Report on experiments with methods introduced in

Discrimination Aware Decision Tree Learning

by Faisal Kamiran, Toon Calders and Mykola Pechenizkiy

Nigul Olspert
nigul.olspert@aalto.fi

October 16, 2015

1 Overview of the methods

The aim of the paper by Kamiran et al. was to reduce a discrimination of a binary classifier w.r.t. given sensitive attribute. For this purpose a method of leaf relabeling in decision tree was introduced. Usually when a new example needs to be classified, it is given the majority class label of the region it falls into. This strategy was not changed by the method, however instead, the idea was to modify the class labels of the leaves so that the discrimination would reduce while trading in as little accuracy as possible. The general algorithm called RELAB was proposed as a suitable implementation for the method and equivalence of it to the well known KNAPSACK problem was stated. The results were compared to the methods mostly based on data preprocessing (massaging and reweighing), but also to a discrimination-free naive Bayes classifier. Establishing the benefit of the leaf relabeling technique the so called baseline method was used. The latter one consists of gradually modifying the initial dataset so that the attributes most correlating with the sensitive one would be removed and subsequently "ordinary" decision trees be constructed.

Another approach introduced in the paper was a generalization of the information gain used in splitting the nodes (IGC-IGS, IGC/IGS). As it did not lead to decrease in discrimination the results from this study were omitted.

2 Reproducing the results

In the scope of given report we implemented a decision tree with leaf relabeling and applied it on Census Income Data. This dataset consists of 32561 samples with a class attribute as gross income separated into two ranges with threshold 50000. The sensitive attribute was taken to be gender. We implemented the algorithms in programming language R and the code can be accsessed online by following the this link.

The accuracy of a decision tree is affected by multiple factors, e.g. by splitting criteria and stopping criteria. Mainly due to the limitations in computational capabilities, we set the following 2 constraints in our decision tree construction algorithm:

- Splitting based on continuous attribute is done exactly at the mean value of the given attribute calculated for the data samples left in the node

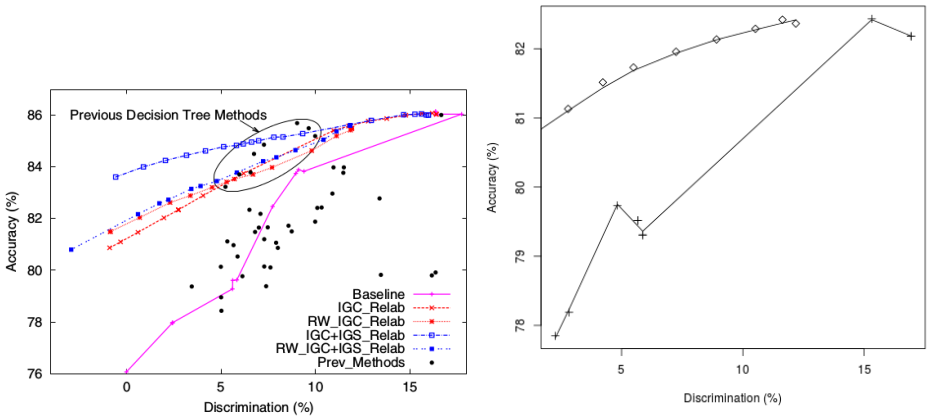

Figure 1: Accuracy-discrimination trade-off for discrimination aware classifiers. On left: original results, on right: reproduced results (diamonds represent leaf relabeling and pluses baseline method).

- Splitting is stopped when the number of samples at the node drops below 1000 or the node entropy is less than 0.001

The efficiency of the leaf relabeling technique is mostly affected by how accurately the RELAB algorithm is implemented. Here again, due to the above mentioned limitations we chose a heuristic approach and did not aim to an optimal solution. More precisely we first randomly relabeled the leaves of the tree until the discrimination dropped under a required value, then we repeated the procedure 1000 times and chose the configuration which had the smallest loss in accuracy.

Likewise in the paper we used 10-fold cross validation and compared the results with the baseline method. While building trees with leaf relabeling we took 9 evenly spaced discrimination thresholds starting from the highest value of the initial tree and ending with 0. In case of baseline trees we gradually omitted up to 9 attributes. We only tested the method with the non-modified information gain (IGC) for splitting, although the implementation of the others approaches were also included [1]. The results from the original study alongside with our results are depicted in Fig. 1. As we see the accuracy achieved by the unmodified tree is approximately 4% smaller in our case, but the accuracy after the discrimination has been nearly eliminated via relabeling is quite comparable in both cases (only compare IGC_Relab with ours). There is a difference in baseline curves as well: in our case the drop in discrimination is significant after removal of 2nd most correlating attribute (which is hours per week), while in the case of original study the drop happens much later. There seems to be a difference in the order of attributes dropped, but as it is unspecified in the original study, it cannot be verified. All the other differences between the results are most likely due to approximations used in our calculations, but possibly also caused by the fact that we did not exclude the data samples with missing attribute values (thus considering an unknown value as a separate value itself).

[1] The authors commented that IGC-IGS approach did not lead to good results. In our experiment we were unable to confirm or reject it, as the computation did not complete. The reasons for that remained however unclear