International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-3 Issue-3, February 2014

A Novel Approach to Prevent the Discrimination in Data Mining

R. Kayalvizhi, S. Sarika

Abstract-Automatic data collection has become the most wanted method in the banking sector to make automatic decisions like loan granting/denial. The discriminations in the dataset will lead to take the decisions in the partiality manner. The discrimination can be either direct or indirect discrimination. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on non-sensitive attributes. To overcome the partiality decisions the proposed system produces the anti-discrimination methodologies. The anti-discrimination methodologies prevent the discriminative decisions in the dataset. The proposed system prevents the discrimination without affecting the data quality.

Index Terms- anti-discrimination. Rule protection and rule generalization.

I. 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 but they are strongly related to the sensitive attributes. To overcome the partiality decisions anti-discrimination method is introduced

Anti-discrimination laws have been adopted by many democratic governments. Some examples are The Caste Disabilities Removal Act, Hindu Succession Act 1956, Scheduled Caste and Scheduled Tribe (Prevention of Atrocities) Act. Several data mining techniques have been adapted with the purpose of detecting discriminatory decisions. Anti-discrimination also plays a vital role in the cyber security to detect the intrusion and crime detection.

The main contribution of this paper are as follows (1) Detect the discrimination in the given dataset (2) Prevent the discrimination without affecting the data quality (3) Remove the discrimination by anti-discrimination methodologies and preserve the data quality (4) large amount of data's can be discriminated with the help of anti-discrimination methods

Rule protection and rule generalization algorithm are mainly used to generate the discriminate decisions.

Manuscript published on 28 February 2014.
* Correspondence Author (s)
R.Kayalvizhi, Department of Computer science and engineering, Sathyabama University/Chennai, India.
S.Sarika, Department of Computer science and engineering, Sathyabama University/Chennai, India.

The algorithm measures the discrimination with the help of support and confidence values produced by them. The algorithm helps to produce the data which are free from discrimination. The data quality is also protected by anti-discrimination method.

II. RELATED ARTICLES

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. The accuracy problem and the discrimination problem are solved with the help of knapsack algorithm. The knapsack algorithm discovers both the discriminatory and non-discriminatory data's. The resulting data's are free from the discrimination with the help of algorithms and decision tree models.

S.Ruggireri et al. (2010) have developed a paper for DCUBE: Discovery of discrimination from the databases. The DCUBE system discovers the direct and indirect discrimination from the given datasets. DCUBE system generally uses the classification rule extraction and analysis for the discovery of discrimination. Several users use the DCUBE method for the discrimination discovery. Query snippets and rule extraction are the main features for the discovery of discrimination and analysis of the process.

Toon` calders et al. (2010) has developed a paper for discrimination free classification with the modified naive Bayesian approach. Experiment is done with both artificial and real life dataset. Three different approaches are made in the proposed system to make the naive Bayesian classifiers free from the discrimination. The Bayesian classifier removes the discrimination and the rules causing the discrimination from the given data set.

Faisal Kamiran et al. (2010) have developed a paper for classification with no discrimination with respect to the preferential sampling.

The sampling scheme is introduced to make the data free from the discrimination. This makes the data to be less intrusive. The results are compared with the stable and unstable classifiers. Discrimination level is reduced by maintaining the high accuracy level in the data's. Data set obtained is free from the discrimination with the high accuracy level.

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.

III. EXISTING SYSTEM

The existing system is effective at removing the direct and indirect discrimination in the original dataset and preserves the data quality. The existing system does not require the standard data mining algorithm. They generally based on the classification rules of inductive part and reasoning on them the deductive part on the basis of the discriminative measures. Discrimination prevention methods are used in terms of the data quality and discrimination removal methods are used for both direct and indirect discrimination.

Drawbacks of the existing system are (1) it takes more time to handle the decision tree. (2) It will not handle more data and cannot predict attributes (3) Mining data's are not trustworthy and system cannot handle more amount of data

IV. 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.

In Proposed system more data can be discriminated. Discrimination is main thing of this process by the way of this process more people can serve without any partiality. Nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion.

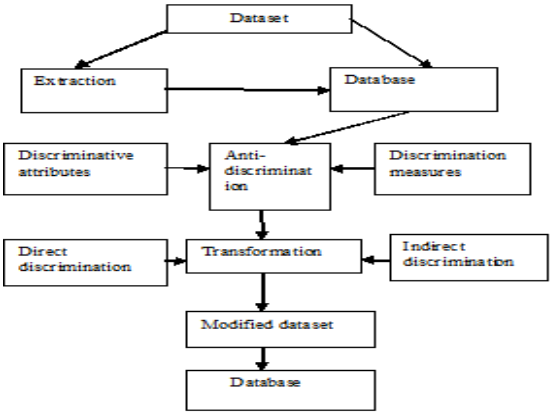


Fig 1.1 proposed block diagram

. [1]It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text

The block diagram of the proposed method was shown in figure 1.1 first the data set are collected. The dataset collected are extracted to the available format and stored in the database. Then the discriminative sensitive attributes are modified into anonymized attributes. Anonymized attributes means it does not behave like sensitive attributes. For example gender, race, religion and nationality are converted into anonymized attributes. Then the anti-discrimination method is introduced.Anti-discrimination methods checks the discrimination attributes and measures the discrimination with the help of rule protection and generalization algorithm. The algorithm generates the support and confidence values for the direct and indirect discrimination. Decisions are made on the attributes by specifying the personal score for each attributes. The transformed dataset is free from both direct and indirect discrimination. The modified dataset is free from discrimination and data quality is also maintained. Finally discrimination free data's are stored in the database

VI. ALGORITHM USED

Rule protection and rule generalization are used to generate the discriminatory values. The algorithm produces the support and confidence values for the discrimination measures. The output of the algorithm is the transformed dataset which is discrimination prevented data set. Rule protection algorithm generally produces support and the confidence values. Rule generalization is generally used to find the relationship between the rules instead of the discrimination measures. Both rule protection and rule generalization are effective at removing the discrimination in the original dataset.

V. IMPLEMENTATION

The implementation part is mainly explained to prevent the discrimination in the dataset and to maintain the data quality for the given dataset.

A. Data Analysis:

Data analysis is to gather the data from the external disk. Dataset contains real life dataset and synthetic dataset. First of all we have to check if all the attributes are placed in a correct manner if any null values are present then those dataset attributes cannot be processed by the metric and other computation process. Data analysis is generally termed as the process of gathering and analysis of dataset individually in a given two dataset.

B. Utility Measures:

Utility measures are taken to remove the discrimination on the given dataset. Dataset are analyzed with certain measures to remove the discrimination from the specified data's. Indirect discrimination removal and measuring data quality of that process are computed by the mathematical functionality like metric and rule protection and rule generalization. With the use of these techniques and algorithm records are filter in short time.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

C. Transform The Source Data:

The purpose is to transform the original data to remove direct or indirect discrimination from the given data set. To attain this process algorithm should be developed to specify which records should be changed during the transmission process. Discrimination data's are specified and changed during the data transmission.

D. Modifying Discriminatory Methods:

In this modified data's are converted into anonymized data. Anonymized data's means it does not behave like sensitive attributes but this data's can be processed. In this module modification are done on the sensitive attributes like gender, race, religion, sex, marital status and so on to anonymized the data's. In this module administrator can make the data to free from the sensitive attributes.

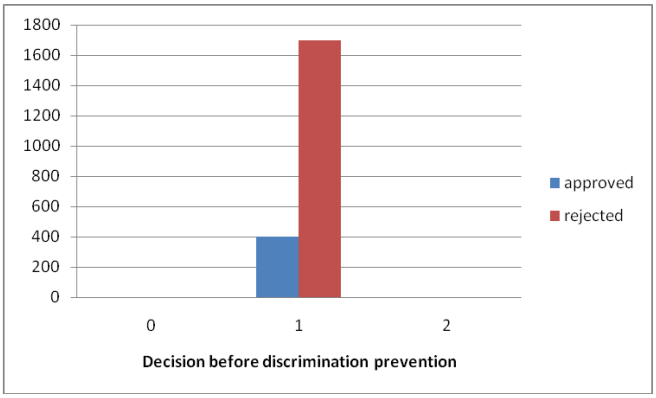
E. Decision Making:

Decisions could be depend on the attributes like gender, race and religion and so on. Each user gets score for their personal attributes in direct discrimination. Indirect discrimination can be done with the help of anonymized data's. Decisions should be done with the help of algorithm. The resulting data's are free from the discrimination removal and the data quality is maintained.

VII. PERFOMANCE EVALUATION

The proposed system is effective at removing the direct and indirect discrimination from the original dataset. Anti-discrimination method was introduced in the proposed method for effective discrimination method. The existing system provides more way to the discrimination approach than the proposed system

.



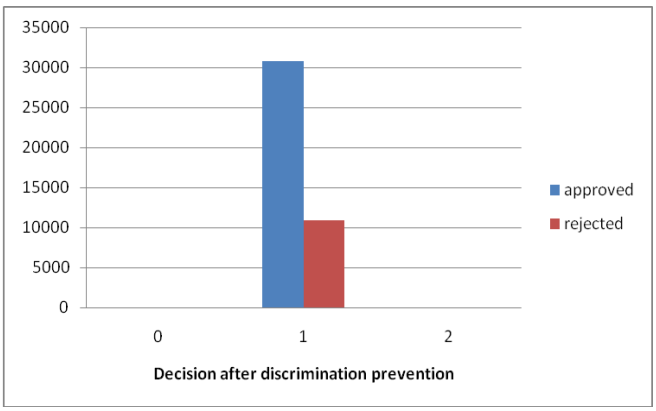


Fig 1.2 performance evaluation of proposed and existing System

The above figure1.2 shows how the proposed and existing systems are performed.

In the existing system due to the discrimination approach more number of data's are rejected in the given dataset.
In proposed method Anti-discrimination approach (removal of discrimination from the original dataset by anonymzing the attributes) is introduced. Less number of dataset is rejected in the proposed method due to the effective discrimination approach.

VIII.CONCLUSION

This paper presents a new pre-processing discrimination prevention method. Different transformations are used for the discovery of discrimination. The process measures the discrimination and identifies the categories by decision-making processes. Discrimination-free data models can be produced from the transformed data set without seriously damaging the data quality. More data's can be handled and the system result is trustworthy.

REFERENCES

[1]. S.Hajain, J.Domingo-Ferrer, and A.Martinez-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, and 2011.
[2]. F.Kamiran, T.Calders and M.Pecheninzkiy,"DiscriminationAware Decision Tree Learning", Proc.IEEE Int'l Conf.Data Mining (ICDM'10), pp.869-874, 2010
[3]. 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.
[4]. 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.
[5]. 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
[6]. 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.P.N.Tan,M.Steinbach and V.Kumar, Introdiction to Data Mining.Addison-Wesely,2006
[7]. S.Hajjain.J.Domingo-Ferrer and A.Martinez-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

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)