"REVIEW ON DISCRIMINATION AND PREVENTION POLICIES OVER WEB DATA"

PROF. NITIN KHACHANE
Department of Computer Science & Engineering, PRMIT Badnera, Amravati, India
khachane.nitin001@gmail.com

ABSTRACT: *Data warehouse is emerging need of web users. Every user wants to store data in centralized location from where it can access the data. And data mining is the process of extracting the data which is most important or knowledgeable. Some time user access the data which is sensitive and on the basis of that discrimination can be occurred. According to the different area, state, country discrimination can be happened. Direct and indirect are the two most observable discrimination processes which are identified. This paper gives literature survey and identifies the some important facts that can consider.*

Keywords: discrimination, preprocessing, classification

1. INTRODUCTION

The aim of data mining is to extract useful information, such as patterns and trends, from large making automated decisions, like loan granting/denial, insurance premium computation, personnel selection, etc. Firstly, automating decisions may give a sense of fairness: cataloguing rules do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system (e.g., loan granting) from the training data. [3] If the training data are fundamentally biased for or against a particular community, the learned model may show an unfair prejudiced behaviour. In other words, the system may infer that just being foreign is a legitimate reason for loan denial and activities of suspects and potential suspects. This can be very useful, but usually at least part of the data on which data mining is applied is confidential and amounts of data. Many governments are gathering large amounts of data to gain insight into methods.

There is several decisions -making tasks which lend themselves to discrimination, e.g. loan granting, education, health insurances and staff selection. In many scenarios, decision - making tasks are supported by information systems. Given a set of information items on a potential customer, an automated system decides whether the customer is to be recommended for a credit or a certain type of life insurance. Automating such decisions reduces the workload of the staff of banks and insurance companies, among other organizations. [4]The use of information systems based on data mining technology for decision making has attracted the attention of many researchers in the field of computer.

Preprocessing: Transform the source data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision rule can be mined from the transformed data and apply any of the standard data mining algorithms. I it can be adapted from the privacy preservation literature. The existing systems perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased data set. This approach is useful for applications in which a data set should be published and in which data mining needs to be performed.[8] In -processing: Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. However, it is obvious that in–processing.

2. LITERATURE SURVEYS

Literature survey 1: Fast Algorithms for Mining Association Rules They consider the problem of association rules discovery between items in a large database of sales transactions. For that they present two new algorithms for solving above mentioned problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. [1] They also show the best features of their two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid. Scale -up experiments show that Apriori Hybrid scales linearly with the number of transactions. AprioriHybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database.

The Apriori and AprioriTid algorithms they propose vary fundamentally from the AIS and SETM algorithms which was proposed in previous methods in terms of which candidate item sets are counted in a pass and in the way that those candidates are generated. In both the AIS and SETM algorithms, which was proposed by the existing methods, candidate item sets are generated on they during the pass as data is being read. Specially, after reading a transaction, it is determined which of the item sets found large in the previous pass are present in the transaction. New candidate item sets are generated by extending these large item sets with other items in the transaction.

3. Problem Definition:

The problem of using their discrimination model is that it is based on assumptions that might not always hold in practice. They remove low frequency counts by pooling any bin that occurs less than 50 times which may lead problem.

One obvious drawback of such a method is that the number of parameters to describe the distribution of S is exponential in the number of attributes Ai. Therefore it would be beneficial to consider other models that could be inserted" into the Bayesian model to replace the probability table, such as, e.g., a decision tree. Obviously they didn't explore why the convergence of Expectation maximization (EM) was relatively poor, even for the synthetic datasets where all conditions for a successful convergence were satisfied.

In other methods they consider removing discriminatory attributes from the dataset to handle discrimination prevention, there may be other attributes that are highly correlated with the sensitive one. Hence, one might decide to remove also those highly correlated attributes as well. Although this would solve the discrimination problem, in this process there is a chance of loss of much useful information. Some of them concentrated on discrimination discovery, by considering each rule individually for measuring discrimination without considering other rules or the relation between them. In proposed system, they introduced anti-discrimination for cyber security applications based on data mining. The proposed solution is based on the fact that the dataset of decision rules would be free of discriminatory claim. The proposed solution in removing all evidence of discrimination from the original dataset is called as degree of discrimination prevention. The impact of the proposed solution on data quality is called as degree of information loss.

A discrimination prevention method should provide a good trade -off between both aspects above. The following is the evaluating their solution measures are proposed as: Discrimination Prevention Degree (DPD), Discrimination Protection Preservation (DPP), Misses Cost (MC), Ghost Cost (GC).[14]

Their contribution concentrates on producing training data while saving their use to detect real intrusion or crime which are free or nearly free from discrimination. In order to control incrimination in a dataset, a first step consists in discovering whether there exists discrimination. If any discrimination is found, the dataset will be mod if field until discrimination is brought below a certain threshold or is entirely eliminated.

Drawback:

They didn't present a unified discrimination prevention approach based on the discrimination hiding idea that encompasses both direct and indirect discrimination. Literature survey 5: Classification with No Discrimination by Preferential Sampling.

In existing system, they introduced the concept of discrimination aware classification and proposed a solution to the problem based on changing the class labels. Preferential Sampling (PS) introduces a less intrusive technique to make the dataset unbiased than changing the class labels. In existing work also have similar motivation towards the solution of the discrimination problem. They concentrate on identifying discriminatory rules that are present in a dataset; hence they learn potential discriminatory guidelines that have been followed in the decision procedure.[13][14].

In the Proposed work they closely related to class imbalance problem. In existing system they introduced a synthetic minority over-sampling technique (SMOTE) for two class problems that over -sampled the minority class by creating synthetic examples rather than replicating examples. In contrast PS concentrates only on border regions.

Classification with No Discrimination by Preferential Sampling is an excellent solution to the discrimination problem. It gives promising results with both stable and unstable classifiers. It reduces the discrimination level by maintaining a high accuracy level. It gives similar performance to "massaging" but without changing the dataset and always outperforms the "reweighing". In existing method, simply removing the discriminatory attribute from the training data in the learning of a classifier for the classification of future data objects is not enough to solve this problem, because often other attributes will still allow for the identification of the discriminated community.

During the investigation of literature survey, some issues were identified and are summarized using the following points:

• The methods focus on the attempt to detect discrimination in the original data only for one discriminatory item and also based on a single measure.
• They do not include any measure to evaluate how much discrimination has been removed and how much information loss has been incurred.
• It focuses either on direct discrimination or indirect discrimination or not on both together.
• The approaches do not shows any measure to evaluate how much discrimination has been removed, and thus do not concentrate on the amount of information loss generated.[12] So the proposed work in data mining proposes preprocessing methods which overcome the above limitations. And introduces new data transformation methods (rule protection and rule generalization (RG)) are based on measures for both direct and indirect discrimination and can deal with several discriminatory items.[17]

4. CONCLUSION

From the above literature survey it can be conclude that existing system has some drawbacks like some researcher works on single attribute, some researcher provide only direct discrimination. The system can be implemented which can be work on both for direct and indirect discrimination and use efficient preprocessing algorithm to overcomes the problems of in processing and post processing.

5. REFERENCE

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases, "Proc. 20th Int'l Conf. Very Large Data Bases,pp. 487 -499, 1994.

[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination -Free Classification, "Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277 - 292, 2010.

[3] European Commission, "EU Directive 2004/113/EC on Anti-Discrimination," http://eur-lex.europa.eu /LexUriServ/ LexUriServ.do?uri= OJ: L: 2004:373:0037:0043: EN: PDF, 2004.

[4] European Commission, "EU Directive 2006/54/EC on Anti-Discrimination," http://eur-lex.europa.eu/ LexUriServ /LexUriServ.do? Uri=OJ:L:2006:204:0023:0 036:en: PDF, 2006.

[5] S. Hajian, J. Domingo-Ferrer, and A. artı´nez-Balleste´, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection, "Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11),pp. 47-54, 2011.

[6] S. Hajian, J. Domingo-Ferrer, and A. artı´nez -Balleste´, "Rule Protection for Indirect Discrimination Prevention in Data ining,"Proc. Eighth Int'l Conf. Modeling Decisionsfor rtificial Intelligence (MDAI '11),pp. 211-222, 2011.

[7] F. Kamiran and T. Calders, "Classification without Discrimina-tion," Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09), 2009.

[8] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.

[9] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning", Proc. IEEE Int'l Conf. Data Mining (ICDM '10),pp. 869-874, 2010.

[10] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining,"Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08),pp. 560-568, 2008.

[11] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially -Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09),pp. 581-592, 2009.

[12] D. Pedreschi, S. Ruggieri, and F. Turini,"Integrating Induction and Deduction for Finding Evidence of Discrimination,"Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09),pp. 157-166, 2009.

[13] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery,"ACM Trans. Knowledge Discovery from Data,vol. 4, no. 2, article 9, 2010.

[14] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases,"Proc. ACM Int'l Conf. Management of Data (SIGMOD '10), pp. 1127-1130, 2010.

[15] P.N . Tan, M. Steinbach, and V. Kumar,Introduction to Data Mining. Addison - Wesley, 2006