ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Discrimination Discovery and Prevention in Data Mining

Jagriti Singh[*1], Prof. Dr. S.S. Sane[2]

[*1] Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra - 422003, India

[2] Head of Department, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra - 422003, India

pritijagriti@gmail.com

Abstract

Data mining is an important technology for extracting useful knowledge in large collections of data. However, the perceptions about data mining are negative as potential privacy invasion and potential discrimination. Unjustified distinction of individuals based on their membership in a certain group or category are referred as discrimination. At first sight, automated data collection and data mining techniques such as classification rule mining may give a sense of fairness and may not guide themselves by personal preferences. However, classification rules are actually learned by the system from the training data and training data sets itself are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc. To extract knowledge without violation such as privacy and non-discrimination is most difficult and challenging. The discrimination discovery and prevention techniques have been discussed in this papers.

Keywords: Discrimination Discovery, Data Mining, Discrimination Prevention Technique

Introduction

Data mining is an important technology for extracting useful knowledge in large collections of data. However, the perceptions about data mining are negative as potential privacy invasion and potential discrimination. Unjustified distinction of individuals based on their membership in a certain group or category are referred as discrimination. The word discrimination originates from the Latin discriminare, which means to “distinguish between".

From a legal perspective, discrimination arises only through the application of different rules or of the same rule or practice to different situations or practices to comparable situations. When any rules or practices explicitly favor one person than another, is known as direct discrimination, sometimes called systematic discrimination or disparate treatment. An apparently neutral provision, practice or criterion which results in an unfair treatment of a protected group called as indirect discrimination and also sometimes called as disparate impact.

To extract knowledge without violation such as privacy and non-discrimination is most difficult and challenging. In this paper we will analysis the potential discrimination, measure them and define algorithm to remove the discrimination.

Literature Survey

Discrimination prevention has been recognized as an issue in a tutorial by (Clifton, 2003) [1] where the danger of building classifiers capable of racial discrimination in home loans has been put forward. Data mining and classififcation models extracted from historical data may discover traditional prejudices for example, mortgage redlining can be easily recognized as a common pattern in loan data but so solution was provided in this tutorial.

Pedreschi et al. (2008) [2]; propose the extraction of classification rules of the form A, B → C, called potentially discriminatory (PD) rules, to unveil contexts B of the dataset where the protected group A suffered from underrepresentation w.r.t the positive decision C or from over-representation w.r.t the negative decision C. A is a non-empty itemset, whose elements belong to a fixed set of protected groups. C is a class item denoting the negative decision, e.g., credit denial, application rejection, job firing, and so on. Finally, B is an itemset denoting a context of possible discrimination. The degree of over-representation is measured by the ER measure (called extended lift).

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Pedreschi et al. (2009) [3] alter the confidence of classification rules inferred by the CPAR algorithm. Calders and Verwer (2010) [4] act on the probabilities of a naïve Bayes model. Kamiran et al. (2010) re-label the class predicted at the leaves of a decision tree. The naive approach of deleting attributes that denote protected groups from the original dataset does not prevent a classifier from indirectly learning discriminatory decisions, since other attributes (sometimes called redundant encodings) that are strongly correlated with them could be used as proxies by the mining algorithm. This has been repeatedly observed in several contexts: credit scoring (Fortowsky and LaCour-Little, 2001 [5]), predictive statistics (Pope and Sydnor, 2007 [6]), and data mining (Ruggieri et al., 2010b; Calders and Zliobaitye, 2013).

Calders and Verwer (2010) [4] consider three approaches to deal with naive Bayes models, two of which consist in modifications to the learning algorithm: training a separate model for each protected group; and, adding a latent variable to model the class value in the absence of discrimination.

Kamiran et al. (2010) [15] modify the entropy-based splitting criterion in decision tree induction to account for attributes denoting protected groups.

Zliobaitye et al. (2011) [11] prevent excessive sanitization by taking into account legitimate explanatory variables that are correlated with grounds of discrimination, i.e., genuine occupational requirements.

The approach of Luong et al. (2011) [12] extends to discrimination prevention by changing the class label of individuals that are labeled as discriminated against.

Using additional background knowledge, Hajian and Domingo-Ferrer (2012) [13] perturb the training set so as to reduce the degree of indirect discrimination, which is measured in terms of the number of rules that could be inferred using the approach of Ruggieri (2010a,b) [14].

Kamishima et al. (2012) [16] measure the indirect causal effect of variables modeling grounds of discrimination on the independent variable in a classification model by their mutual information. Then they apply a regularization (i.e., a change in the objective minimization function) to probabilistic discriminative models, such as logistic regression.

Kamiran and Calders (2012) [10] compare sanitization techniques such as changing class labels based on prediction confidence, instance re-weighting, and sampling.

Discrimination Discovery

At first sight, automated data collection and data mining techniques such as classification rule mining may give a sense of fairness and may not guide themselves by personal preferences. However, classification rules are actually learned by the system from the training data and training data sets itself are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc. As a result actual discovery of discrimination situations, practices may be extremely difficult task. The reasons are mainly as:

- Personal data in decision records are highly dimensional. Due to this, a huge number of possible contexts may, or may not, be the theater for discrimination.
- Complexity in indirect discrimination: the feature that may be the object of discrimination, e.g., the race, is not directly recorded in the data.

The possibility of accessing to historical data concerning decisions made in socially-sensitive tasks is the starting point for discovering discrimination.

Measure for Discrimination

Assume a set of attributes

- $A = \{A_1, A_2, \ldots, A_n\}$ is a set of attributes with domains $dom(A_i)$, $i = 1, \ldots, n$.
- A tuple X is an element of $dom(A_1) \times \ldots \times dom(A_n)$ over the $schema(A_1, \ldots, A_n)$.The value of X for attribute A_i is denoted by $X(A_i)$.
- A dataset D is a finite set of tuples over scheme $(A_1, \ldots, A_n)$.
- A label dataset over schema (A1, . . ., An, class) is a finite set of tuples.
- Assume class has binary domain dom(class) = { - , +} where "+" is a desirable class.
- A special attribute $S \in A$, is sensitive attribute with multiple values.
- P is set of favored community values and Q is set of deprived community values.
- Domain of S is dom(S) = {P,Q}.

Replacing all values of P with new dedicated value *w* and Q with new dedicated value *b*

The discrimination can be defined as follows:

Definition 1*(Discrimination in labeled dataset)*: The discrimination in given labeled dataset D with respect to the group S = b, is denoted by $disc_{S=b}(D)$, define as:

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

$$\text{discS} = \text{b}(\text{D}) \coloneqq \frac{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{w},\ \text{X}(\text{class}) = +\}|}{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{w}\}|} - \frac{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{b}, \text{X}(\text{class}) = +\}|}{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{b}\}|}$$

Definition 2*(Discrimination in a classifier's predictions):* The discrimination of the classifier C with respect to the group S = b in unlabeled dataset D is denoted by $\text{disc}_{S=b}(C,D)$, define as:

$$\text{discS} = \text{b}(\text{C}, \text{D}) \coloneqq \frac{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{w},\ \text{C}(\text{X}) = +\}|}{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{w}\}|} - \frac{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{b}, \text{C}(\text{X}) = +\}|}{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{b}\}|}$$

Discrimination Prevention Technique in Data Mining

There is need of disruptive technologies for the construction of human knowledge discovery systems that, by design, over native technological safeguards against discrimination. To ensure this, these computational models should be free from discrimination and data preprocessing technique for classifier is one way.

There are three different approaches for discrimination prevention in data mining:

- *Preprocessing:* Removing of discrimination from original source data in such a way that no unbiased rule can be mined from the transformed data and applying any standard algorithm. This preprocessing approach is useful in such cases where data set should be published and performed by external parties.
- *In-processing:* Change of knowledge discovery algorithm in such a way that resulting model do not contain biased decision rules. In-processing discrimination prevention depends on new special purpose algorithm. In this standard data mining algorithm cannot be used.
    - *Postprocessing:* Instead of removing biases from original data set or modify the standard data mining algorithm, resulting data mining models are modified. This approach does not allow the data set to be published, only modified mining models can be published. So this can be performed only by data holder.

Although some of the methods have already been proposed for each of the above mentioned approach, but still this is a challenge to remove the discrimination from the original data set. One might be able to think that direct implementation of preprocessing techniques can remove the discriminatory attribute from the original data set and solve the problem of discrimination but this would solve the direct discrimination but not indirect discrimination. This would also cause much of information loss from the original data set.

Preprocessing Algorithm for Discrimination Prevention

The propose solution is to learn a non-discriminating classifier which use the sensitive attribute S only learning time and not at prediction time. The solution is for removing discrimination from training dataset.

In original dataset D sensitive attribute is categorical and its domain is non-binary. In dataset Class has binary domain dom(Class) ={-,+} where "+" is desirable class. The values of sensitive attribute for favored community is replaced with new dedicated value w and the values for deprived community with new dedicated value b.

Algorithm

1: **Input:** Dataset D, Sensitive attribute S, Class
2: **Output:** Classifier learned on reweighed D
3: **for** s ∈{b,w} do
4: **for** c ∈ {-,+} do
5: **Let**
W(s,c):=
$$\frac{|\{\text{X} \in \text{D}|\text{X}(\text{S}) = \text{s}\}| \times |\{\text{X} \in \text{D}|\text{X}(\text{Class}) = \text{c}\}|}{|\text{D}| \times |\{\text{X} \in \text{D}\ |\text{X}(\text{Class}) = \text{c and X}(\text{S}) = \text{s}\}|}$$
6: **end for**
7: **end for**
8: Dw :={}
9: **for** X in D
10: Add (X, W (X (S), X (Class))) to D_W
11: **end for**
12: Train a classifier C on training set D_W, taking onto account the weights
13: **return** Classifier C

The propose system preprocess the data to remove discrimination before a classifier is learned.

Discussion and Future Scope

Classification models usually make predictions on the basis of training data. If the training data is biased towards certain groups or classes of objects, e.g., there is racial discrimination towards black people, the learned model will also show discriminatory behavior towards that particular

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

community. This partial attitude of the learned model may lead to biased outcomes when labeling future unlabeled data objects. Presented preprocessing approach allow for removing discrimination from the dataset more efficiently than simple method such as, e.g., removing the sensitive attribute from the training dataset.

The work has been done till yet was restricted only one binary sensitive attribute. But the propose work is extended for non-binary categorical a sensitive attribute. This can be extended for more than one sensitive attribute in as data set.

Conclusion

Discrimination is a very important issue when considering the legal and ethical aspects of data mining. Most of us do not want to be discriminated based on their gender, religion, nationality, age and so on, especially when these attributes are used for making decisions about our jobs, loans, insurance etc. Discrimination must be detected and removed to get the unbiased results.

This paper has pressnted the type of discriminations, measure for discrimination, discriminatioin prevention techniques and a method of pre-processing technique for removing the discrimination and subsequently a classifier is learned on this unbiased data in data mining.

Reference

1. *C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003), Tutorial, Washington, DC (USA), 2003.*
2. *D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In Y. Li, B. Liu, and S. Sarawagi, editors,Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2008), pages 560–568.ACM, 2008.*
3. *D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In Proc. Of the SIAM Int. Conf. on Data Mining (SDM 2009), pages 581–592. SIAM, 2009.*
4. *T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. Data Mining &Knowledge Discovery, 21(2):277–292, 2010.*
5. *E. Fortowsky and M. LaCour-Little. Credit scoring and disparate impact. Working paper, Wells Fargo Home Mortagage, 2001. http://fic.wharton.upenn.edu/fic.*
6. *D. G. Pope and J. R. Sydnor. Implicit statistical discrimination in predictive models. Working Paper 2007-09-11, Risk Management and Decision Processes Center, The Wharton School of the University of Pennsylvania, 2007. http://opim.wharton.upenn.edu.*
7. *C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003), Tutorial, Washington, DC (USA), 2003. http://www.cs.purdue.edu.*
8. *S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. ACM Trans. on Knowledge Discovery from Data, 4(2):Article 9, 2010b.*
9. *T. Calders and I. Zliobaitye. Why unbiased computational processes can lead to discriminative decision procedures. In B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky, editors, Discrimination and Privacy in the Information Society, volume 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics, pages 43–57. Springer, 2013.*
10. *F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33:1–33, 2012.*
11. *I. Zliobaitye, F. Kamiran, and T. Calders. Handling conditional discrimination. In Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2011), pages 992–1001. IEEE Computer Society, 2011.*
12. *B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In C. Apt´e, J. Ghosh, and P. Smyth, editors, Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011), pages 502–510. ACM, 2011.*
13. *S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining.IEEE Transactions on Knowledge and Data Engineering, page to appear, 2012.*
14. *S. Ruggieri, D. Pedreschi, and F. Turini. Integrating induction and deduction for*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

finding evidence of discrimination.Artificial Intelligence and Law, 18(1):1–43, 2010a.

15. *F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In G. I. Webb, B. Liu,C. Zhang, D. Gunopulos, and X. Wu, editors, Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2010), pages 869–874. IEEE Computer Society, 2010.*
16. *T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Proc.of the Eur. Conf. on Machine Learning and on Principles and Practice of Knowledge Discovery in Databases ECML-PKDD 2012), volume 7524 of LNCS, pages 35–50. Springer, 2012.*

Author Biblography

	**Jagriti Singh** Post Graduate Student of Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra, India. She received her B.Tech Degree in Information Technology from Uttar Pradesh Technical University, Uttar Pradesh. (e-mail: pritijagriti@gmail.com)
	**Prof. Dr. S. S. Sane** Head of Computer Engineering Department, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra, India. (e-mail: sssane65@yahoo.com).