Report for T-61.6010 - Special Course in Computer and Information Science I: Non-discriminatory Machine Learning

Polina Rozenshtein
student number 338125

November 18, 2015

Project topic

For a course project I have chosen to implement algorithms from paper by Hajian et al. [1], which I have presented in class. The authors address problem of prevention of direct and indirect discrimination in frequent association rule mining. The proposed approaches are based on altering records in the database (DB), such that a number of discriminatory rules, obtained for a transformed database, is minimized.

Implementation

My Python implementation of all 4 proposed algorithms can be found at course's GitHub [1]. Frequent itemsets and association rules are part of input for all algorithms and I use code for Apriori algorithm from `https://github.com/asaini/Apriori`.

For this report I experimented with algorithms for direct discrimination protection: DIRECTED RULE PROTECTION (METHOD 1) and DIRECTED RULE PROTECTION (METHOD 2). I refer to them as DRP1 and DRP2. Both algorithms iterate through direct α-discriminatory frequent association rules $A, B \rightarrow C$, where A is a subset of predetermined discriminatory items (DI) and C is a classification label. Algorithm DRP1 modifies records of DB, which support $\neg A, B \rightarrow \neg C$ by flipping $\neg A$ to A. Algorithm DRP2 modifies the same records by flipping $\neg C$ to C.

Negation of the itemset is a crucial ingredient of these approaches. However, the authors do not provide clear information on whether they consider only 1-itemsets to be negated and how to deal with non-binary items. Thus, I implemented the most general case, when negation for itemset $A = (a = 1, b = 1)$ is a set of itemsets $\{(a = 0, b = 1), (a = 1, b = 0), (a = 0, b = 0)\}$. In addition, if a is not binary, e.g. $a \in \{1, 2, 3\}$, then $\neg(a = 1)$ is defined as set $\{a = 2, a = 3\}$. To create a negation of $A = (a = 1, b = 1)$ with $a \in \{1, 2, 3\}$ and $b \in \{0, 1\}$ we need to construct a set of all possible negations $\{(a = 2, b = 1), (a = 3, b = 1), (a = 1, b = 0), (a = 2, b = 0), (a = 3, b = 1)\}$ and then try all of them as $\neg A$.

Experiments

As a test dataset I selected one of dataset used in the paper, namely Adults dataset of income from `http://archive.ics.uci.edu/ml/machine-learning-databases/adult/`. The dataset contains 32561 records. I used 3 versions of it: original with continuous attributes, binarized (the criteria of binarization can be explicitly seen in the code) and semi-binarized (only age is binarized by thresholding against 30).

The authors set DI={gender: Female, age: ≤ 30} and run Apriori with minsupport=0.02 and minconfidence=0.1. This results in 5092 of classification rules of shape $X \rightarrow C$. This is rather questionable, why the authors aim on protection for rules with so low confidence. Moreover, increase of minconfidence yields to significant reduction in number of α-discriminative rules (up to 10-100). Thus, it is rather suspicious, that minconfidence is kept that low in order to report significant improvement, achieved by the proposed approaches. To verify the capability of the protective algorithms, here I use larger minsupport and minconfidence=0.6. Corresponding number of frequent rules and used discriminatory sets can be found in Table .

To evaluate the results, I calculate proposed metrics: DDPD - a fraction of α-discriminative rules, which are not α-discriminative in the modified dataset; DDPP - a fraction of α-protective rules, which remain α-protective; MC - a fraction of frequent rules, which are not frequent in the transformed dataset; GC - a fraction of frequent rules in the transformed dataset, which are not frequent in the original data.

As a baseline (BL) I use naive approach, which removes discriminating attributes from the dataset. The results are shown in Table 2.

Table 3 presents performance of DRP2 on original dataset for different α.

Conclusions

For original and semi-binary dataset both approaches exhibit behavior, similar to the one presented in the paper: they far outperform baseline, remove all α-discriminative rules and preserve set of α-protective rules. The performance degrades on binarized dataset, however this may be explained by crude binarization of such non-binary fields as job-title or county of origin.

[1] `https://github.com/fairml/aalto-seminar-2015/tree/master/Hajian13_byPolina`

Similarly to published results, DRP2 tends to outperform DRP1. However it introduces more ghost rules. With increase of α accuracy of DRP2 increases, because total number of α-discriminative rules decreases. In addition, it needs to modify fewer records to fix each rule.

To sum up, implemented approaches are clearly intuitive and mechanical: they greedily edit database records until α-based constraint is met. Although there is no guarantee on preserving frequent rules, the experimental results do not show signification distortion. Thus, this is an effective approach to preprocess a dataset for decision rules mining.

dataset	minsupp	minconf	num. of rules	DI
original	0.09	0.6	1031	{sex=Female; marital-status: Never-married}
semi-binary	0.2	0.6	282	{sex=Female; age: <=30}
binary	0.4	0.6	663	{education: No-Degree; marital-status: Not-Married}

Table 1: Datasets characteristics.

dataset	method	α	num. of α-disc. rules	DDPD	DDPP	MC	GC	num. of modified rec.
original	BL	NA	92	NA	NA	0.57	0	NA
	DRP1	1.2	92	1	0.94	0.06	0.10	3130
	DRP2	1.2	92	1	1	0	0.20	3717
sime-binary	BL	NA	12	NA	NA	0.50	0	NA
	DRP1	1.2	12	0.92	1	0	0.01	773
	DRP2	1.2	12	1	1	0	0.17	1399
binary	BL	NA	22	NA	NA	0.44	0.84	NA
	DRP1	1.2	22	1	0.01	0.99	0	590
	DRP2	1.2	22	1	0.08	0.92	0.13	538

Table 2: Utility measures.

α	num. of α-disc. rules	DDPD	DDPP	MC	GC
1	237	0.74	1	0	0.44
1.1	175	0.98	1	0	0.31
1.2	92	1	1	0	0.20
1.3	23	1	1	0	0.12
1.4	0	NA	NA	NA	NA

Table 3: Results of DRP2 for different α

Bibliography

[1] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 25(7):1445–1459, 2013.