Discrimination-aware classification for imbalanced datasets

Goce Ristanoski
The University of Melbourne
NICTA Victoria Laboratory
Melbourne. Australia
g.ristanoski@student.unimelb.edu.au

Wei Liu
The University of Melbourne
Melbourne, Australia
wei.liu@unimelb.edu.au

James Bailey
The University of Melbourne
NICTA Victoria Laboratory
Melbourne, Australia
baileyj@unimelb.edu.au

ABSTRACT

The problem of learning a discrimination aware model has recently received attention in the data mining community. Various methods and improved models have been proposed, with the main approach being the detection of a discrimination sensitive attribute. Once the discrimination sensitive attribute is identified, the methods aim to develop a strategy that will include the useful information from that attribute without causing any additional discrimination. Our work focuses on an aspect often overlooked in the discrimination aware classification - the scenario of an imbalanced dataset, where the number of samples from one class is disproportionate to the other. We also investigate a strategy that is directly minimizing discrimination and is independent of the class balance. Our empirical results indicate additional concerns that need to be considered when developing discrimination aware classifiers, and our proposed strategy shows promise in overcoming these concerns.

Categories and Subject Descriptors

H.3.1 [**Knowledge Management**]: Data mining theory, methods, and applications

Keywords

Discrimination aware classification, imbalanced datasets

1. INTRODUCTION

Machine learning models such as Decision Trees, Naive Bayes, Support Vector Machines and others, are used to provide unbiased optimization in the process of decision making. When the term unbiased is used to describe these methods, it refers to the methods learning from data in a way that is not influenced by the user. This, however, tells very little about the model gaining some form of bias due to the data itself. An example is a dataset that predicts whether or not a customer should be given loan in a bank, and a model prefers male customers with higher income over female customer with medium income. Though this might be a reasonable judgement, it tells us very little about the point where bias turns into discrimination.

The discrimination in datasets is generally referred to having the samples belong to two groups, and one of the groups having a greater ratio of positive vs. negative samples than the other group. A very common belief when it comes to the groups is that the group that is being discriminated against will also be less present in the number of samples. After all, if a group is largely present, it would not seem to be possible to have that group as a discriminated group. An aspect that is often overlooked is if we have imbalanced number of positive and negative samples, should we include this information in our training as well?

Our work provides three contributions in the analyses of discrimination aware classification:

- We consider imbalanced datasets as a special case of discrimination aware classification, and identify potential challenges other methods may encounter.
- We analyse the nature of discrimination and propose a strategy of direct minimization of the model aided discrimination that can be applied to imbalanced datasets.
- We consider the case of the discriminated group being more present in the dataset and the effects this might have on the learning process.

2. RELATED WORK

Discrimination aware classification has been studied in the machine learning community, and discrimination is often related to one or more attributes in the dataset - gender, job title, ethnicity etc. In the next section we will present the background of discrimination in model training, the most common ways of dealing with discrimination, and overlooked aspects of the discrimination aware learning process.

2.1 Presence of discrimination in models

Different forms of discrimination can be present in the training dataset: gender discrimination when applying for work, ethnic discrimination when requesting a loan, discrimination originating from social status and so on. Though it is not a primary concern for the machine learning community to label this discrimination as a key factor in training a model, it can be of great concerns to users: companies can get bad publicity if they unjustifiably prefer male over female job candidates; banks can be sued for racism if their decisions are largely based on ethnic components. Let us assume we have two groups of samples in the dataset, Group 1 and Group 2, and Group 2 is being marked as discriminated. We define a discrimination score as (number of samples in Group 1 with positive class/number of samples in Group 1) - (number of samples in Group 2 with positive class/number of samples in Group 2). If we build a model with all the samples included in the training, we would expect similar value for the discrimination score for the test dataset as we have for the training dataset: most of the positive samples to belong to Group 1, and some in Group 2. However, if we have a greater value for the discrimination score calculated on the predictions for the test set, we may infer there is some form of discrimination in the model.

2.2 Dealing with discrimination

As discrimination is mostly considered to be originating from a specific attribute, or several attributes, methods for dealing with discrimination mainly focus on this specific attribute. Some simple strategies are removing the discrimination sensitive attribute, removing samples from the dataset or swapping class labels so that there is no discrimination to begin with. This may result in valuable information being lost or additional noise. Removing the discrimination sensitive attribute does not necessarily mean removing discrimination: that attribute can be correlated with some of the other attributes, known as the red-lining effect.
The additional discrimination that a learned model will add leads to the definition of the total discrimination as

$$D_{total}=D_{explanatory}+D_{bad} \tag{1}$$

where D_{bad} is the additional discrimination added by the model[6], and $D_{explanatory}$ is the already existing preference to one of the groups, equal in value to the discrimination score. It is more convenient to attempt to minimize D_{bad} only. Advanced methods of dealing with discrimination include using discrimination aware information gain for training decision trees [3], fair classification regularizer component in the logistic regression loss [4], discrimination aware conditional probabilities for Naive Bayes [2].

3. DISCRIMINATION AT IMBALANCED DATASETS

Imbalanced datasets present a challenge on their own when it comes to building a model. In a dataset with 10% of the samples having positive class, we can easily learn a suboptimal model. In the next section we analyse how the imbalanced dataset will affect the discrimination aware model.

3.1 Imbalanced datasets groups distribution

We will refer to the positive class as the less present class in the dataset through the paper, though it can have any class labelling in general. When there is a high amount of presence of the positive class, both in percentage and in samples, discrimination can be considered using the standard approaches. An example is given in Figure 1.a: we have 25% of the samples with a positive class label, and we can observe the preference of the positive class towards the samples of the male group. If the number of positive samples is 15%, as shown in Figure 1.b, the situation is more alarming: there are barely any samples with positive class left in the female group, and the number of samples with positive class in the male group is reduced too.
There are two concerns regarding discrimination that we are interested in investigating: will the model produce any predictions for samples in the female groups, leading to a higher discrimination score than expected, and whether the model will not produce enough predictions with positive class for samples in the male group, leading to a lower discrimination score then expected. Our assumption is that both of these scenarios have a high likelihood of occurring, leading to poor performance in the attempt to handle discrimination.

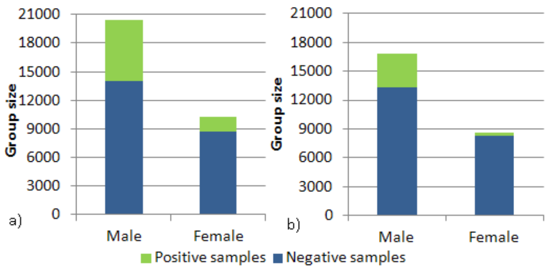

Figure 1: Positive and negative samples per gender group.

3.2 Discriminated group as a more present group

A situation overlooked in most of the discrimination aware learning literature is the case when we have the discriminated group being more present in the total number of samples. This means that only in the overall number of samples we have more samples belonging to the discriminated group, but in terms of samples with positive class, they are more present in the preferred group. Examples of such type of discrimination can be bank loans given to clients with higher education in a country where most of the population does not have higher education, or perhaps where members of an elite minority have more influence over the majority of the population belonging to a less privileged ethnic group. When we have the total number of positive samples being around 25% in this case, the distribution of samples will be as presented in Figure 2.a. If the total number of positive samples is 15% (Figure 2.b), we can see that both groups do not have too many positive samples.

3.3 Discrimination loss definition

Concerns that arise when applying discrimination aware models to imbalanced datasets originate from the fact that the number of positive samples may not be high enough. This may mean that:

- In the case of the discriminated group, due to the small number of positive samples, we will have fewer positive predictions for that group, leading to a greater discrimination score.
- In the case of the preferred group, also due to the small number of positive samples, we will not have many positive predictions, leading to a small discrimination score. This applies in particular when the preferred group is the less present group.

These concerns have their origin in the fact that many of the discrimination aware models use some form of conditional probabilities in the attempt to reduce D_{bad}. The state

Table 1: Discrimination aware loss definition. The standard empirical loss is still optimized, but with addition of the discrimination aware loss.

<table>
<tr><th>Function</th><th>Empirical Loss(EL)</th><th>E. Loss Derivative(DEL)</th><th>Discrimination Loss(DL)</th><th>D. loss derivative(DDL)</th></tr>
<tr><td>SVM</td><td>max(0, 1-$y_i(w^T x_i)$)</td><td>-y_i, $y_i(w^T x_i)$<1, else 0</td><td>0</td><td>0</td></tr>
<tr><td>SVMDisc</td><td>max(0, 1-$y_i(w^T x_i)$) ,</td><td>-y_i, $y_i(w^T x_i)$<1, else 0</td><td>max(0,$w^T x_i$),x_i in G_{pref}
-max(0,$w^T x_i$),x_i in G_{disc}</td><td>1, sign($w^T x_i$)=1, else 0
-1, sign($w^T x_i$)=1, else 0</td></tr>
<tr><td colspan="5">Standard empirical total loss $R_{emp}(\mathbf{w}) = F_1 = \sum_{i=1}^{n} EL(x_i, w, y_i)$, $DF_1 = \sum_{i=1}^{n} DEL(x_i, w, y_i)$</td></tr>
<tr><td colspan="5">Discrimination total loss $F_2 = n(\frac{\sum_{i=1}^{n} w^T x_i, sign(w^T x_i)=1,\ x_i\ in\ G_{pref}(=sum_1)}{\sum_{i=1}^{n} 1,\ x_i\ in\ G_{pref}}$ - $\frac{\sum_{i=1}^{n} w^T x_i, sign(w^T x_i)=1,\ x_i\ in\ G_{disc}}{\sum_{i=1}^{n} 1,\ x_i\ in\ G_{disc}}$
$-(\frac{\sum_{i=1}^{n} y_i, y_i=1,\ x_i\ in\ G_{pref}}{\sum_{i=1}^{n} 1,\ x_i\ in\ G_{pref}} - \frac{\sum_{i=1}^{n} y_i, y_i=1,\ x_i\ in\ G_{disc}}{\sum_{i=1}^{n} 1,\ x_i\ in})$)</td></tr>
<tr><td colspan="5">Discrimination aware empirical loss $R_{emp}^{disc}(\mathbf{w}) = \sqrt{\frac{F_1^2 + \lambda F_2^2}{1+\lambda}}$, $\partial_w R_{emp}^{disc}(\mathbf{w}) = \frac{1}{2}(\frac{F_1^2+\lambda F_2^2}{1+\lambda})^{-\frac{1}{2}}(\frac{2*(F_1*DF_1+\lambda F_2*DF_2)}{1+\lambda})$</td></tr>
</table>

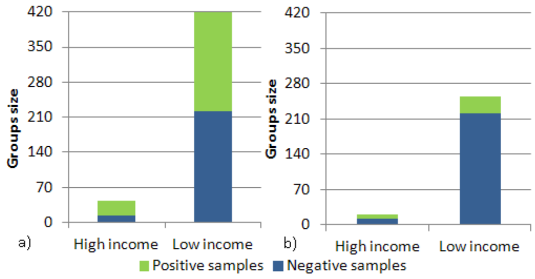

Figure 2: Positive and negative samples per income group. The group with high income is less present, but preferred for assigning loan that the group with low income.

of an imbalanced dataset may lead to calculations which use very small values for the conditional probabilities, resulting in an discrimination optimisation process with reduced influence. The paradigm of having a D_{bad} component is something we adopt as well, but instead of looking into an algorithm which results in D_{total} being minimized, we are choose a direct approach - optimizing D_{bad} itself. As D_{bad} can be considered as the difference between the produced discrimination score and actual discrimination score, we choose to include D_{bad} in the empirical risk function minimisation. The standard regularized risk function has the form of

$$L(\mathbf{w}^*) = argmin_{\mathbf{w}}\ \phi \mathbf{w}^T \mathbf{w} + R_{emp}(\mathbf{w}) \qquad (2)$$

with $\mathbf{w}$ as the weight vector, ϕ is a positive parameter that determines the influence of the structural error in Equation 2, and $R_{emp}(\mathbf{w}) = \sum_{i=1}^{n} l(\mathbf{x}_i, y_i, \mathbf{w})$ is the loss function with $l(\mathbf{x}_i, y_i, \mathbf{w})$ as a measure of the distance between a true label y_i and the predicted label from the forecasting done using $\mathbf{w}$ Our approach is to modify R_{emp} in a discrimination aware manner, and apply the new loss function to Support Vector Machine. If we define the two groups which we use for definition on the discrimination as G_{pref} and G_{disc}, then D_{bad} will be the difference between the predicted and actual discrimination score. We define this as discrimination total loss, and alongside the standard empirical loss we can derive a new discrimination aware empirical loss $R_{emp}^{disc}(\mathbf{w})$, as presented in Table 1. We choose to use a quadratic mean form, as it has already been successfully applied for balancing different losses [5], and it is solved by using the Bundle Method, a linear optimization method that uses the subgradients of the loss to iteratively update the w in a direction that minimizes the quadratic loss (Table 1). The new form of the regularized risk function will now be:

$$L^{disc}(\mathbf{w}^*) = argmin_{\mathbf{w}}\ \phi \mathbf{w}^T \mathbf{w} + R_{emp}^{disc}(\mathbf{w}) \qquad (3)$$

4. EXPERIMENTS AND RESULTS

We conducted experiments on two datasets commonly used in discrimination aware learning: *Adult* and *Census-Income (KDD)* dataset. We worked with a randomly chosen 17% of the *Census-Income (KDD)* due to its size. The discriminated group is the female group in the gender attribute in both cases, and this group is less present in general. We also tested with the *Statlog (German Credit Data)*, where we edited the saving attribution into binary attribute describing if a person has more or less than 1000 former German Marks on the account. All of the original files of the dataset can be found on [1].

We compare our Support Vector Machine implementation of the discrimination aware loss (or SVMDisc), with standard SVM, and advanced discrimination aware Naive Bayes(NB) and J48[6]. The evaluation metrics we use are the F-measure and Discrimination score distance, which is the absolute difference between the expected discrimination score and the achieved discrimination score.

From the testing with the *Adult* and *Census-Income (KDD)* datasets (Table 2), where the discriminated group is less present group as well, we can observe the results of the original datasets (number of positive class samples is 25%) and the imbalanced dataset (number of positive class samples is down to 15%). We can see that in the case of original datasets all the methods produce small distribution score distances. For the *Adult* datatset, when imbalanced, our method produced Discrimination score distance higher than J48 and Naive Bayes, but also produced greater F-measure value. In the case of imbalanced *Census-Income (KDD)*, all the methods perform well and produced small distribution score distances.

When tested on the *Statlog (German Credit Data)* dataset (Table 3), we can observe our method showing more clearer improvements over the discrimination aware J48 and Naive

Bayes. Though Naive Bayes performed well with the original dataset, J48 produced high distribution score distance. When we tested on the imbalanced dataset, both J48 and Naive Bayes produced very high distribution score distances and small F-measures, while SVMDisc produced higher F-measure and significantly smaller distribution score distance. Overall, as we imbalance the dataset, J48 and Naive Bayes have the discrimination score distance increase more rapidly than the one for SVMDisc, while all methods have the AUC value drop severely, as presented in Figure 3, when tested on the *Statlog (German Credit Data)* dataset.

Table 2: Performance of SVMDisc, standard SVM and discrimination aware J48 and Naive Bayes(NB) when the discriminated group is less present groups. We can observe that the F-score had a large drop in the case of the *Adult* dataset for J48 and Naive Bayes(NB), while it was less of a drop for SVMDisc.

Adult dataset, expected D-score=0.2				
Standard dataset				
Method	**Prec.**	**Recall**	**F-meas.**	**D.s. dist.**
SVMDisc	0.3421	0.7322	0.4663	0.01
SVM	0.7454	0.5254	0.6164	0.033
J48	0.7528	0.6132	0.6759	0.02
NB	0.6615	0.5822	0.6193	0.012
Imbalanced dataset				
Method	**Prec.**	**Recall**	**F-meas.**	**D.s. dist.**
SVMDisc	0.4101	0.3032	0.3487	0.14
SVM	0.8669	0.2324	0.3666	0.15
J48	0.1967	0.0764	0.1100	0.09
NB	0.1978	0.0966	0.1298	0.05
Census-Income (KDD) dataset, expected D-score=0.2				
Standard dataset				
Method	**Prec.**	**Recall**	**F-meas.**	**D.s. dist.**
SVMDisc	0.6624	0.6720	0.6671	0.01
SVM	0.7478	0.6468	0.6937	0.068
J48	0.7537	0.7424	0.7480	0.044
NB	0.7043	0.6845	0.6943	0.04
Imbalanced dataset				
Method	**Prec.**	**Recall**	**F-meas.**	**D.s. dist.**
SVMDisc	0.5947	0.6257	0.6098	0.01
SVM	0.7655	0.4877	0.5958	0.04
J48	0.7889	0.5617	0.6562	0.02
NB	0.6702	0.5694	0.6157	0.0

5. CONCLUSION

Discrimination aware learning is a rising research area in the data mining community, and several discrimination aware approaches have been suggested. In this paper we introduce two new concepts overlooked in the discrimination aware learning literature: the scenario of the discriminated group being the more present group in the dataset, and the case of an imbalanced dataset. We analyse the effects these two aspects may have on the discrimination aware learning, and suggest a direct discrimination score optimisation that aims to be less sensitive to imbalanced datasets.
Our experiments suggest that direct discrimination score optimisation technique has potential when tested on imbalanced datasets, especially in the case when the discriminated group is more present. We extended SVM to design this optimisation method, but our approach can be applied to other machine learning methods based on iterative learning.

Table 3: Performance of SVMDisc, standard SVM, discrimination aware J48 and Naive Bayes(NB). The discriminated group is more present group.

Statlog (German Credit Data) dataset, D-score=0.22				
Imbalanced dataset				
Method	**Prec.**	**Recall**	**F-meas.**	**D.s. dist.**
SVMDisc	0.6897	0.3922	0.5000	0.01
SVM	0.6800	0.6667	0.6733	0.11
J48	0.6222	0.5490	0.5833	0.17
NB	0.7200	0.7059	0.7129	0.0
Imbalanced dataset				
SVMDisc	0.1613	0.5000	0.2439	0.083
SVM	0.5000	0.1000	0.1667	0.04
J48	0.0000	0.0000	0.0000	0.2
NB	1.0000	0.1000	0.1818	0.21

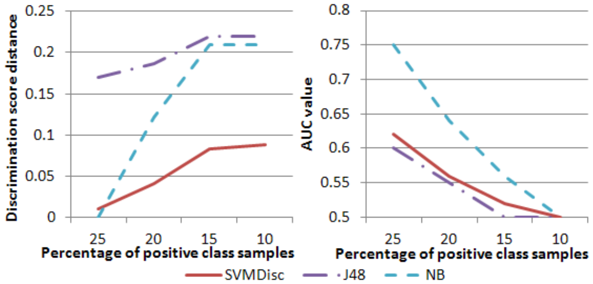

Figure 3: Discrimination score distance and AUC value as we imbalance the *Statlog (German Credit Data)* dataset. The SVMDisc Discrimination score distance increases less intensely than the one for J48 and NB, and all of the methods have the AUC score drop significantly as the dataset gets imbalanced.

6. REFERENCES

[1] http://archive.ics.uci.edu/ml/datasets.html, 2013.

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

[3] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *Proceeding of IEEE 10th International Conference on Data Mining (ICDM)*, pages 869–874. IEEE, 2010.

[4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer Berlin Heidelberg, 2012.

[5] W. Liu and S. Chawla. A quadratic mean based supervised learning model for managing data skewness. In *In Proceedings of the Eleventh SIAM International Conference on Data Mining*, pages 188–198, 2011.

[6] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *IEEE 11th International Conference on Data Mining (ICDM)*, pages 992–1001. IEEE, 2011.