriminatory techniques.

1. Introduction

The concept of *Classification with No Discrimination (CND)* was formally introduced in (Kamiran & Calders, 2009) as follows: *Classification models are trained on historical data for the prediction of class labels of unknown data samples. Often, however, the historical data is biased towards certain groups or classes of objects. For example, throughout the years, in a certain organization black people might systematically have been denied from jobs. As such, the historical employment information of this company concerning job applications will be biased towards giving jobs to white people while denying jobs from black people. In order to reduce this type of racial discrimination, new laws requiring equal job opportunity have been enacted by the government. As such, the organization receives instructions in the form of, e.g., minimum quota for black employees. Suppose now that the company wants to partially automate its recruitment strategy by learning a classifier that predicts the most likely candidates for a job. As the historical recruitment data of the company is biased, the learned model may show unlawfully prejudiced behavior. This partial attitude of the learned model leads to discriminatory outcomes for future unlabeled data objects.*

Simply removing the discriminatory attribute from the training data in the learning of a classifier for the classification of future data objects is not enough to solve this problem, because often other attributes will still allow for the identification of the discriminated community (Kamiran & Calders, 2009).

In (Kamiran & Calders, 2009), a method was proposed based on first "massaging" the data by changing class labels of selected objects in the training data in order to obtain a discrimination free dataset. For massaging the data, first a ranker for predicting the class attribute without taking into account discrimination is learnt. This ranker is then used to rank the data objects according to their probability of being in the desired class, e.g., job = yes. The class labels of the most likely *victims* (training instances of the discriminated community with a negative label but a high positive class probability) and *profiters* (training instances of the favored community with a positive label but a low positive class probability) are changed. The modified data is then used for learning a classifier with no discrimination for future decisions.

One disadvantage of "massaging" the data, however, is that it is very intrusive. In this paper, therefore, we introduce a *Preferential Sampling (PS)* scheme to make the dataset bias free. Instead, *PS* changes the distribution of different data objects for a given data to make it discrimination free. The idea is that the data objects close to the decision boundaries are more prone to be the victim of discrimination. We change the distribution of these borderline objects to make the dataset discrimination free. To know the least certain elements, we use a ranking function, learned on the original data, to identify the data objects close to the borderline. Then, based on the sanitized data, a non-

Appearing in *Proceedings of the 19^{th} Machine Learning conference of Belgium and The Netherlands*. Copyright 2010 by the author(s)/owner(s).

discriminatory model can be learned. The fact that this model is learned on non-discriminatory data reduces the prejudicial behavior for future classification.

The *PS* scheme was implemented and tested on the Census Income dataset (Asuncion & Newman, 2007). Using our proposed *PS* scheme we are able to learn classifiers that no longer discriminate future data, without loosing too much accuracy.

In summary, the contributions of this paper is as follows: *A new method called Preferential Sampling (PS) for creating discrimination free training set based on sampling is introduced that does not require the change of any class label.*

The paper is organized as follows: in Section 2 we revisit the definitions of (Kamiran & Calders, 2009) and introduce the discrimination measure. In Section 3 we propose a solution for the problem based on *Preferential Sampling* and the results of different experiments are shown in Section 4. Section 5 discusses related work and Section 6 concludes and gives some directions for future work.

2. Problem Formulation

We use the original formulation of (Kamiran & Calders, 2009). We assume a set of attributes $A = \{a_1, ..., a_m\}$, and a binary set of class labels $C = \{c_1, c_2\}$. $dom(a_i)$ refers to the domain of the ith attribute. A *labeled dataset over A with labels from C* is defined as a finite set of tuples $(x_1, \ldots, x_n, c)$ with

$$x = (x_1, ..., x_m) \in dom(a_1) \times ... \times dom(a_m) \ ,$$

and the *class label* $c \in C$. We will often use $x.a_i$ to refer to the component x_i of x corresponding to the attribute a_i, and to its class label as $x.c$. Let $D = \{(x^1, c^1), \ldots, (x^n, c^n)\}$ be a labeled dataset where $(x^i, c^i) = (x^i_1, \ldots, x^i_m, c^i)$.

We assume that a special attribute $SA \in A$, called the *Sensitive Attribute*, and a special value $s \in dom(SA)$, called *Sensitive Attribute Value* have been given. The semantics of SA and s is that they define the discriminated community; e.g., $SA = Ethnicity$ and $s = Black$. For reasons of simplicity we will assume that the domain of SA is binary; i.e., $dom(SA) = \{s, \overline{s}\}$.

Obviously, we can easily transform a dataset with multiple attribute values for SA into a binary one by replacing all values $v \in dom(SA) \setminus \{s\}$ with a new dedicated value $\overline{s}$. Furthermore, we assume that a *desired class* $+ \in C$ has been given. In the credit evaluation example, e.g., $+$ would be the *Good* credit class. Let now

$$\begin{aligned} s &:= |\{x \in D \mid x.SA = s\}| \\ \overline{s} &:= |\{x \in D \mid x.SA \neq s\}| \end{aligned}$$

We divide the dataset into four groups DP (<u>D</u>iscriminated community with <u>P</u>ositive class labels), DN (<u>D</u>iscriminated community with <u>N</u>egative class labels), PP (<u>P</u>rivileged community with <u>P</u>ositive class labels) and PN (<u>P</u>rivileged community with <u>N</u>egative class labels) on the basis of SA- and $Class$- values:

$$\begin{aligned} DP &:= \{x \in D \mid x.SA = s \wedge x.c = +\} \\ DN &:= \{x \in D \mid x.SA = s \wedge x.c = -\} \\ PP &:= \{x \in D \mid x.SA = \overline{s} \wedge x.c = +\} \\ PN &:= \{x \in D \mid x.SA = \overline{s} \wedge x.c = -\} \end{aligned}$$

The *discrimination in* D *of* s *towards* $+$, denoted $Disc(D, SA, s, +)$, is now defined as:

$$Disc(D, SA, s, +) \ := \ conf(\overline{s} \rightarrow +) - conf(s \rightarrow +) \ ,$$

where

$$\begin{aligned} conf(\overline{s} \rightarrow +) &:= \frac{|PP|}{|PP| + |PN|}, \text{ and} \\ conf(s \rightarrow +) &:= \frac{|DP|}{|DP| + |DN|}. \end{aligned}$$

The goal of this paper is now to develop a classification model such that when it is trained on a biased dataset D, it does show impartial behavior on future data. For a given dataset D_f, $CND(D_f)$ denotes the labeled dataset resulting from applying CND on D_f. More formally, the *Classification with No Discrimination (CND)* problem is defined as follows (Kamiran & Calders, 2009):

Given a biased dataset D, a sensitive attribute SA and value s, and a desired class +. The output is a classifier CND such that, even though the discrimination in the training data $Disc(D, SA, s, +)$ might be high, on unseen future data objects D_f, $Disc(CND(D_f), SA, s, +)$ must be low, while at the same time the predictive accuracy of CND on this data D_f must be high.

Example 1 *Consider the fictive database given in Table 1, storing credit scores decisions. Each data object has 4 attributes which are Marital Status, Age, Housing, Credit History and one class attribute Credit Class with class values Good and Bad. The domain of the attribute Age consists of the values Young and Aged. Using our discrimination identifier function for this database, with Age as SA, Young as s and Good as desired class, we get:*

$$\begin{aligned} & Disc(D, Age, Young, Good) = \\ & \qquad Conf(Aged \rightarrow Good) \\ & \qquad - Conf(Young \rightarrow Good) \ = \ 40\%, \end{aligned}$$

Table 1. Run-Through Example: Test data for classification.

Marit_stat	Age	House	Cred_hist	Class
Single	Young	Rent	No credit	*Bad*
Single	Young	Rent	Bad	*Bad*
Single	Young	Own	No credit	*Good*
Married	Young	Own	Excellent	*Good*
Married	Young	Rent	Excellent	*Bad*
Divorced	Aged	Rent	Bad	*Good*
Divorced	Aged	Own	Excellent	*Good*
Divorced	Aged	Own	No credit	*Good*
Married	Aged	Rent	Bad	*Bad*
Married	Aged	Rent	No credit	*Good*

which shows that Aged people have 40% more chance to be assigned the credit class Good than Young people.
Hence the data of Table 1 is biased in favor of Aged people. Suppose now that, nevertheless, we want to build a classification model such that the accuracy of assigning an account holder with the correct credit class is high, but at the same time we want to remove the Age-discrimination, e.g., because we are interested in attracting young people to our bank. In this situation we want a classification model which is learnt on the data of Table 1 but classifies future loan applicants without discriminating on Age.

3. Problem Solution

Our approach consists of sampling the data objects with replacement to make the dataset bias free. The main idea is to chose those data objects which are the best choice for the removal of discrimination from the dataset. The modified data is then used for learning a bias free classifier for future decisions.

3.1. Preferential Sampling

In *Preferential Sampling (PS)* we use the idea that data objects close to the borderline are more prone to have been discriminated or favored due to discrimination in the dataset and give preference to them for sampling. To identify the borderline objects, *PS* starts by learning a ranker on the training data. *PS* uses this ranker to sort the data objects of DP and PP in ascending order, and the objects of DN and PN in descending order; both w.r.t. the positive class probability. Such arrangement of data objects makes sure that the higher the rank an element occupies, the closer it is to the borderline.
PS starts from the original training dataset and iteratively duplicates (for the groups DP and PN) and removes objects (for the groups DN and PP) in the following way:

Decreasing the size of a group is always done by removing the data objects closest to the borderline; i.e., the top element.

Algorithm 1 *Preferential Sampling*

```
Input $(D, SA, c)$ Output Classifier $CND$ learnt on $D$ without discrimination
For all the data objects with $SA = s$: add in $DP$ if $c = +$ else add in $DN$
For all the data objects with $SA = \bar{s}$: add in $PP$ if $c = +$ else add in $PN$
Learn a ranker which arranges the data objects w.r.t. their probability of being in the desired class
Arrange the elements of DP and PP in ascending order and elements of DN and PN in descending order
Calculate the expected size for each combination of $v \in SA$ and $c \in C$ by $\frac{|v| \times |c|}{|D|}$
Change the sample size for each group by either re-substitution or skipping of top order elements iteratively
Move the top order elements with their duplicates (if exists) to the bottom of ranking after each iteration
Train a classifier $CND$ on the re-sampled $D$
Return $CND$
```

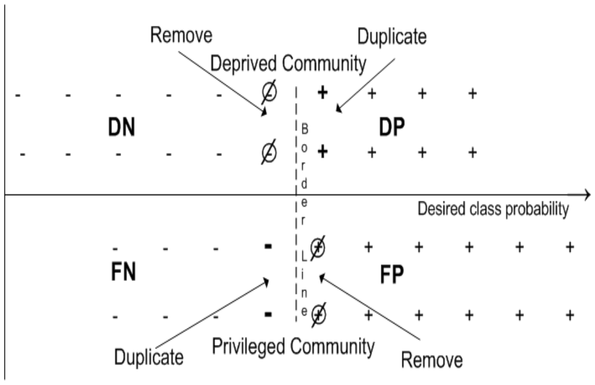

Figure 1. Pictorial representation of *Preferential Sampling* scheme. The re-substituted data points are in bold while the encircled ones are skipped.

Increasing the sample size is done by duplication of the data object closest to the borderline. When an object has been duplicated, together with its duplicate, it is moved to the bottom of the ranking. We repeat this procedure until the desired number of objects is obtained.

In most cases, only a few data objects have to be duplicated or removed.

Figure 1 gives an illustration of *Preferential Sampling (PS)*, showing 40 data point. Data points of the desired class and the negative class are represented by $+$ and $-$ symbols respectively. The X-axis shows the probability of each data object to be in the desired class: the higher up, the higher the probability. The data points plotted along the positive Y-axis and the negative Y-axis represent the discriminated and the privileged communities respectively. *PS* works in the following steps:

(1) Divide the data objects into the four groups, DP, DN, PP, and PN. Steps 1 and 2 of Algorithm 1 describe the division of the dataset into four groups.
(2) Learn a ranker on the complete training data to arrange the data objects according to their probability of being in class $+$. Any ranking algorithm may be used for calculating the class probability of each data tuple. This ranking will be used to identify the borderline data objects. In Figure 1, the rank of the objects is depicted by their horizontal position: the more to the right, the higher the rank. Steps 3 and 4 of Algorithm 1 describe it.
(3) Calculate the expected size for each group to make the dataset bias free. For the data given in Figure 1, Table 2 shows the actual and expected sample size for each group to make the data unbiased. Step 5 of Algorithm 1 gives the formula for expected size calculation.

Table 2. Actual and expected size of each group of data shown in Figure 1.

Sample Size	**DP**	**DN**	**PP**	**PN**
Actual	8	12	12	8
Expected	10	10	10	10

(4) Finally apply sampling with replacement to increase the size of DP and PN. The data points of Figure 1 in bold are duplicated to increase the sample size of DP and PN, and the sizes of DN and PP are reduced by removing the data objects close to the borderline. The encircled data points of Figure 1 will be removed. Steps 6 and 7 of Algorithm 1 describe it. Now this re-sampled dataset is used for learning a discrimination free classifier.

3.2. Running Example

We consider the data given in Table 1 and apply the *PS* scheme to make it discrimination free. The *PS* method divides this dataset into four groups according to the value of Age and class attribute. Our DP and DN include *Young* loan applicants with class *Good* and *Bad* respectively while PP and PN include *Aged* loan applicants with class *Good* and *Bad* respectively. The data objects of the DP and PP are arranged by the ranker in ascending order w.r.t. positive class probability while of DN and PN are arranged in descending order. To estimate the number of objects to be re-substituted or skipped, the *PS* method calculates the expected size of each group for unbiased dataset. We calculate the expected size of DP:

$$\begin{aligned} E(SA = Young \mid c = Good) &= \frac{|Young| \wedge |Good|}{|D|} \\ &= \frac{5 \times 6}{10} = 3 \ . \end{aligned}$$

We see that if the data is unbiased, the sample size of DP should be 3 but actually it is 2. Therefor, we will have to increase the sample size from 2 to 3 by re-substitution of an object. Similarly the size of DN in an unbiased dataset would be 2 instead of 3. The *PS* method arranges the data objects of DP in ascending order and of DN in descending order w.r.t. their probability of being in *Good* class by using a ranker as shown in Table 3. To change the sample size, we will re-substitute or skip the top order data object (in bold) of DP and DN.

Table 3. Data objects of DP and DN arranged in ascending and descending order respectively according to the desired class probability.

Marit_stat	**Age**	**House**	**Cred_hist**	**Class**	**Prob**
Single	**Young**	**Own**	**No credit**	***Good***	**70%**
Married	Young	Own	Excellent	*Good*	73%

Marit_stat	**Age**	**House**	**Cred_hist**	**Class**	**Prob**
Married	**Young**	**Rent**	**Excellent**	***Bad***	24%
Single	Young	Rent	No credit	*Bad*	22%
Single	Young	Rent	Bad	*Bad*	8%

Similarly, *PS* calculate the expected size of PP and PN which is 3 and 2 respectively. It means, we will have to skip and re-substitute one borderline data object of PP and PN respectively. The *PS* arranges the PP in ascending order while PN in descending order and changes their sizes. The top data object of PN (borderline object) is re-substituted to increase its size while the top data object of PP (borderline object) is skipped to reduce its size, as shown in bold in Table 4.

Table 4. Data objects of PP and PN arranged in ascending and descending order respectively according to the desired class probability.

Marit_stat	**Age**	**House**	**Cred_hist**	**Class**	**Prob**
Married	**Aged**	**Rent**	**No credit**	***Good***	**59%**
Divorced	Aged	Rent	Bad	*Good*	65%
Divorced	Aged	Own	Excellent	*Good*	97%
Divorced	Aged	Own	No credit	*Good*	98%

Marit_stat	**Age**	**House**	**Cred_hist**	**Class**	**Prob**
Married	**Aged**	**Rent**	**Bad**	***Bad***	**32%**

So, after the application of *PS*, we get an unbiased data shown in Table 5 which can be used for discrimination free classifier learning. The duplicated data objects of the unbiased dataset are shown in bold.

Table 5. Unbiased dataset after the application of the *PS* scheme.

Marit_stat	Age	House	Cred_hist	Class
Single	Young	Rent	No credit	*Bad*
Single	Young	Rent	Bad	*Bad*
Single	Young	Own	No credit	*Good*
Single	**Young**	**Own**	**No credit**	***Good***
Married	Young	Own	Excellent	*Good*
Divorced	Aged	Rent	Bad	*Good*
Divorced	Aged	Own	Excellent	*Good*
Divorced	Aged	Own	No credit	*Good*
Married	Aged	Rent	Bad	*Bad*
Married	**Aged**	**Rent**	**Bad**	***Bad***

4. Experiments

In our experiments, we compare the following four approaches:

1. We learn a classification algorithm directly on the original data without applying any preprocessing technique on it. We refer this approach to as *No Preprocessing*.

2. We learn a classification algorithm directly on the original data after removing the *Sensitive Attribute*. We refer this approach to as *No SA*.

3. We make the dataset discrimination free by applying the “massaging” method of (Kamiran & Calders, 2009) and then learn a classifier on this unbiased dataset. We use a Naive Bayesian model as ranker. We refer this approach to as *Massaging*.

4. For comparison, we change the sample size of different groups, i.e., DP, DN, PP and PN, by randomly removing or re-substituting the data objects to make the dataset discrimination free. Then we learn a classification algorithms on the discrimination-free data. We refer to this approach as *Uniform Sampling*. *Uniform Sampling* is exactly similar to the *Preferential Sampling* scheme except in *Uniform Sampling* each data object has the uniform probability to be duplicated or removed while in the *PS* scheme the data objects close the decision boundaries are more likely to be duplicated or removed.

5. We apply the “reweighing” method of (Calders et al., 2009) to make the dataset discrimination free. The “reweighing” method is very similar to *Uniform Sampling*, but instead of sampling the dataset, weights are assigned to data objects directly to make the dataset unbiased. We refer this approach to as ”Reweighing”.

6. Our proposed approach; i.e., we apply *PS* on the training data to make it discrimination free. The ranking function will be based on a Naive Bayesian model learned on the raw data. Then we learn a classification algorithms on the discrimination-free data. We refer to this technique as *Preferential Sampling*.

4.1. The Census Income dataset

In our experiments we used the Census Income dataset which is available in the UCI ML-repository (Asuncion & Newman, 2007). Census Income has 48842 instances of which we used only a random sample of 1/3 for reasons of efficiency. Census Income contains demographic information about people and the associated prediction task is to determine whether a person makes over 50K per year or not, i.e., income class *High* or *Low* will be predicted. We will denote income class *High* as $+$ (desired class) and income class *Low* as $-$. Each data object is described by 14 attributes which include 8 categorical and 6 numerical attributes. We excluded the attribute *fnlwgt* from our experiments (as suggested in the documentation of the dataset). We use Sex as SA and $female(f)$ as s. In our sample of the dataset, 5421 citizens have $Sex = f$ and 10860 have $Sex = m$. The discrimination between $Sex = f$ and $Sex = m$ for $c = +$ is as high as 19.13%. The goal is now to learn a classifier that has minimal discrimination between $Sex = m$ and $Sex = f$ for class *High* income while maintaining a high accuracy. All reported accuracy numbers in the paper were obtained using 10-fold cross-validation and reflect the true accuracy; that is, on unaltered (un-sampled) test data.

4.2. Performance Comparison

We test the accuracy and the discriminatory attitude of the proposed classification model with *Sex* as SA. We apply a decision tree learner: the Weka implementation of the C4.5 classifier (label J48), a Naïve Bayes Classifier (label NBS), two nearest neighbor classifiers with respectively 1 and 7 neighbors (label IBk-1 and IBk-7) as base classifiers that are learned on the original data (label No Preprocess), original data without *Sensitive Attribute* (label No SA), unbiased data by applying “massaging” scheme of (Kamiran & Calders, 2009) (label Massaging), unbiased data by applying “reweighing” scheme of (Calders et al., 2009) (label Reweighing), unbiased data with *Uniform Sampling* (label US) and unbiased data with *Preferential Sampling* (label PS). Table 6 shows the results. We observe that all the classification algorithms classify the future data objects with less discrimination after the application of *Preferential Sampling* technique, e.g., discrimination level reduces to 1.5% from 16.19%. *PS* always outperforms all other competitive methods used in our experiments except Massaging. Moreover, *PS* gives better accuracy scores than Massaging and is less intrusive because it does not change any class label. In all our experiments *PS* always outperforms the “reweighing” scheme of (Calders et al., 2009) which is less intrusive. We hypothesize that the change in the training data has a more impact on an unstable classifier, e.g., J48 than stable classifiers, e.g., NBS. The results given in Table 6 support this hypothesis. We observe that the resultant discrimination level has reduced drastically with unstable classifiers, e.g. J48 while the effect is relatively small over the stable classifiers, e.g., NBS.

Table 6. Results of 10-fold cross validation with *Sex* as SA.

Classifier	Preprocessing	Disc.	Accuracy
J48	No Preprocess	16.19 ± 1.9	85.79 ± 0.94
	No SA	16.07 ± 2	85.73 ± 0.9
	Massaging	0.35 ± 2.3	83.28 ± 1.1
	Reweighing	6.91 ± 2	84.93 ± 0.87
	US	8.87 ± 1.9	85.07 ± 0.91
	PS	1.5 ± 2.4	83.94 ± 1.26
NBS	No Preprocess	15.56 ± 1.18	82.84 ± 1.25
	No SA	13.38 ± 1.47	82.78 ± 1.21
	Massaging	9.98 ± 1.76	81.88 ± 1.18
	Reweighing	10.32 ± 1.45	82.43 ± 1.06
	US	10.1 ± 1.77	82.45 ± 1.1
	PS	10.95 ± 1.76	83.98 ± 1.12
IBK-1	No Preprocess	16.76 ± 2.33	79.74 ± 0.64
	No SA	16.35 ± 2.45	79.92 ± 0.78
	Massaging	0.36 ± 2.75	79.95 ± 1.07
	Reweighing	16.29 ± 2.48	79.64 ± 0.62
	US	12.92 ± 2.75	79.64 ± 0.82
	PS	3.74 ± 2.77	80.68 ± 0.89
IBK-7	No Preprocess	18.63 ± 1.10	82.71 ± 0.88
	No SA	18.34 ± 1.14	82.77 ± 0.89
	Massaging	2.59 ± 2.26	81.24 ± 1.14
	Reweighing	2.63 ± 2.11	81.36 ± 0.96
	US	6.12 ± 2.38	81.79 ± 0.96
	PS	0.17 ± 2.64	80.56 ± 1.23

Table 6 also shows the accuracy scores by above mentioned four approaches. We find that the accuracy for *Preferential Sampling* scheme drops to some extent but the difference in our experiments is negligible. We observe that stable classifiers give more accurate results but do not reduce the discrimination so much after the application of *Preferential Sampling* than other four methods.

5. Related Work

Due to space restriction, we discuss the most important work published in this research area. (Kamiran & Calders, 2009) introduced the concept of discrimination aware classification and proposed a solution to the problem based on changing the class labels. *PS* introduces a less intrusive technique to make the dataset unbiased than changing the class labels. The work of (Pedreschi et al., 2008; Pedreschi et al., 2009) has similar motivation towards the solution of the discrimination problem. They concentrate on identifying discriminatory rules that are present in a dataset, hence they learn potential discriminatory guidelines that have been followed in the decision procedure.

Our work is also closely related to class imbalance problem. (Chawla et al., 2002) introduced a synthetic minority over-sampling technique (SMOTE) for two class problems that over-sampled the minority class by creating synthetic examples rather than replicating examples. In contrast *PS* concentrates only on border regions. It changes the representation of data objects of each class according to the value of SA and class attribute.

6. Conclusion and Future Work

Classification with No Discrimination by Preferential Sampling is an excellent solution to the discrimination problem. It gives promising results with both stable and unstable classifiers. It reduces the discrimination level by maintaining a high accuracy level. It gives similar performance to “massaging” but without changing the dataset and always outperforms the “reweighing” scheme of (Calders et al., 2009).

As future work we are interested in extending the discrimination model itself; in many cases, it is acceptable from an ethical and legal point of view to have some discrimination, as long as it can be explained by other attributes. This extension of the model will help us to justify that the discrimination can be removed from those regions only where it is legally or ethically unacceptable. Therefore it would be interesting to refine our model to *Classification with Conditional Discrimination.*

References

Asuncion, A., & Newman, D. (2007). UCI machine learning repository. .

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *IEEE ICDM Workshop on Domain Driven Data Mining*. IEEE press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, *16*, 321–357.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *Proc. IC4 09*. IEEE press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. *Proceedings of SIAM Data Mining conference*.