INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

DIRECT AND INDIRECT DISCRIMINATION PREVENTION AND RULE GENERATION IN DATA MINING

MS BHAGYASHRI SAWANA[1], PROF. P. K. BHARNE[2]

Dept. of Computer Sci, SSGMCE SHEGAON.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract: In social and economic science, discrimination is the subject which has been extensively studied in data mining. Discrimination can be categorized into direct and indirect. Decisions based on sensitive attributes are termed as direct discrimination and the decisions which are based on non-sensitive attributes are termed as indirect discrimination which is strongly correlated with biased sensitive once. There are many new techniques propose for solving discrimination prevention problems by applying direct or indirect discrimination prevention individually or both at the same time. New metrics to evaluate the utility were proposed and are comparing with approaches. The propose work discusses how privacy preservation and prevention between discrimination is implement with the help of post processing approach. The Classification Based on Multiple-class Association Rules (CMAR) is a kind of association classification methods which combines the advantages of both associative classification and traditional rule-based classification which is used to prevent discrimination prevention in post processing.

Keywords: Discrimination, Direct discrimination, indirect discrimination, Rule Based classification etc.

Corresponding Author: MS BHAGYASHRI SAWANA

Access Online On:

www.ijpret.com

How to Cite This Article:

Bhagyashri Sawana, IJPRET, 2016; Volume 4 (9): 1025-1036

PAPER-QR CODE

INTRODUCTION

Discrimination is termed as the act of unequally treating people on the basis of their belonging to a specific group. For example individuals may be discriminated because of their gender, ethnicity or nationality… etc. Different decision making tasks leads to the discrimination.eg loan granting/denial in the banking application .Discrimination is classified into two types. They are direct and indirect discrimination. Direct discrimination occurs when decisions are made based on the sensitive attributes. Indirect discrimination occurs when decisions are made based on the non-sensitive attributes. The aim of data mining is to extract useful information, such as patterns and trends, from large amounts of data. Many governments are gathering large amounts of data to gain insight into methods 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 privacy sensitive. Examples are race, religion, gender, nationality, disability, marital status, and age, etc. This raises the question how privacy.In many scenarios, an information system is used for decision making tasks. Given a set of information items on a potential customer, an automated system decides whether the customer is to be recommended for a scheme or a certain type of life insurance. Such automatic decisions reduce the workload of the staff of banks and insurance companies, among other organizations. The decision making is takes place by the use of information systems based on data mining technology has attracted the attention of many researchers in the field of computer science. In a plethora of data mining techniques such as association/classification rule mining have been designed and are currently widely used for making automated decisions. Consequence, automated data collection

I. LITERATURE SURVEY

J. Domingo-Ferrer et al. (2011) have developed a paper for rule protection for the indirect discrimination prevention in data mining. The datasets are trained and developed to make the classification rules to be extracted. Indirect discrimination rules cannot be extracted from the trained dataset. (i.e.) the trained datasets are free from indirect discrimination. Datasets are modified if any indirect discrimination occurs. Standard data mining algorithms are used to prevent the indirect discrimination from the training dataset.

Mykola Pechenizkiy et al. (2010) have developed a paper for discrimination aware decision tree learning. The decision tree models leads to the lower discrimination than the other models but with a little loss in the accuracy. The decision tree models are effective at removing the

discrimination from the original datasets. The problem is the datasets are cleaned away for discrimination before the discovery of the classifier in the dataset[1][2].

Sara Hajian et al. (2011) have developed a paper for prevention of discrimination in data mining for intrusion and crime detection. Data mining algorithm are used to prevent the direct and indirect discrimination. The data set obtained is free from the discrimination. In addition to detect the discrimination intrusion fraud and crime is also detected in the given dataset[3].

A classification rule is a procedure in which the elements of the population set are each assigned to one of the classes. A perfect test is such that every element in the population is assigned to the class it really belongs. An imperfect test is such that some errors appear, and then statistical analysis must be applied to analyse the classification [4].

Apriori Algorithm Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis[5].

III RELATED WORK

GOAL AND OBJECTIVE

1. To survey different discrimination discovery methods in data mining

2. To comparing other methods with work done to get better results.

3. To identify how discrimination take place while mining data.

4. To reduce the information loss by maintain data quality

SCOPE OF STATEMENT

The change is continues process with this consideration the scope of work is deals with many future aspects while using base idea for single class scope is designed for multiple classes. Developing for not only in binary form but also others, for testing presence of discrimination.

MOTIVATION

While reading regarding new technologies data mining needs privacy for preventing negative trends or crime. While mining the large data or giving cyber security, mining itself creating discrimination. Part while mining in cyber security where computational intelligence technologies used for different decision making scenarios. Therefore giving new patterns, rule or models or security to detect or prevent data mining for discrimination in large data set. So reading continued and idea flashed in mind about improving it with different technologies. The types of in/direct discrimination and some rules, patterns algorithms from studying all these got an idea about topic given name as “Prevention Discrimination in Data Mining”.

PROBLEM STATEMENT

To develop a system of discrimination prevention with post processing approach for no loss of data from original dataset by Classification based on multi-class association rules (CMAR).

EXISTING SYSTEM

- 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 focus 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.

PROPOSED SYSTEM

The proposed method can handle more data and discriminate them with the help of rule protection and generalization method. Preprocessing approach is used here. Different possible methods are compared for both direct and indirect discrimination method.

Anti-discrimination methodology is introduced. Different measures of discriminating power of the mined decision rules are defined by the anti-discrimination. The unwanted memory space and the buffering memory are reduced. Discrimination free data models can be produced from the transformed dataset without seriously damaging the data quality

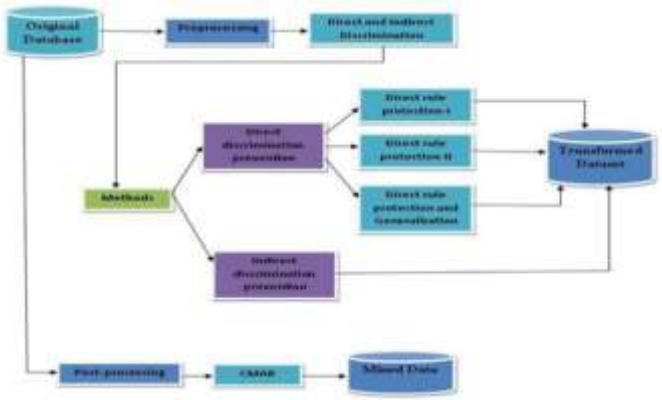

Fig 1 Proposed System Flow

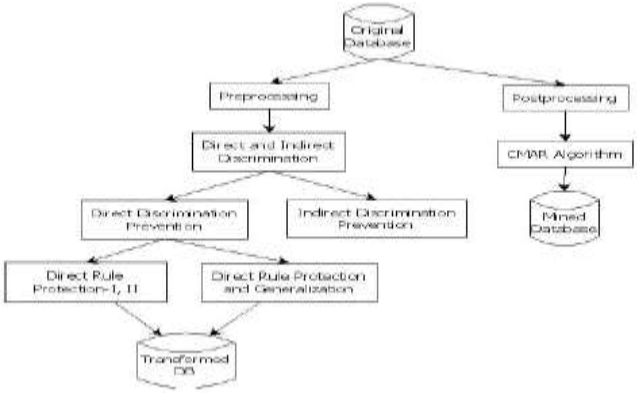

Fig 2 System Architecture

The above figure shows the overall view of system design. It gives idea how the work will take place to get desired output. In how many phases the shifting of data or operations to do. Here the main task is getting pre-processing method done and then go for post processing. Here in the pre-processing the Rule protection algorithms method1,2 are used and prevention is done. The post-processing method sing CMAR algorithm finishes the task.

MATHEMATICAL MODEL

Let DS as Discrimination System which is a set of some items, functions and other sets of operations and database.

DS = fs;DB;R;DI;NDI;D;ND; f; ^I_;DB^a; eg

Where,

s - Start of system

e - End of system

DB - original database

R - Set of rules

DI - Discriminatory items set

NDI - non-Discriminatory items set

f = set of functions like discover (), calculate () according conditions. = is range value [0, 10] where it decide the presence of discrimination if _ >s threshold value.

DB0 = Transformed DB.

Step 1:

Explained with general Example, DB = fa, v g

Where,

a = set of attributes

Ex. fName; Gender; Age;Race;Zip;Marital;Ref:g

v = set of only particular values

Ex. fany;M=F; Intno:; Pol=RP; 6digitno:; Y=Ng

Where,

a 2 DI,

DI = fName; Gender; Age;Race;Maritalg

And may be a 2 NDI,

NDI = fZip;Refg

Step 2: Relation of attributes with value is shown

Here a fg having value v and value v belongs to discrimination, then it considers as

DI.

As fa = v; andv 2 DIg ! PD

Here a fg having value v and value v belongs to non-discrimination, then it considers

as NDI.

As fa = v; andv 2 NDIg ! PND

Step 3:

Here, from R rules given for DB helps to find that attributes are discriminatory or not

f (check (a))=

fa2R!a2DIthendiscriminatory;a;v2D

a!2R!a2NDI0otherwisea;v2ND A classification rule X ! C is potentially discriminatory (PD) when X = A, B with A is the subset of DI a non-empty discriminatory item set and B a nondiscriminatory

item set.A classification rule X!C is potentially non-discriminatory (PND) when X = D, B is a non-discriminatory item set.From above the rules are also classified in PD and PND rules.

Step 4:

From above function can get main formulas for support and for confidence calculation formulas are [4]:Support (A ! B) = P (A B), where A, B¨I_DI (1)

Confidence(X! C) = support (X, C) / support (X) (2)

PROPOSED ALGORITHM

First phase: Rule Generation

Step1: scanning of database for finding item sets

Step2: descending sorting applied on attributes to construct frequent pattern tree

Step3: generation of subset of CARs based on F-list without overlapping

Step4: pruning the FP-tree by distributing class label

Step5: CR-tree is constructed to index the rules and potential sharing of rules and saving space since the rules that have common frequent items share the part of path

Step6: Pruning rules with high and low confidence to select subset of rules based on database coverage

Second phase: classification based on multiple rules (CMAR).

Step1: dividing the rules into groups according to class labels

Step2: measuring the combined effect of each group to compare the strength of the groups, is very tough, CMAR adopted weighted X2 measure.

MODULE DESCRIPTION

Data Preprocessing

Initially user uploads the dataset or we used trace the real time dataset from the ecommerce feedback website. This dataset contain seven attributes with 200 records. This dataset is in the form of ARFF format. Rules are generated in this stage. Apriori algorithm is used for generating the frequent rules. The apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules.

PRE PROCESSING METHOD

In this module user prevent the dataset from discrimination by using preprocessing method. In this stage initially user provides the alpha value and PD and PND rules are generated. After implementing the algorithm for preventing discrimination, we implement four algorithms. This is direct rule protection I, II, Direct Rule Protection and Rule Generalization, Indirect Rule Protection.

POST PREPROCESSING METHOD

In this module use prevent data from discrimination by using post processing method.

We used CMAR algorithm for post processing method. This algorithm overcome the disadvantage of the existing system, i.e. maintain the quality of the dataset by successfully removing the discrimination of the dataset.

RESULTS AND DISCUSSION

1. File Processing

Here, data set generate frequent item set with the help of apriori mine rule. The transaction file is generated which is in binary format.

2. Pre processing

In pre-processing with the help of α value, a potential discriminated and non-potential discriminated are checked.

3. Direct rule protection algorithm

In direct rule protection algorithm, each α discriminatory rule into an α protective rule, based on the direct discriminatory measure. And we get transformed file of original file as an output table.

4. Rule Generalization

It is the another data transformation for direct discrimination prevention method. It is based on the fact that if each α discriminatory rule in the database of decision

TECHNOLOGY AND ASSOCIATED PLATFORM

Hardware Specification

- Hard Disk : 80 GB
- RAM : 512 MB
- Processor : Intel Pentium 4 and above

Software Specification

- Technology Used: Core Java.
- Tools : JDK 1.5 or above, Netbeans
- Operating System : Windows XP or above

DATASET

For implementing the algorithm and calculating the effectiveness of algorithm we consider the dataset. Here grown-up data set is used; this data set is with seven attributes and number of records. The dataset contains general information about the individual: sex, training, marital status, race, nation, compensation. We utilized train dataset, the errand associated with the dataset of grown-ups, in which the individual makes the compensation more than fifty thousand. The discriminated standards are prevented from the dataset by using the CMAR algorithm.

EVALUATION RESULT

In this module, have to evaluate the whole result and also displayed in table. The above mentioned methods have to evaluate and finally they have to present with the parameter alpha which is the fixed value.

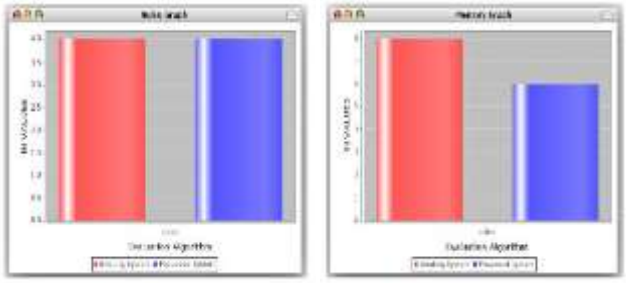

Figure 3: Rules Graph

Figure 3 Represent the comparison of existing system and proposed system in terms of memory utilization. Proposed system utilizes less memory than existing system.

OUTPUT

IV CONCLUSION

Now, we have clear idea about discrimination. The influence of discriminative behaviors has been the subject of studies in economics, law and social sciences. Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. The thing is obvious that most people do not want to be discriminated because of their gender, religion, nationality, age, and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. facilities. Here we have studied many ways to discover discrimination and handle with different technologies, algorithms, methods, rules and so on. The purpose of this study is to develop a new pre-processing discrimination prevention methodology including different data transformation methods that can prevent and don't make any information loss with importance attributes which are helpful to make decisions.

FUTURE ENHANCEMENT

The main approach is to finding out what cause the discrimination in mining as opinion mining has different perspective age, sex, area, and in E- Commerce these are totally different like quality, place, delivery time etc. Here the main problem of dataset in binary input so changing it is a new task to expand. The study is done on this topic is having different aspects. It may be difficult to identify attributes where not present but if such data cause discrimination then?? The future work is always for better performance also the new technologies will be there to handle this. Main task is finding discrimination or factors of E-commerce causing it are important.

V REFERENCES

1. S. Hajain, J.Domingo Ferrer, and A.Martinez [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," Uri=OJ:L:2006:204:0023:0036:en:PDF, 2006.
5. S. Hajian, J. Domingo-Ferrer, and A. Martı´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. 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.
7. F. Kamiran and T. Calders, "Classification without Discrimination," Proc. IEEE Second Int'l Conf. Computer, Control and Comm.
8. D. Pedreschi, S. Ruggieri, F. Turini, ˆa Discrimination-aware data miningˆa, in Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
9. D. Pedreschi, S. Ruggieri, and F. Turini, ˆaIntegrating Induction and Deduction for Finding Evidence of Discrimination,ˆa Proc. 12th ACM Intˆal Conf. Artificial Intelligence and Law, pp. 157-166, 2009.
10. S. Ruggieri, D. Pedreschi, F. Turini, "DCUBE: Discrimination Discovery in Databases," Proc. ACM Int'l Conf. Management of Data (SIGMOD '10), pp. 1127-1130, 2010.
11. S. Ruggieri, D. Pedreschi, F. Turini, "Data Mining for Discrimination Discovery", ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.
12. Andrea Romei, Salvatore Ruggieri , Franco Turini, " Discrimination Discovery in scientific Projects : A Case Study", Expert Systems with Applications 6064ˆa6079, Elsevier Ltd, 2013.