Discrimination-Aware Association Rule Mining for Unbiased Data Analytics

Ling Luo[1,2(✉)], Wei Liu[2,3], Irena Koprinska[1], and Fang Chen[2]

[1] School of Information Technologies, University of Sydney, Sydney, Australia
{ling.luo,irena.koprinska}@sydney.edu.au
[2] NICTA ATP Laboratory, Sydney, Australia
fang.chen@nicta.com.au
[3] Faculty of Engineering and IT, University of Technology, Sydney, Australia
wei.liu@uts.edu.au

Abstract. A discriminatory dataset refers to a dataset with undesirable correlation between sensitive attributes and the class label, which often leads to biased decision making in data analytics processes. This paper investigates how to build discrimination-aware models even when the available training set is intrinsically discriminating based on some sensitive attributes, such as race, gender or personal status. We propose a new classification method called Discrimination-Aware Association Rule classifier (DAAR), which integrates a new discrimination-aware measure and an association rule mining algorithm. We evaluate the performance of DAAR on three real datasets from different domains and compare it with two non-discrimination-aware classifiers (a standard association rule classification algorithm and the state-of-the-art association rule algorithm SPARCCC), and also with a recently proposed discrimination-aware decision tree method. The results show that DAAR is able to effectively filter out the discriminatory rules and decrease the discrimination on all datasets with insignificant impact on the predictive accuracy.

Keywords: Discrimination-aware data mining · Association rule classification · Unbiased decision making

1 Introduction

The rapid advances in data mining have facilitated the collection of a large amount of data and its uses for decision making in various applications. Although automatic data processing increases efficiency, it can bring potential ethical risks to users, such as discrimination and invasion of privacy. This paper focuses on building discrimination-aware classification models to eliminate potential bias against sensitive attributes such as gender and race.

Discrimination refers to the prejudicial treatment of individuals based on their actual or perceived affiliation to a group or class. People in the discriminated group are unfairly excluded from benefits or opportunities such as employment, salary or

S. Madria and T. Hara (Eds.): DaWaK 2015, LNCS 9263, pp. 108–120, 2015.
DOI: 10.1007/978-3-319-22729-0_9

education, which are open to other groups [1]. In order to reduce the unfair treatment, there are anti-discrimination legislations in different countries such as the Equal Pay Act of 1963 and the Fair Housing Act of 1968 in the US, and the Sex Discrimination Act 1975 in the UK [1]. Therefore, it is imperative to consider eliminating discrimination in applications such as decision support systems, otherwise the companies might be sued or penalized for acting against the law.

Our problem can be formally stated as follows. Suppose we are given a labeled dataset D with N instances, m nominal attributes $\{A_1, A_2,..., A_m\}$, from which the attribute $S = \{s_1, s_2,...,s_p\}$ has been identified as a sensitive attribute (e.g. race, gender, etc.), and a class attribute C. D is a *discriminatory dataset*, if there is an undesirable correlation between the sensitive attribute S and the class attribute C. For example, when performing credit history checks, if the probability P(*credit history = good | race = white*) is much higher than P(*credit history = good | race = black*), it is said that this dataset is biased against black people. The *discrimination severity of a classifier* is measured by computing the *discrimination score DS* (see Sect. 3.2) defined as:

$$DS = |P(C = C_t | S = S_1) - P(C = C_t | S = S_2)|, \text{ if } S \text{ is a binary nominal attribute,}$$

$$DS = \frac{1}{m} * \left(\sum_{i=1}^{m} \left|P\left(C = C_t | S = S_i\right) - P\left(C = C_t | S = S_{others}\right)\right|\right), \text{ if } S \text{ is a multi-value}$$

nominal attribute. This score is computed on the testing set using the predicted class labels. The goal is to learn a classifier with low discrimination score with respect to S, with minimal impact on the classification accuracy.

As an example assume that we are designing a recruitment system for a company to predict if a new candidate is suitable for a job or not. If the historical data contains more males than females, the prediction model may tend to favor the attribute gender. A prediction rule using gender or sensitive attribute like marital status, may achieve high accuracy, but it is not acceptable as it is discriminating, which is both unethical and against the law. Sensitive attributes such as gender, race and religion should be taken as an information carrier of a dataset, instead of distinguishing factors [2]. Females may be less suitable for a given job as on average they might have less work experience or lower educational level. It is acceptable to use work experience and educational level in the prediction model.

In this paper we investigate discrimination-aware classification that aims to decrease the discrimination severity for sensitive attributes, when the training data contains intrinsic discrimination. Our proposed method DAAR improves the traditional association rule classifier by removing the discriminatory rules while maintaining similar accuracy. DAAR also keeps the sensitive attributes during the classifier training of the classifier, which avoids information loss. Our contributions can be summarized as follows:

- We illustrate the discrimination problem and the importance of minimizing discrimination in real world applications.
- We propose a new measure, called Discrimination Correlation Indicator (DCI), which examines the discrimination severity of an association rule. DCI is applied as an effective criterion to rank and select useful rules in discrimination-aware association rule classification tasks.

- We extend the standard definition of the Discrimination Score measure (DS) from binary to multi-level nominal sensitive attributes.
- We propose DAAR, a new Discrimination-Aware Association Rule classification algorithm. We evaluate DAAR on three real datasets from different domains: traffic incident management, assessment of credit card holders and census income. We compare its performance with three methods: the standard association rule classifier [3], the state-of-the-art association rule classifier SPARCCC [4] and a discrimination-aware decision tree [5].

2 Related Work

2.1 Discrimination-Aware Methods

The discrimination-aware classification problem was introduced by Kamiran and Calders [6] and Pedreshi et al. [1], who formulated the direct and indirect discrimination definitions, and raised the attention of the Data Mining community to this problem. The existing discrimination-aware methods can be classified into two groups: methods that modify the dataset and methods that modify the algorithm.

The first group focuses on modifying the dataset during the pre-processing phase to eliminate the discrimination at the beginning. This includes removing the sensitive attribute, resampling [7] or relabeling some instances in the dataset to balance class labels for a certain sensitive attribute value [6, 8]. These methods typically lead to loss of important and useful information and undermine the quality of the predictive model that is learnt from the modified dataset. Additionally, just removing the sensitive attribute doesn't help due to the so called red-lining effect - the prediction model will still discriminate indirectly through other attributes that are highly correlated with the sensitive attribute [1, 9, 10].

The second group includes methods that integrate discrimination-aware mechanisms when building the classifier. Previous work [2, 5, 11] have adapted various widely used classification algorithms, including decision tree, naïve Bayes and support vector machine to deal with potential discrimination issues. The Discrimination-Aware Decision Tree (DADT) [5] uses a new splitting criterion, IGS, which is the information gain relative to the sensitive attribute, together with the standard information gain which is relative to the class (IGC). After generating a preliminary tree, the leaves are relabeled to decrease the discrimination severity to less than a non-discriminatory constraint $\varepsilon \in [0, 1]$ while losing as little accuracy as possible.

2.2 Association Rule Methods

An association rule takes the form $X \rightarrow Y$, where X and Y are disjoint item sets. X contains the set of antecedents of the rule, and Y is the consequent of the rule [12]. Given a dataset containing N instances and an association rule $X \rightarrow Y$, the support and confidence of this rule are defined as follows:

$$\text{Support } (X \rightarrow Y) = \sigma(X \cup Y)/N, \text{Confidence } (X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$$

where $\sigma(\cdot)$ is the frequency of an item set $(\cdot)$. When learning association rules, we are interested in rules with high support and high confidence.

Firstly introduced in [3], Classification Based on Association rules (CBA) uses association analysis to solve classification problems. In CBA, only class attributes can appear in *Y*. When classifying a new instance, if there are multiple matching rules, the rule with the highest confidence will be used to determine the class label. This method will also be referred as "standard AR" in the rest of the paper. Our proposed method is based on CBA, as the rule-based classifier can produce easy-to-interpret models.

SPARCCC is a relative new variation of CBA, which adds a statistical test to discover rules positively associated with the class in imbalanced datasets [4]. SPARCCC introduced the use of p-value and Class Correlation Ratio (CCR) in the rule pruning and ranking. CCR is defined as:

$$\text{CCR}(X \rightarrow y) = \text{corr}(X \rightarrow y)/\text{corr}(X \rightarrow \neg y)$$

$$\text{corr}(X \rightarrow y) = (\sigma(X \cup y) * N)/(\sigma(X) * \sigma(y))$$

where $\sigma(\bullet)$ is the frequency of an item set $(\bullet)$. The method retains rules with $corr\ (X \rightarrow y) > 1$ and $CCR > 1$, which condition guarantees that they are statistically significant in the positive associative direction $X \rightarrow y$, rather than in the opposite direction $X \rightarrow \neg y$. SPARCCC has been shown to significantly outperform CCCS [4].

3 The Proposed Method DAAR

Our proposed method DAAR uses the new measure Discrimination Correlation Indicator (DCI), together with confidence and support, to efficiently select representative and non-discriminatory rules that can be used to classify new instances.

DAAR offers the following advantages: (1) unlike naïve methods which simply remove the sensitive attribute to deal with discrimination, DAAR keeps the sensitive attribute in the model construction to avoid losing useful information; (2) the new measure DCI is easy to compute and capable of filtering out discriminatory rules with minimal impact on the predictive accuracy; (3) DAAR generates a smaller set of rules than the standard AR and these discrimination-free rules are easy to use by the users.

3.1 DCI Measure

DCI is designed to measure the degree of discrimination for *each rule*. Given the rule $X \rightarrow y$, DCI is defined as

$$\text{DCI} = \begin{cases} \frac{\left|P(C=y|S=S_{\text{rule}}) - P(C=y|S=S_{\text{others}})\right|}{\left(P(C=y|S=S_{\text{rule}}) + P(C=y|S=S_{\text{others}})\right)} \\ 0 \qquad \textit{if either of the above } P(\cdot) \textit{ is } 0 \end{cases}$$

$P\left(C = y|S = S_{\text{rule}}\right)$ is the probability of the class to be *y* given the value of the sensitive attribute *S* is S_{rule}. If *S* is a binary attribute, S_{rule} is the value of *S* in the target rule and

S_{others} is the other value of S. If S is a multi-value nominal attribute, S_{others} includes the set of all attribute values except the one which appears in the target rule.

For example, if the target rule is "gender = female, housing = rent → assessment = bad", where gender is the sensitive attribute, then S_{rule} is female and S_{others} is male. The DCI for this rule will be:

$$\text{DCI} = \frac{|\text{P}(\text{C} = \text{low}|\text{gender} = \text{female}) - \text{P}(\text{C} = \text{low}|\text{gender} = \text{male})|}{\text{P}(\text{C} = \text{low}|\text{gender} = \text{female}) + \text{P}(\text{C} = \text{low}|\text{gender} = \text{male})}$$

If the sensitive attribute does not appear in that rule at all, we define DCI to be 0. Therefore, the range of DCI is [0, 1). When DCI equals to 0, which means the probability of the class value to be y is the same given different sensitive attribute values, the rule is considered to be free of discrimination. Otherwise, DCI is *monotonically increasing* with the discriminatory severity of a rule, which means that the larger DCI is, the more discriminatory the rule is with regard to the sensitive attribute S.

3.2 Discrimination Score

The Discrimination Score measure (DS) has been used in previous research [2, 5, 11] to *evaluate the discrimination severity of a classifier*. The conventional definition is only for the binary sensitive attribute case. If the sensitive attribute S is binary with values S_1 and S_2, DS is defined as:

$$\text{DS} = \left|\text{P}\left(\text{C} = \text{C}_{\text{target}}|\text{S} = \text{S}_1\right) - \text{P}\left(\text{C} = \text{C}_{\text{target}}|\text{S} = \text{S}_2\right)\right|$$

DS computes the difference between the probabilities of the target class C_{target} given the two values of the sensitive attribute $S = S_1$ or $S = S_2$, on the testing dataset. C_{target} can be any attribute value of the class label.

We extend this definition for the case with multi-value nominal attribute with m values, $m > 2$. We propose that DS is computed for each S_{value} and then averaged over the m scores. For each computation, it takes S_{value} as S_i and all the other values as S_{others}, and is defined as follows:

$$\text{DS} = 1/\text{m}^* \left(\sum_{\text{i=1}}^{\text{m}} \left|\text{P}\left(\text{C} = \text{C}_{\text{target}}|\text{S} = \text{S}_{\text{i}}\right) - \text{P}\left(\text{C} = \text{C}_{\text{target}}|\text{S} = \text{S}_{\text{others}}\right)\right| \right)$$

The best case is when DS is zero, which means that the probabilities of the class value to be C_{target}, for all different values of the sensitive attribute, are the same, i.e. there is no discrimination. Otherwise, higher DS corresponds to higher discrimination severity. As the testing dataset has been labeled by the classifier, higher discrimination in the dataset indicates the classifier is biased, which should be prevented.

The purpose of DS and DCI is different. We note that DS cannot be used to filter discriminatory rules in DAAR instead of DCI, as DS is not for a single rule as required by DAAR. More specifically, DS is designed to measure the quality of a

classifier (any classifier, not only AR) based on a testing dataset that has been labeled by the classifier. In contrast, DCI is computed for each rule (hence, it requires a rule-based classifier) and then compared against a threshold to check whether the rule is discriminatory or not. Another difference between DCI and DS is that DCI is a single ratio, as there is only one possible attribute value of S in one rule, so S_{rule} and S_{others} are fixed once we know the target rule. On the other hand, DS is the average score over m sub-scores, as it computes a sub score for each possible attribute value S_{value} in the dataset, which will be more than one for non-binary attribute S.

3.3 DAAR Algorithm

DAAR integrates DCI and the association rule classification to select discrimination-aware rules from all rules that have passed the minimum confidence and support thresholds. DAAR's algorithm is shown in Table 1.

Table 1. DAAR Algorithm

Algorithm Build Discrimination-Aware Classifier	
Input	dataset *D* with sensitive attribute *S*;
	max_length = *k*; thresholds *conf*, *spt* and *dci*;
Output	non-discriminatory rules
1	**for** i = 2 to k // generate *i*-item rules
2	**if** i = 2: generate 2-item rule, which is the base case;
3	**else** : merge (*i*-1)-item RuleSet and 2-item RuleSet;
4	prune rules to keep anti-monotonic;
5	**end if**
6	set up contingency tables to calculate confidence, support, DCI;
7	filter rules using thresholds *conf*, *spt* and *dci*;
8	store rules in *i*-item RuleSet;
9	**end for**
10	sort rules in *k*-item RuleSet by DCI in ascending order;
11	**return** *k*-item RuleSet as the classifier;

The algorithm defines the maximum length of the rule as k in the input, so as a result, all rules will contain at most $k - 1$ antecedents on the left and one class label on the right. In the loop, the algorithm merges the $(i–1)$-item rule set which was generated in the last round with the 2-item rule set (the base case), to get the i-item rule set. In line 10, the set of rules is sorted by DCI in ascending order for clear presentation to users, and this sorting does not affect the classification results. The majority voting is then used to classify new instances; if the vote is tied (e.g. the same numbers of rules support each class), the sum of DCI of all rules supporting each class is calculated and compared to determine the final class. As discussed in Sect. 3.1, the severity of discrimination is lower when DCI is smaller; therefore the voting will select the class value with lower sum value (i.e. which is less discriminatory).

4 Datasets and Experimental Setup

Three real datasets from different domains, such as public transport and finance management, are used to evaluate the performance of DAAR.

Traffic Incident Data was collected by the road authority of a major Australian city.[1] Each instance has 10 attributes, including the time, location and severity of the incident, the manager and other useful information. The *incident manager* is selected as the sensitive attribute S, and the class label is the *duration of the incident*, which takes two values: *long* and *short*. Our task is to predict whether an incident would be difficult to manage based on the available information. The *incident duration* is considered as a proxy to incident difficulty level – an incident with long duration corresponds to a difficult-to-manage incident and an incident with short duration corresponds to an easy-to-manage incident. Our experimental evaluation tests whether the proposed method can reduce the discrimination based on the sensitive attribute *incident manager* in predicting the *difficulty level of the incident*.

The data was preprocessed in two steps. Firstly, we noticed that there were more than 90 distinct manager values appearing in the full dataset, but most of them were only associated to less than 10 incidents. For simplicity, the managers were sorted by the number of associated incidents, and only instances handled by the top 5 managers were used in the experiment. Then, the majority class in the dataset was under-sampled to keep the dataset balanced with respect to the class attribute. This resulted in a dataset of 4,920 incidents, half of which were *difficult* and the other half were *easy* to manage.

German Credit Card Data is a public dataset from the UCI Machine Learning repository [13]. The dataset consists of 1,000 examples (700 *good* and 300 *bad* customers), described by 20 attributes (7 numerical and 13 categorical). The sensitive attribute is the *personal status and sex*, which shows the gender of a customer and whether he or she is single, married or divorced. Since it can be discriminatory to assess customers by their gender and marital status, we would like to decrease the discrimination based on this attribute when classifying new customers.

As the original dataset is strongly biased towards the class *good*, with ratio *good*:*bad* = 7:3, we randomly removed 400 good customers to keep the balance of the dataset. This resulted in 600 examples, 300 from each of the two classes.

Census Income Data is also a public dataset from the UCI Machine Learning repository. It contains 40 attributes (7 numerical and 33 categorical), which are used to predict the income level of a person. If the income is over $50 K, the person is classified as having *high income*, otherwise as having *low income*. The attribute *race* (with values: *white*, *black*, *asian or pacific islander*, *amer indian aleut or eskimo* and *other*) is the sensitive attribute. We randomly selected a smaller portion of the original dataset containing 1,200 examples; half with *high income* and half with *low income*.

[1] Data Collected by NSW Live Traffic: https://www.livetraffic.com/desktop.html#dev.

We evaluate the performance of DAAR and the other three methods in terms of predictive accuracy and discrimination score. The three baseline methods are CBA (the standard AR), SPARCCC and DADT. CBA was chosen as it is a standard association rule classifier. SPARCCC was selected as it is a state-of-the-art association rule classifier for imbalanced datasets. The discriminatory dataset can be considered as a special type of imbalanced dataset. In a discriminatory database, the bias is against a certain class label within a group, having the same value for a sensitive attribute, e.g. *race* = *black*, while in an imbalanced dataset the bias is against a class over the whole dataset. DADT was selected as it is a successful discrimination-aware classifier.

All three association rule mining methods (standard AR, DAAR and SPARCCC) use confidence and support thresholds to remove the uninteresting rules. These thresholds are controlled as a baseline, while the other measures, the CCR threshold in SPARCCC and the DCI threshold in DAAR, are varied to generate comparison conditions. For example, for the traffic data, we used the following pairs of confidence and support values (*conf* = 0.6, *spt* = 0.01; *conf* = 0.6, *spt* = 0.03; *conf* = 0.6, spt = 0.05; *conf* = 0.5, *spt* = 0.1).

The number of rules generated by the classifier is another important factor to consider. It affects both the accuracy and discrimination score, and is also very sensitive to the chosen thresholds for confidence, support, CCR and DCI. In order to compare the results fairly, it is important to make sure that the number of rules of comparable conditions are in the same range. Hence, once the thresholds for confidence and support are fixed for the standard AR, the thresholds for CCR and DCI (between [0,1)) are configured such that SPARCCC and DAAR can generate 4-5 conditions where the number of rules is of the same order of magnitude.

DADT, the discrimination-aware decision tree [5], uses the addition of the accuracy gain and the discrimination gain, IGC + IGS, as a splitting criterion, and relabeling of some of the tree nodes to reduce the discrimination. The non-discriminatory constraint $\varepsilon \in [0, 1]$ is tuned to generate comparison conditions.

5 Results and Discussion

All reported results are average values from 10-fold cross validation. The *p* value is the result of the independent two-sample *t*-test, which is used to statistically compare the differences in performance between DAAR and the other methods.

Traffic Incident Data. The sensitive attribute for this dataset is the *incident manager*, and the class label is the *incident difficulty level*. Table 2 shows the average predictive accuracy and discrimination score of the proposed DAAR method and the three methods used for comparison (standard AR, SPARCCC and DADT).

Table 2 shows that overall, considering both accuracy and discrimination, DAAR is the best performing algorithm – it has the second highest accuracy and the second lowest discrimination score, and also small standard deviations. DAAR is statistically significantly more accurate than SPARCCC ($p = 0.041$) and slightly more accurate than DADT ($p > 0.05$). Although AR is more accurate than DAAR, this difference is not statistically significant. In terms of discrimination score, DAAR has

Table 2. Accuracy and Discrimination Score for Traffic Incident Data

Methods	Accuracy		Discrimination Score	
	mean	std	mean	std
Standard AR	77.22%	0.021	0.236	0.010
SPARCCC	73.21%	0.059	0.266	0.064
DADT	74.19%	0.045	0.187	0.022
DAAR	76.32%	0.025	0.213	0.012

significantly lower discrimination score than both the standard AR ($p = 0.009$) and SPARCCC ($p = 0.0016$). DADT has the lowest discrimination score but its accuracy is impacted – it has the second lowest accuracy after SPARCCC. SPARCCC is the worst performing method – it has the lowest accuracy and highest discrimination score, and the largest variation for both measures.

Figure 1 presents a scatter plot of the accuracy and discrimination score. Ideally, we would like to see points in the top-left corner of the figure, which corresponds to high accuracy and low discrimination score simultaneously. However, there is a trade-off between the two measures, as the filtering out of discriminating rules will normally lower the predictive accuracy. Given this trade-off, our aim is to select the method which has lower discrimination score but no significant impact on accuracy.

The results in Fig. 1 are consistent with the results in Table 2. All individual results of DAAR (triangles) are clustered in the left part of the graph which corresponds to low discrimination score, and these scores are always lower than the results for the standard AR (diamonds). SPARCCC is more diverse – it has points with relatively low and very large discrimination score and some others with a large discrimination score, which explains the overall low discrimination score and the large standard deviation. DADT's results (circles) are also in the left part of the graph (low discrimination score) but the accuracy varies, which explains the overall lower accuracy and its higher standard deviation.

Table 3 presents some of the rules generated by the DAAR, together with their confidence and DCI values. These rules are easy to understand by users, which is one of the advantages of applying association rule classification.

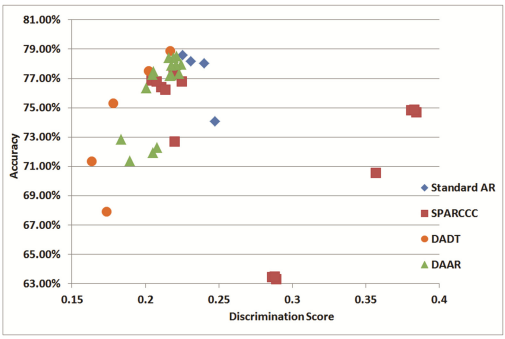

Fig. 1. Scatter plot for traffic incident data

Table 3. Examples of Rules Generate by DAAR for Traffic Incident Data

Rules	Confidence	DCI
type=road development, time = [20:00,23:59] → difficult incident	1	0
severity=2, time =[20:00,23:59], direction= north/south → difficult incident	0.95	0
day_of_week=7, direction = north /south → difficult incident	0.90	0
location= Cahill Expressway, Sydney → easy incident	0.76	0

Table 4 shows examples of rules that discriminate based on the manager with high confidence. These rules were filtered out by the proposed DAAR method.

Table 4. Examples of Discriminating Rules Filtered Out by DAAR

Rules	Confidence	DCI
manager=Frank, incident_severity=2 → difficult incident	0.87	0.273
manager=Charles →difficult incident	0.75	0.245
manager=Henry → easy incident	0.76	0.259

German Credit Card Data. For this data set, the aim is to eliminate the discrimination on *personal status and sex* when determining whether a customer is *good* or *bad*.

The results are shown in Table 5. We can see that the standard AR is the most accurate method, followed by SPARCCC, DAAR and DADT. The statistical testing results show that the accuracy of DAAR is not significantly lower than SPARCCC ($p > 0.05$) and that it is significantly higher than DADT ($p = 5.5e\text{-}5$). The accuracy range on this dataset (60-68 %) is lower compared to the traffic dataset (73-77 %). This might be due to the small size of the credit card dataset and the large number of its attributes. In terms of discrimination score, DADT is the best performing algorithm, followed by DAAR, SPARCCC and the standard AR. The t-test results show that our method DAAR has statistically significantly lower discrimination score than both the standard AR ($p = 8.0e\text{-}6$) and SPARCCC ($p = 0.0004$). Again, we can see that DAAR provides a good balance in terms of accuracy and discrimination score.

Table 5. Accuracy and Discrimination Score for German Credit Card Data

Methods	Accuracy		Discrimination Score	
	mean	std	mean	std
Standard AR	68.40%	0.012	0.329	0.006
SPARCCC	66.60%	0.029	0.305	0.071
DADT	60.13%	0.008	0.157	0.007
DAAR	64.90%	0.027	0.208	0.055

The scatter plot in Fig. 2 illustrates clearly the trade-off between accuracy and discrimination score. We can see that all DAAR's points (triangles) are on the left with respect to the standard AR points, but the accuracy of these points is lower than the accuracy of the standard AR points due to the removal of the discriminating rules. It is also interesting to observe that although SPARCCC has higher average accuracy than DAAR, the scatter plot demonstrates that for the same discrimination score (between 0.15 and 0.3), SPARCCC has lower accuracy than DAAR. The SPARCCC points are grouped into two main clusters: one in the middle that has similar discrimination score as DAAR and three points at the right corner that have high accuracy but large discrimination scores. DADT points are clustered at the bottom left of the graph, so that its accuracy lower than DAAR's for similar discrimination scores.

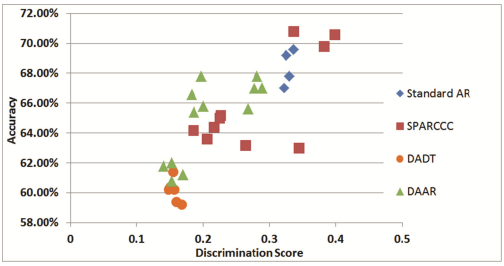

Fig. 2. Scatter Plot for German Credit Card Data

Census Income Data. The sensitive attribute in this dataset is *race* and the class label is *income level*. The aim is to avoid predicting the income level (*high* or *low*) of a person based on their race.

The results are presented in Table 6. We can see that in terms of average accuracy, all methods except DADT perform very similarly achieving accuracy of about 79-81 %, which is higher than the accuracy on the previous two datasets. SPARCCC is slightly more accurate and DAAR is slightly less accurate. The standard deviations of all three association rule methods are about 1 %. As to discrimination score, DADT again is the best performing method, followed by DAAR, which is consistent with the previous results. The standard AR comes next and the worst performing algorithm is SPARCCC. The t-test shows that the discrimination score of DAAR is significantly lower than both the standard AR ($p = 0.016$) and SPARCCC ($p = 1.12e\text{-}6$).

Table 6. Accuracy and discrimination score for Census Income Data

Methods	Accuracy		Discrimination Score	
	mean	std	mean	std
Standard AR	80.81%	0.011	0.285	0.009
SPARCCC	81.20%	0.010	0.289	0.010
DADT	68.57%	0.132	0.197	0.077
DAAR	79.65%	0.011	0.265	0.007

6 Conclusions

In this paper, we proposed DAAR, a discrimination-aware association rule classification algorithm that provides unbiased decision making support in data analytics processes. We have shown that DAAR is able to address the discrimination issues occurred on sensitive attributes, while having a minimal impact on the classification accuracy. DAAR uses DCI, a new discrimination measure, to prune rules that discriminate based on sensitive attributes, such as race and gender. The rules that pass the confidence-support-DCI filter will form the final DAAR rule set. To classify new instances, DAAR uses majority voting and a sum of DCI scores.

We empirically evaluated the performance of DAAR on three real datasets from traffic management and finance domains, and compared it with two non-discrimination-aware methods (a standard AR classifier and the state-of-the-art AR classifier SPARCCC), and also with the discrimination-aware decision tree DADT. The experimental results on all datasets consistently showed that DAAR is capable of providing a good trade-off between discrimination score and accuracy – it obtained low discrimination score while its accuracy was comparable with AR and SPARCCC, and higher than DADT. An additional advantage of DAAR is that it generates a smaller set of rules than the standard AR; these rules are easy to use by the users, in helping them make discrimination-free decisions.

Future work will include integrating DAAR in decision support applications such as assessment of social benefits. From a theoretical perspective, we plan to investigate the case with multiple sensitive attributes and the use of DCI in ensemble classifiers.

References

1. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560–568. ACM (2008)
2. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**, 277–292 (2010)
3. Ma, Y., Liu, B., Yiming, W.H.: Integrating classification and association rule mining. In: Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1998), pp. 80–86 (1998)
4. Verhein, F., Chawla, S.: Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In: Proceedings of the 7th IEEE International Conference on Data Mining, pp. 679–684. IEEE (2007)
5. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Proceedings of the 10th IEEE International Conference on Data Mining, pp. 869–874. IEEE (2010)
6. Kamiran, F., Calders, T.: Classifying without discriminating. In: International Conference on Computer, Control and Communication, pp. 1–6. IEEE (2009)
7. Kamiran, F., Calders, T.: Classification with no discrimination by preferential sampling. In: Proceedings of the Benelearn (2010)
8. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: IEEE International Conference on Data Mining Workshops, pp. 13–18. IEEE (2009)

9. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. **25**, 1445–1459 (2013)
10. Pedreschi, D., Ruggieri, S., Turini, F.: Integrating induction and deduction for finding evidence of discrimination. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 157–166. ACM, Barcelona, Spain (2009)
11. Ristanoski, G., Liu, W., Bailey, J.: Discrimination aware classification for imbalanced datasets. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1529–1532. ACM (2013)
12. Simon, G.J., Kumar, V., Li, P.W.: A simple statistical model and association rule filtering for classification. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 823–831. ACM, 2020550 (2011)
13. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml