ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Technique for Prevent Straight and Indirect Bias in Information Mining

Arati Shrimant Mote[1], Prof. G. P. Chakote[2]

PG Student, Dept of Comp. Engineering, MSS's College of Engineering and Technology, Dr. BAMU University ,Jalna, Maharashtra, India[1]

Assistant Professor, Dept of Comp. Engineering, MSS's College of Engineering and Technology, Dr BAMU University, Jalna, Maharashtra, India[2]

ABSTRACT: For coercing the accommodating understanding disguised in the huge accumulation of database the information mining technology is utilized. There are some negative methodologies happened about the data mining technology, among which the potential protection incursion and potential separation. The last comprises of unreasonably treating people on the premise of their having a place with an exact group. Data mining and automated data collection techniques like the arrangement secured the route for making the computerized judgment like granting or denial the loan, race, religion, and so on. On the off chance that the training information set are one-sided in what respects oppressive traits like gender, race, religion, and so forth., discriminator decision may guarantee. As a result of this reason the data mining technology presented antidiscrimination strategies with including discrimination discovery and avoidance. The discrimination can immediate or circuitous. At the point when any decisions are made on the sensitive attributes around then direct discrimination are happened. While the indirect discrimination are happened when the decision are made on the premise of non-sensitive attributes which are unequivocally connected with the sensitive. Here in this paper we manage discrimination avoidance in data mining and proposed novel system for discrimination prevention action with the post transforming methodology. We proposed Classification based on predictive association rules (CPAR) algorithm, which is a sort of association classification methods. The calculation joins the preferences of both association classification methods and traditional rule based classification. The algorithm used to avoid discrimination anticipation in post handling. We compute the utility of the proposed approach and contrast and the existing methodologies. The exploratory appraisal demonstrated that the proposed strategy is successfully deleting the direct or indirect discrimination biases in the first data set for keeping up the nature of data.

KEYWORDS: Direct discrimination prevention, post-processing, indirect discrimination prevention, antidiscrimination, rule protection, rule generalization, privacy.

I. INTRODUCTION

Discrimination is the counterproductive treatment of an individual taking into account their participation in the certain membership of their enrollment in a certain group or categories. It immerses negating to individuals from one gathering open opportunities which are accessible to another groups. There are a few laws which are intended for keeping the discrimination on the premise of quantities of characteristics like race, religion, gender, nationality, disability, marital status and age in the distinctive settings, for example, employment and training, access to public services, credit and insurance, and so forth. We should take one illustration; the European Union executes the law of comparable treatment among men and ladies in the entrance to and supply of goods and services in [1] or in matters

10.15680/ijircce.2015.0304072

of livelihood and occupation in [2]. Although there are a few standards as opposed to discrimination, these standards are receptive, not proactive. Technology can add proactively to enactment by contributing discrimination discovery and avoidance techniques.

Services in the data society permit and routine collection of extensive measure of data. These data are accustomed to training association or classification in perspective of settling on robotized decision, similar to loan allowing or dissent, insurance premium computation, personnel selection, and so forth. At first sight, automating decisions may give a feeling of reasonableness: classification rules don't control themselves by personal inclination. Be that as it may, at a more classification rules, one understands that order standards are found out by the framework from the training data. On the off chance that the training data are innately one-sided for or against an especially community, the model may demonstrate a discriminatory prejudiced activities. In another words, the real explanation for denying the loan is that the person is have a place from another nation. In this way there is a need to dispose of such a potential inclinations from the training data without influencing the decision making utility. Everybody need to keep their information from turning into the source of discrimination, because of data mining assignments producing discriminatory model from one-sided data sets as a piece of automated decision making. In [3], its closed data mining can be both a source of discrimination and an asset for finding discrimination.

The discrimination can be immediate or roundabout. The immediate discrimination subsists of set of laws or motivation which plainly expressed minority or impediment groups in light of delicate oppressive highlights identified with group memberships. The roundabout discrimination subsists of set of laws or plans which is not plainly demonstrating the discriminatory features, intentionally or unintentionally create discriminatory decision. Stereotypical sample of the roundabout discrimination is redlining by the financial organization. With a minor abuse of language for the benefit of density. In this paper roundabout discrimination is assign as redlining, and the principles which are bringing on roundabout called as redlining rules [3]. In view of the accessibility of the some foundation learning roundabout discrimination happen, how about we take one illustration a sure zip code coincide to a fizzling region or a territory with ordinarily black population.

II. RELATED WORK

Despite the fact that the expansive organization of the data framework in view of the data mining technology in the decision making, the theme of antidiscrimination does not give attention consideration until 2008 [3]. A few techniques are gotten ready for discovery and measures of discrimination. Different strategies manage the discouragement of discrimination.

In paper [3], [4] Perdreschi was initially proposed the revelation of discrimination decisions. The strategy is taking into account mining the classification rules and examination on them on the premise of noteworthy measures of discrimination that formalize legal meaning of discrimination. The US equivalent pay act [5] expressed that the confirmation for unfriendly effect is the selection rate for any race, sex, or ethic gathering which is under four fifth of the rate for the group with the most astounding rate. The approached has been reached out to statistical significance of the removed patterns of discrimination [6] and the reason about steady activity and nepotism [7]. In [8] the system has been executed as a oracle based tool. Right now the methodology of discrimination technique is considered as every guideline freely for processing the discrimination without considering different rules or the connection among them. Besides in this paper we additionally consider the connection among guidelines for the disclosure of discrimination which was in light of the presence or nonexistence discriminatory features.

The preventing of discrimination is the other significant antidiscrimination point in the data mining which subsists of incited example that don't prompt discriminatory decisions regardless of the fact that the first preparing data sets are preferential. The accompanying three methodologies are conceivable;

a). Preprocessing: Transform the source information in such a route, to the point that the prejudicial predispositions contained in the first information are evacuated so that no unreasonable choice standard can be mined from the changed information and apply any of the average data mining algorithms. The preprocessing methodologies of data transformation and hierarchy based generalization can be customized from the privacy preservation literature. [9], [10] accomplished a controlled bending of the training data from which the classifier is found out by making insignificantly meddlesome modification prompting a impartial dataset. The preprocessing methodology is helpful for application in which a dataset should to be reported or in which data mining should to be performed by the third parties.

b). In processing: The information mining algorithm is redesigned in such way that the unfair decision rule does not contained by the resulting models. For instance, for cleaning the discrimination from the first data set is proposed in [11], however which the nondiscriminatory limitation is inserted into a decision tree learner by modifying its gashing criterion and pruning procedures amid a novel leaf relabeling methodologies. However, it is open that in-processing discrimination avoidance systems must anticipate on new specific reason data mining algorithms.

c). Post-preparing: Rather than cleaning the honest to goodness data set or changing the data mining algorithm procedure of post-transforming changed the resulting data mining models. Case in point in [6], a confidence adjusting approached is proposed for the classification rule construed by the CPAR algorithm. The power to distribute the information is not given by the post-processing. They distributed just the altered information mining models. Since the methodology of data mining can be performed by information holder only.

The proposed framework concentrates on the avoidance of discrimination by utilizing the post-preparing methodology. As the existing framework concentrate on the preprocessing methodology. One of the drawbacks of the framework which was proposed by the preprocessing methodology is that the technique does not say the data lost amid the discrimination prevention. While in our methodology the subsequent data mining model is changed as opposed to cleaning the first data set or changing the data mining algorithm.

The CPAR calculation keeps the discrimination by utilizing the post processing methodologies. In this methodology as opposed to cleaning the first data set, post-processing methodology change the data mining model. Killing the discriminated attributes from the database for discrimination prevention. The proposed algorithm utilized the Apriori or FP-growth algorithm for getting the incessant result, these association rule mining algorithm create the complete set of association rule and attain to the high order precision than the traditional classification methodologies.

III. BACKGROUND

In this section we will discussed about the background knowledge essential for this paper. Initially we will remember some basic definition related with the data mining. After that we elaborate on measuring and discovering discrimination.

1. Basic Definitions:

- Data set is a collection of data objects and their attributes. Suppose DB is the original database.
- An item is characteristic along with its value, for example Nationality – Indian.
- An item set Y is collect than of more than items, for example {Nationality – Indian, State - Gujarat}
- Classification rule is an expression Y→ C, where C is a class item and Y is an item set.

10.15680/ijircce.2015.0304072

- Support of an item set, supp(Y), it is the division of records which contains the item sets Y. we can say that rule Y→ C is completely sustain by the record if Y and C both are present in the record.
- Confidence of a classification rule, conf (Y→ C) evaluates the how frequently the class item C appears in record which contains Y. Therefore, if supp (Y) > 0 then,

- $conf\ (Y\rightarrow C) = \frac{supp\ (Y,C)}{supp\ (Y)}$ (1)
- Frequent classification rule is a classification rule with the support and confidence value which is greater than the specific lower bounds. In which support is the measure of statistical significance and confidence is the measure of strength of rule.

2. PD and PND Classification Rule:

Assume DC be the arrangement of predetermined discriminatory items in the DB. FP is the frequent classification rule which fall into the accompanying classes**:**

1. Classification rule Y→ C is PD when Y = A, B with A is the subset of DC a nonempty discriminatory item set and B is the nondiscriminatory item set.
2. Classification rule Y→ C is PND where Y = D, B is a nondiscriminatory item set.

The expression "potentially" implies a PD manages presumably prompt discriminatory decision. Subsequently, for evaluating the direct discrimination potential a few measures are required.

3. Preprocessing Approach:

The method of preprocessing approach for prevention of direct and indirect discrimination is split up into two phases:

- Measurement of discrimination:

Direct and indirect discrimination detection contains obtaining the alpha discriminatory rules and redlining rules. At first, PD and PND tenets are created on the premise of unfair things in the database DB and FP the regular classification principle. After that by utilizing the direct discriminatory measures and the oppressive edge the direct separation is measured by getting the alpha discriminatory rules with the PD rules. After that, same as the direct discrimination, indirect discrimination is measured by getting the redlining guidelines with the PND rules joining with the background knowledge, utilizing an indirect discrimination measures and the discrimination threshold.

- Transformation of discrimination:

Changing the original database DB in such path that to remove direct or indirect discriminatory, together with slightest effect on the data and on legitimate decision rules, so that no unjustifiable decision rules can be mined from the exchange database.

IV. PROPOSED APPROACHES

1. Algorithm Strategy:

In this paper we are examined about discrimination prevention by post processing methodology. We presented CPAR algorithm for keeping database from the discrimination rules and keeping up the nature of database. CPAR algorithm is the sort of association classification system. The algorithm consolidates the points of interest of both association classification methods and traditional rule based classification. As contrasting and the current algorithm of association classification technique like First Order Inductive Learner (FOIL) and Predictive Rule Mining (PRM), CPAR is more productive than these algorithms. The essential distinction between those calculations is the procedure of rule generation. FOIL at the time producing the tenet loses some vital rule. PRM beats the hindrance of FOIL, it remove the standards yet with the expense of repetition. CPAR utilized the idea of PRM for creating more standards with some repetitive rule, however it can test more than one ascribe at once to umpire whether the characteristics can issue some valuable rule or not. As contrasting and PRM, CPAR obliged more rules and less calculation. There are a few stages which was requires at the time of usage of CPAR algorithm. The steps are as per the following:

i. Generation of Rules
ii. Estimate accuracy of rules
iii. Classification and Result analysis

CPAR take up greedy algorithm for producing the guidelines specifically from the training data. For avoiding issue of missing vital rules, CPAR create and test a greater rule of standards than traditional rule based classifier. The algorithm utilizes the normal exactness for computing the every rule and uses the best k rule in expectation for evading the issue of over fitting. Like PRM, CPAR manufactures rule by including the literals one by one. Along these lines by creating the capable rule this algorithm is utilized for discrimination counteractive action by post processing methodology.

CPAR chooses various attributes if those attribute have comparable best pick up, this was finished by relating the Gain_Similarity_Ratio and by assessing the minimum gain. Binary value dataset is taken as the input by the CPAR and produces CARs. The algorithms additionally require increase steady which is entered by the user, rot component and Total_Weight_threshold. The following information is as connection rundown of principles requested by Laplace accuracy.

V. EXPERIMENTAL RESULT

i) Dataset:

For executing the algorithm and assessing the effectiveness of algorithm we obliged the dataset. We utilized adult data set, this information set subsists of seven attribute and number of records. The dataset comprises of general data about the person : sex, work class, education, marital status, Race, native country, salary. We utilized train dataset, the undertaking connected with the dataset of adults, in which the individual makes the pay more than fifty thousand or not. Number of rule is produced from this dataset and the discriminated rules are kept from the dataset by utilizing the CPAR algorithm.

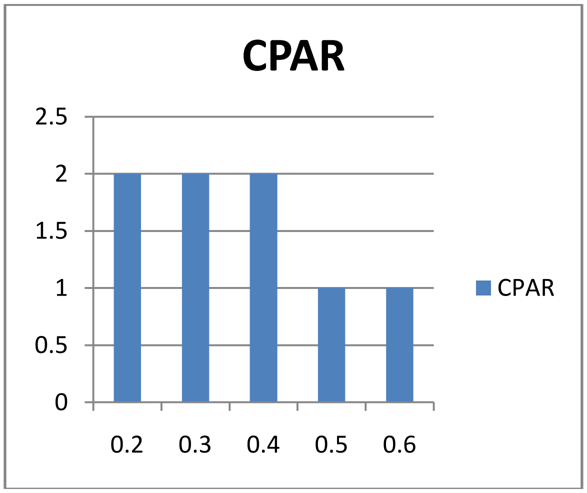
ii) Results:

We used adult dataset for implementing the proposed system. The adult dataset contains discriminated rules; in the following results user may input the k integer value for generating the strong rules for generating the strong rules and user also enter the TWT value. In the below table, the CPAR rules generated for different TWT values. The graph is plot from different TWT values. The resultant graph and table are as follows:

TWT	CPAR
0.2	2
0.3	2
0.4	1
0.5	1
0.6	1

Table 1: CPAR rules generated for different TWT

In the accompanying chart we will demonstrate the examination between CPAR rules generated for different TWT values. In this diagram x axis is different TWT values and y axis is rules generated for those values. From the chart it is presume that the post preprocessing system is more effective for the discrimination avoidance than the preprocessing technique. By utilizing the CPAR calculation for discrimination aversion the strong rule are created with no discriminated items. It conquers the drawbacks of preprocessing technique. Post preprocessing strategy keeps up the nature of dataset and keeps the dataset from discriminated rules.



Graph 1: Rules generated by different TWT values.

VI. CONCLUSION

As discrimination is an essential issue of data mining. The reason for this paper was to grow new post-processing discrimination prevention. As future work we are investigating measures of discrimination not the same as the ones considered in this paper alongside security protection in data mining. Proposed CPAR Algorithm attains to high exactness and efficiency, which can be credited to the accompanying recognized highlights: First, it utilizes greedy approach as a part of guideline era, which is significantly more proficient than creating all candidates rules. Second, it utilizes an dynamic programming way to deal with repeated calculation in rule generation. Third, it chooses various literals and fabricates multiple rules at the same time. Fourth, it uses anticipated that exactness would evaluate rules, and uses the best k rule in forecast. Proposed calculation speaks towards approach towards efficient and high quality, by actualizing proposed algorithm which shows preferred effectiveness over FOIL and PRM and utilization to anticipate discrimination aversion in post-preprocessing.

REFERENCES

[1]. 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.

[2].European Commission, “EU Directive 2006/54/EC on Anti- Discrimination,” http://eur-lex.europa.eu/LexUriServ/ LexUriServ.do?uri=OJ: L:2006:204:0023:0036:en:PDF, 2006.

[3]. 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.

[4]. 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.

[5]. United States Congress, US Equal Pay Act, http://archive.eeoc.gov/epa/anniversary/epa-40.html, 1963.

[6]. 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

[7]. 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.

[8]. 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.

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

[10]. F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 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]. S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Rule Protection for Indirect Discrimination Prevention in Data Mining," Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11), pp. 211-222, 2011.

[13]. WalidAdlyAtteya, Keshav Dahal and M.AlamgirHossain, "Distributed Bit Table multi agent Association Rules Mining Algorithm", Springer-Verlag, KES 2011, Part I, LNAI 6881.

[14] V. Verykios and A. Gkoulalas-Divanis, "A Survey of Association Rule Hiding Methods for Privacy," Privacy-Preserving Data Mining: Models and Algorithms, C.C. Aggarwal and P.S. Yu, eds., Springer, 2008.

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

BIOGRAPHY

Arati Shrimant Mote is persuing PG in Computer Science and Engineering from Matsyodari Shikshan Sanstha's College of Engineering and Technology, Jalna, She has completed B.E. in Computer Science and Engineering from Marathwada Institute of Technology in 2010. She has 2.7 years industrial experience as Oracle developer.

Prof. G. P. Chakote is working as Professor at Matsyodari Shikshan Sanstha's College of Engineering and Technology, Jalna, Maharashtra, India

10.15680/ijircce.2015.0304072