ISSN:0975-9646

Survey of Approaches for Discrimination Prevention in Data Mining

Harshali M. Saindane[#1], Mrs. V. L. Kolhe[*2]

[#]*Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune, Savitribai Phule Pune University, India*

[*]*Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune, Savitribai Phule Pune University, India*

***Abstract*— Data mining is an important technology for extracting useful patterns from large amount of data. Two major prevalent issues in data mining are privacy violation and discrimination. Discrimination arises when people are given unfair treatment on the basis of their sensitive features like gender, race, religion etc. Types of discrimination are direct and indirect discrimination. Direct discrimination consists of rules based on sensitive attributes like religion, race, community etc. Indirect discrimination occurs when decisions are based on non sensitive attributes which are closely related to sensitive attributes. Automated data collection and data mining techniques such as classification rule mining are used for making automated decisions by decision support systems. These systems are used for personnel selection, loan granting etc. If the training data sets are biased with respect to the sensitive features, discriminatory decisions may occur. Antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining. The main purpose of this survey paper is to understand the existing approaches for discrimination prevention.**

***Keywords*— Data Mining, Discrimination prevention, Anti-discrimination, Direct and indirect discrimination, privacy preservation, discrimination measures.**

I. INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that deal with the automatic extraction of useful patterns from large amounts of data. Data mining techniques are used in business and research and are becoming more and more popular with time. There are two issues related to data mining. These issues are privacy violation and potential discrimination. Discrimination is a very important issue when considering the legal and ethical aspects of data mining. It can be viewed as the act of illegally treating people on the basis of their belonging to a specific group[1]. People may be discriminated because of their race, ideology, gender, etc, if those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. Antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining. Services in the information society allow for automatic and routine collection of large amounts of data. For a given set of information attributes about a customer, an automated system decides whether the customer is to be recommended for a credit or a job selection. Automated such decision systems for these tasks reduces the workload of the staff of banks and insurance companies, among other organizations. Those data are often used to train association/classification rules in view of making automatic decisions, like loan granting/denial, insurance premium computation, personnel selection, etc. classification rules are actually learned by the system from the training data. If the training data is biased for or against a particular community (e.g.,foreigners), the learned model may show a discriminatory behaviour towards that community. The system may interpret that just being foreign is a legitimate reason for loan denial. Find such potential biases and eliminating them from the training data without harming their decision-making utility is therefore important. Data mining must not become a source of discrimination, since automated decision systems learn from data mining models.

Data mining can be both a source of discrimination and a means for discovering discrimination. Types of discrimination are direct and indirect discrimination[2]. Direct discrimination consists of set of rules that are obtained from sensitive discriminatory attributes like gender, religion etc. Indirect discrimination consists of set of rules that are inferred from attributes closely related to the sensitive ones. Beyond discrimination discovery, making knowledge-based decision support systems free from making discriminatory decisions (discrimination prevention) is a more challenging issue. The challenge increases if there is need to prevent not only direct discrimination but also indirect discrimination or both. There are various approaches available for discrimination prevention in data mining. In order to be able to classify the various approaches, two orthogonal dimensions are used, based on which the existing approaches exist. The first dimension considers whether the approach deals with only direct discrimination, or indirect discrimination, or both at the same time. Based on this dimension, the discrimination prevention approaches are separated into three groups: direct discrimination prevention approaches, indirect discrimination prevention approaches, and direct/indirect discrimination prevention approaches. The second dimension in the classification relates to the phase of the data mining where discrimination prevention occurs. Accordingly, approaches for discrimination prevention fall into following three types as shown in following Fig 1.

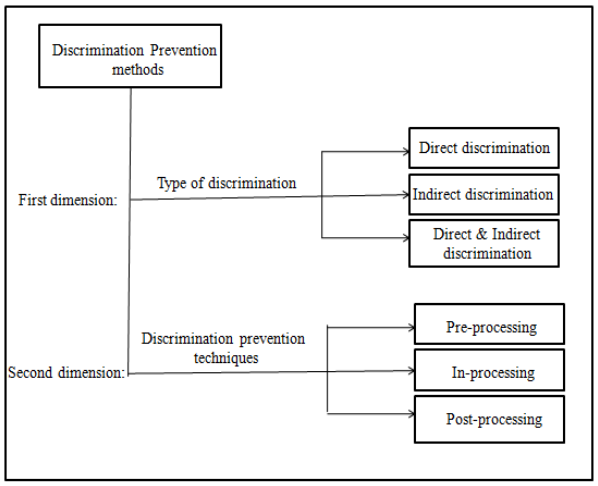

Fig. 1 Taxonomy of discrimination prevention methods[2]

1) Pre-processing:- Approaches belonging to this group 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; any of the standard data mining algorithms can then be applied.

2) In-processing:- Approaches belonging to this group change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. Special purpose data mining algorithms must be used, since standard data mining algorithms cannot prevent discrimination.

3) Postprocessing:- Approaches belonging to this group modify the resulting data mining models, instead of cleaning the original dataset or changing the data mining algorithms.

Although there are some methods for each of the above mentioned techniques, discrimination prevention is a topic of research. This paper gives an overview of the available literature on discrimination prevention approaches in data mining.

The rest of the paper is organized as follows. Section II provides the available literature on approaches for discrimination prevention in data mining. Section III gives the comparative analysis of these techniques. Section IV gives the summary for the topic.

II. Literature Survey

Despite the wide deployment of information systems based on data mining technology in decision making, discrimination in data mining did not receive much attention until 2008. Thus, beyond discrimination discovery, a more challenging issue is to prevent knowledge-based decision support systems from making discriminatory decisions. Some of these approaches are related to the discovery and measure of discrimination. The other approaches deal with discrimination prevention.

A. Classification Rules using α Protection Measure

D. Pedreschi et al.[15] are the first researchers to address the discrimination problem from the point of view of knowledge discovery from databases. This approach belongs to pre-processing method of discrimination prevention. They have shown that discrimination may be hidden in knowledge discovery models extracted from databases, and classification rule models are considered. α-protection is introduced as a measure of the discrimination power of a classification rule containing one or more discriminatory items. The idea is to define such a measure as an estimation of the gain in precision of the rule due to the presence of the discriminatory items. The α parameter is the key for tuning the desired level of protection against discrimination. This approach is based on mining classification rules and reasoning on them on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. The limitation of this approach is that it needs to manually check the α-protection levels.

Application areas: Credit assessment, Job selection, Insurance scoring.

B. Discriminatory Classification Rules

S. Ruggieri et al.[9] have introduced the problem of discovering contexts of discriminatory decisions against protected-by-law groups, and provided a knowledge discovery process for solving it. This approach is based on pre-processing method of discrimination prevention. This approach is based on coding the involved concepts (potentially discriminated groups, contexts of discrimination, measures of discrimination, background knowledge, direct and indirect discrimination) in a coherent framework based on item-sets, association rules and classification rules. In direct discrimination, the extracted rules can be directly mined in on basis of discriminatory attributes. In indirect discrimination, some background knowledge is needed by the mining process which when combined with the extracted rules may lead to discriminatory decisions. This approach cannot be applied to continuous attributes.

Application areas: Credit assessment, Job selection, Insurance companies .

C. Three Naive Bayes Approach

T. Calders et al.[11] have presented a modified Naive Bayes classification approach. This approach belongs to post-processing method of discrimination prevention. The classification task is performed by focusing on independent sensitive attributes. This type of behaviour occurs, when the decision process that leads to the labels in the dataset is biased with respect to sensitive attributes. This approach is motivated by many case studies of decision making, where laws deny a decision that is partly based on discrimination. Three methods based on Bayesian classifier are used for discrimination-aware classification. In the first method, the observed probabilities in a Naive Bayes model are modified in such a way that its predictions become discrimination-free. The second method involved learning two different model. In the third and most involved method a latent variable L is introduced reflecting the latent “true” class of

an object without discrimination. The This approach is not able to work with numerical attributes e.g.:- income as a sensitive attribute.

Application areas: Credit assessment, Job selection, Insurance companies .

D. Classification with No Discrimination Model

Kamiran F. et al.[13] have introduced a Classification with No Discrimination model(CND). This approach belongs to the pre-processing method of discrimination prevention. Classification models usually make predictions on the basis of training datasets. If these data is biased with respect to certain communities or groups, the learned model will also show discriminatory behaviour towards that particular community. This type of prediction behaviour may lead to biased decisions when labelling future unlabeled data. The classification should produce bias free results, also required by law for future data objects in spite of having biased training data. A new classification scheme is used for learning unbiased models on biased training data. This method is based on massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset. A non-discriminating classifier is then learned on this modified dataset. CND classifies the future data with minimum discrimination and high accuracy. This approach also helps in addressing redlining rules.

Application areas: Credit approval, Financial institutions, Insurance companies .

E. Preferential Sampling

Kamiran F. et al.[12] have developed a new solution to the CND problem by introducing a sampling scheme for making the data discrimination free instead of relabeling the dataset. This approach is based on pre-processing method of discrimination prevention. It consists of sampling the data objects with replacement to make the dataset bias free. A Preferential Sampling (PS) scheme is introduced to make the dataset bias free. This approach deals with changing the distribution of different data objects for a given data to make it discrimination free. The idea is that the data objects close to the decision boundaries are more prone to be the victim of discrimination. The distribution of these borderline objects is changed to make the dataset discrimination free. To find the least certain elements, a ranking function, learned on the original data, is used to identify the data objects close to the borderline. Then, based on the sanitized data, a non-discriminatory model can be learned. Since this model is learned on non-discriminatory data, it reduces the prejudicial behaviour for future classification. It helps in obtaining good results with both stable and unstable classifiers. It mitigates the discrimination level by maintaining high accuracy level.

Application areas: Labour selection, granting mortgage, insurance companies.

F. Decision Tree Learning

F. Kamiran, et al[8] have presented the construction of a decision tree classifier without discrimination. This approach belongs to in-processing method of discrimination prevention. They have considered discrimination aware classification as a multi-objective optimization problem. They have constructed the decision trees with non-discriminatory constraints. This is a different approach for addressing the discrimination-aware classification problem. In this approach, the non-discriminatory constraint is pushed deeply into a decision tree learner by changing its splitting criterion and pruning strategy by using a novel leaf relabeling approach. It outperforms the other discrimination aware techniques by giving much lower discrimination scores and maintaining the accuracy high. This approach is suitable for cases wherein training set is discriminatory and test set is non-discriminatory.

Application areas: Car insurance, Personnel selection, Banking sectors .

G. Decision Theory approach

K. Asim et al.[5] have developed two flexible and easy solutions for discrimination-aware classification based on an intuitive hypothesis: discriminatory decisions are often made close to the decision boundary because of decision maker's decisions. Decision theoretic concepts of prediction confidence and ensemble disagreement have been used for this purpose. Their first approach is called Reject Option based Classification (ROC). It makes use of the low confidence region of a single or an ensemble of probabilistic classifiers for discrimination reduction. It invokes the reject option and labels instances belonging to deprived and favoured groups in a manner that reduces discrimination. Second approach is called Discrimination-Aware Ensemble (DAE). It makes use of the disagreement region of a classifier ensemble to relabel deprived and favoured group instances for reduced discrimination. This approach gives better control and interpretability of discrimination-aware classification to decision makers.

Application areas: Personnel selection, Crime detection

H. DCUBE tool for discrimination discovery

Salvatore et al.[10] have developed the DCUBE system which is based on an existing approach of discrimination prevention. It is based on classification rule extraction and analysis, by analysing using an Oracle database. DCUBE is an analytical tool supporting the interactive and iterative process of discrimination discovery. The intended users of DCUBE include: owners of socially sensitive decision databases, anti-discrimination authorities and auditors, researchers in social sciences, economics and law. DCUBE tool helps in guiding the users about the legal issues about discrimination hidden in data, and through several legally-grounded analysis to deal with discriminatory situations. These tool helps in providing knowledge to users about discrimination facts in a user friendly manner.

Application areas: Anti-discrimination organizations, researchers in social sciences.

I. Discrimination for Crime and Intrusion Detection

S. Hajian et al.[6] have introduced anti-discrimination in the context of cyber security. They have introduced a new discrimination prevention method based on data transformation that can consider several discriminatory attributes and their combinations. This approach concentrates on producing training data which are free or nearly free from discrimination while preserving their usefulness to detect real intrusion or crime. In order to control discrimination in a dataset, the first step consists of discovering discrimination. If any discrimination is found, the original training dataset is modified until discrimination is brought below a certain discriminatory threshold or is entirely eradicated. They have introduced some measures for evaluating this method in terms of its success in discrimination prevention and its impact on data quality. The drawback of this approach is that it considers only with direct discrimination.

Application areas: Crime and intrusion detection.

J. Integrating Induction and Deduction for Discrimination Discovery

D. Pedreschi et al.[14] have presented a reference model for finding evidence of discrimination in automatic Decision Support Systems. This approach consists in first extracting frequent classification rules from the set of decisions taken by the DSS over an input pool dataset, with an inductive approach based on data mining. The set of rules is considered as a model of the historical decisions of the DSS. The induced rules are then loaded and the key legal measures are used and reasoning for discovering patterns of direct and systematic discrimination is done. This is the deductive part of the approach. Reference model provides a framework for discrimination analysis by translating key concepts from the legal viewpoint into quantitative measures and deduction rules over classification and association rules extracted from a training set and/or from background knowledge.

Application areas: Personnel selection, financial institutions, credit assessment.

K. Rule Protection for Indirect Discrimination

S. Hajian et al.[7] have introduced a new pre-processing approach for indirect discrimination prevention based on data transformation that can consider several discriminatory attributes and their combinations. This approach belongs to pre-processing method of discrimination prevention. They have used some measures for evaluating this approach in terms of its success in discrimination prevention and its impact on data quality. This is the first approach that uses a discrimination prevention method for indirect discrimination. In order to prevent indirect discrimination in the training dataset, first step consists in discovering indirect discrimination. If any discrimination is found, the original training dataset is modified until discrimination is brought below a certain threshold or is entirely eradicated. It aims at generating training data which are free or almost free from indirect discrimination while preserving their usefulness to data mining algorithms. This approach can deal with only indirect discrimination.

Application areas: Credit assessment, financial institution,insurance companies.

L. Direct and indirect Discrimination Prevention

S Hajian et al.[1] have addressed the problem of direct and indirect discrimination prevention in data mining. This approach belongs to pre-processing method of discrimination prevention. They have developed new techniques applicable for direct or indirect discrimination prevention individually or simultaneously. This approach is an extension to their earlier work. The traditional approaches for discrimination prevention consider only one discriminatory attribute. They can deal with either direct discrimination or indirect discrimination but not both. This approach uses various methods to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (non-discriminatory) classification rules. It works in following phases: Discrimination measurement and Data Transformation. The first phase deals with direct and indirect discrimination discovery that includes identifying α-discriminatory rules and redlining rules. The redlining rules require background knowledge that might be obtained from the original data set itself because of the existence of non-discriminatory attributes that are closely related with the sensitive attributes in the training data set.

The elift and elb measures are used to evaluate the value of direct and indirect discrimination respectively. The values of these measures must be less than value of discriminatory threshold(α)for ensuring that the dataset is free from discrimination. The data transformation methods used are Direct Rule Protection(Method1 that deals with modifying discriminatory attribute value), Direct Rule Protection(Method2 that deals with modifying class item value), Direct Rule Protection and Rule Generalization and Direct Indirect Discrimination Prevention. They have used new metrics to evaluate the utility of these approaches. These approaches are effective at removing direct and/or indirect discrimination biases in the original training data set while preserving data quality. They have introduced utility measures to evaluate the amount of discrimination removal. They have also added two metrics namely Misses Cost and Ghost Cost that help in measuring the information loss. This approach can deal with only binary negations.

Application areas: Credit assessment, insurance companies, crime and intrusion detection, job selection, financial institutions.

III. COMPARATIVE ANALYSIS

This section illustrates comparative analysis of various approaches used for discrimination prevention along with their advantages and limitations. The comparative study is shown in the following table1.

TABLE I
COMPARATIVE ANALYSIS

Paper Title	Approach	Advantages	Limitations
Discrimination aware data mining [15]	α protection measure	Removing the discriminatory attributes with least possible changes.	This method is intrusive.
Classification with no Discrimination by Preferential Sampling[12]	Preferential Sampling scheme	High accuracy level	Concentrates only on the borderline data
Discrimination aware decision tree learning[8]	Decision tree classifier	Lower discrimination score	Construction of decision tree is complex.
Three Naive Bayes Approaches for Discrimination-Free Classification[11]	Naive Bayes approach	Modified Naive Bayes classifier for combating discrimination	It does not considers numeric attributes.
DCUBE: Discrimination Discovery in Databases[10]	DCUBE tool	Interactive and iterative database tool for discrimination discovery.	User need to be well acquainted with SQL database queries.
Decision Theory for Discrimination-aware Classification[5]	ROC and MAE approach	Easy applicability to multiple sensitive attributes	Knowledge of decision theoretic concepts is required.
Discrimination prevention for crime and intrusion detection[6]	Data transformation methods	Preserves data quality	It cannot handle indirect discrimination.
Rule protection for indirect discrimination[7]	Indirect rule protection	Considers several discriminatory attributes	Does not consider background knowledge.
Direct and indirect discrimination prevention methods[2]	Pre-processing	Discovery and discrimination prevention	No measure to evaluate data quality.
A methodology for direct and indirect discrimination in data mining[1]	Data transformation methods	Deals with simultaneous direct and indirect discrimination	It cannot handle attributes with ambiguity in negations.

IV. CONCLUSION

This paper presents a survey of various approaches for discrimination prevention in data mining. From the survey, it can be observed that discrimination prevention is indeed a major issue in data mining. From the survey, it can be observed that approaches based on pre-processing methods are flexible to use than the other two methods since, pre-processing involves transforming dataset so as to remove discriminatory biases from it. The approach[1] is more efficient than the other mentioned approaches since it can handle direct as well as indirect discrimination simultaneously along with preserving data quality.

ACKNOWLEDGMENT

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] S. Hajian and J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, pp. 1445-1459, July 2013.

[2] S. Hajian and J. Domingo-Ferrer, Direct and indirect discrimination prevention methods," in Discrimination and Privacy in the Information Society, pp. 241-254, Springer, 2013.

[3] D. Pedreschi, S. Ruggieri, and F. Turini, The discovery of discrimination," in Discrimination and Privacy in the Information Society, pp. 91-108, Springer, 2013.

[4] Andrea Romei and Salvatore Ruggieri. Discrimination data analysis: A multi-disciplinary bibliography. In Discrimination and Privacy in the Information Society, pages 109-135, Springer, 2013.

[5] F. Kamiran, A. Karim, and X. Zhang, Decision theory for discrimination-aware classification.," in ICDM, pp. 924-929, 2012.

[6] S. Hajian, J. Domingo-Ferrer, and A. Martinez-Balleste, Discrimination prevention in data mining for intrusion and crime detection," in Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on, pp. 47-54, IEEE, 2011.

[7] S. Hajian, J. Domingo-Ferrer, and A. Martinez-Balleste, Rule protection for indirect discrimination prevention in data mining," in Modeling Decision for Artificial Intelligence, pp. 211-222, Springer, 2011.

[8] F. Kamiran, T. Calders, and M. Pechenizkiy, Discrimination aware decision tree learning, "in Data Mining (ICDM), 2010 IEEE 10th International Conference on, pp. 869-874, IEEE, 2010.

[9] S. Ruggieri, D. Pedreschi, and F. Turini, Data mining for discrimination discovery," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 2, p. 9, 2010.

[10] S. Ruggieri, D. Pedreschi, and F. Turini, Dcube: Discrimination discovery in databases," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 1127-1130, ACM, 2010.

[11] 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.

[12] F. Kamiran and T. Calders, Classification with no discrimination by preferential sampling," in Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.

[13] F. Kamiran and T. Calders, Classifying without discriminating," in Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on, pp. 1-6, IEEE, 2009.

[14] D. Pedreshi, S Ruggieri, F Turini. "Integrating induction and deduction for finding evidence of discrimination." Proceedings of the 12th International Conference on Artificial Intelligence and Law. ACM, 2009.

[15] D. Pedreshi, S. Ruggieri F. Turini, Discrimination-aware data mining," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 560-568, ACM, 2008.