International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 4 Issue 11 ǁ November. 2016 ǁ PP.56-60

Algorithmic Analysis of Direct and Indirect Prevention Measure with Modular Ulift Measures

Pooja Shivcharan Rawte
Prof. Pragati Patil
Mtech CSE (Final Year)

ABSTRACT: *Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made on non sensitive attributes which are strongly related with sensitive. In this paper the system tackles discrimination in data mining by using methods like Rule protection, Rule generalization and eligibility of the client is done by using the Preferential algorithm. This is applicable for both direct and indirect discrimination. The cleaning training data sets in such a way that direct and/or indirect discriminatory decision rules are converted to nondiscriminatory classification rules is discussed in this paper we proposed novel approach algorithms which use data mining support and confidence and introduced ulift measures to find data quality.*

Index Terms*: Antidiscrimination, data mining, direct and indirect discrimination prevention, Rule protection, Rule generalization, Preferential sampling.*

I. INTRODUCTION

The word discrimination originates from the Latin word dis-criminare which means to “distinguish between”. In social sense, however, discrimination refers specifically to an action based on prejudice resulting in unfair treatment of people on the basis of their membership to a category, without regard to individual merit. As an example, U.S. federal laws[4] prohibit discrimination on the basis of race, color ,religion, nationality, sex, marital status and age in a number of settings, including: credit/insurance scoring(Equal Credit Opportunity Act):sale, rental, and financing of housing (fair Housing Act).

Concerning the research side, the issue of discrimination in credit, mortgage, insurance, labor market, education and other human activities has attracted much interest of researchers in economics and human science.

Data mining can be both a source of discrimination and a means for discovering discrimination [2]. Direct discrimination is evident when a person is treated unfavorable because of his or her personal attribute like sex, race, age, disability or parental status. This is straightforward and can affect the person being discriminated seriously. Indirect discrimination occurs when a certain policy appears to deal with all people equally but has the result of affecting a number of certain people where the requirement is unreasonable. The policy may seem harmless, but it has a discriminatory result against certain individuals.

Discrimination is defined by the process of unfairly treating people on the basis of their belonging to a specific group, namely race, ideology etc. This involves denying opportunities to members of one group that are available to other group of people. Here some antidiscrimination acts are used, which are laws designed to prevent discrimination on the basis of a set of attributes (e.g., race, religion, gender, nationality, disability and marital status) in various settings (e.g. credit and insurance, employment and training, access to public services, etc.).

Several decision-making tasks are there which lend themselves to discrimination, such as health insurances loan granting, education, and staff selection. In many applications, decision-making tasks are supported by information systems. Given a set of information items about a normal customer, an automated system decides whether the customer is to be recommended for a credit or a certain type of life insurance. This type of automated decisions reduces the workload of the staff of banks and insurance companies, among other organizations. The use of these information systems in data mining technology for decision making has attracted the attention of many persons in the field of computer applications. In consequence, automated data collection and data mining techniques such as association/classification rule mining have been designed and are currently widely used for making automated decisions. Automating decisions may give a sense of fairness classification rules (decision rules) do not guide themselves by personal preferences. However in a closer look, one realizes that classification rules are actually learned by the system based on training data. If this training data are inherently biased for or against a particular community (for example, foreign workers), then there is a chance of occurring discriminatory characteristics.

Discrimination is two types:
Direct and Indirect.

Direct discrimination consists of procedures or rules that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership.

Indirect discrimination consists of procedures or rules that, not explicitly mentioning discriminatory attributes, directly or indirectly generate discriminatory decisions. An example of indirect discrimination is refusing to grant mortgages or insurances in urban areas they consider as deteriorating although certainly not the only one. In this paper indirect discrimination will also be referred to as redlining and rules causing indirect discrimination will be called redlining rules [1].

II. RELATED WORK

Problem Statement

The existing system consists of preprocessing approach of just removing the discriminatory attributes from the data set. Although this would solve the direct discrimination problem, it would cause much information loss and it would not solve indirect discrimination.[1]

The propose system uses post processing approach to solve the above problem using Classification based on predictive association rules(CPAR)which is a kind of association classification methods which combines the advantages of both associative classification and traditional rule-based classification used to prevent discrimination prevention in post processing.

Solving approach and Efficiency issues:

Here the objective is to minimize the information loss by maintaining data quality. For performing this, there are various algorithms like Direct Rule Protection, Direct Rule generalization and Rule Protection, Direct and Indirect Discrimination prevention and CPAR algorithm. Out of which CPAR is more efficient than remaining algorithms in terms of information loss and data quality

Existing System

In Existing system the initial idea of using rule protection and rule generalization for direct discrimination prevention, but we gave no experimental results. We introduced the use of rule protection in a different way for indirect discrimination prevention and we gave some preliminary experimental results. In this paper, we present a unified approach to direct and indirect discrimination prevention, with finalized algorithms and all possible data transformation methods based on rule protection and/ or rule generalization that could be applied for direct or indirect discrimination prevention. We specify the different features of each method. Since methods in our earlier papers could only deal with either direct or indirect discrimination, the methods described in this paper are new.

Goals

Goal of the proposed system is to introduce a new algorithm for preventing discrimination in post processing approach using Classification based on Predictive Association Rules (CPAR) algorithm. Objective is to reduce the information loss by maintaining data quality through the use of various data transformation methods like Direct Rule Protection, Direct Rule Generalization, Direct and Indirect Discrimination Prevention and CPAR algorithm.

1. Misses cost (MC)
2. Ghost cost (GC)

Scope

Classification based on predictive association rules (CPAR) in data mining we can prevent discrimination using post processing approach. In Post processing the resulting data mining model is modified instead of cleaning the original data set or changing the data mining algorithm. As it is a challenging approach we can prevent discrimination by removing discriminatory attributes from datasets. To obtain FR(frequent Result) the CPAR algorithm will be used, which uses association rule mining algorithm, such as Apriori or FP-growth, to generate the complete set of association rules and achieve higher classification accuracy than traditional classification approaches. It combines the advantages of both associative classification and traditional rule-based classification.

Motivation

Discrimination is big issue in data mining. In order to tackle the discrimination problem various new techniques were proposed using direct and indirect discrimination prevention. For solving direct and indirect discrimination prevention preprocessing approach was proposed. Preprocessing approach considered to be more flexible one, it does not require changing the standard data mining algorithm. But this approach cannot guarantee that the transformed dataset is discrimination free. In order to make dataset discrimination free various other algorithms were proposed. One which is considering in this paper is post processing approach for removing direct and indirect discrimination from the dataset

Proposed System

We propose new utility measures to evaluate the different proposed discrimination prevention methods in terms of data quality and discrimination removal for both direct and indirect discrimination. Based on the proposed measures, we present extensive experimental results for two well known data sets and compare the different possible methods for direct or indirect discrimination prevention to find out which methods could be more successful in terms of low information loss and high discrimination removal. The approach is based on mining classification rules (the inductive part) and reasoning on them (the deductive part) on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. Proposed work is focus on to develop a new preprocessing discrimination prevention methodology including different data transformation methods that can prevent direct discrimination, indirect discrimination or both of them at the same time. To attain this objective, the first step is to measure discrimination and identify categories.

Module Description

Module Split up:

1. Pre-processing
2. Discrimination measurements
3. Utility Calculations
4. Rules Generation
5. Evaluation Result

Pre-processing

In this module, have to pre-process the dataset. Here the adult dataset is taken. Train and test dataset is considered. This data set consists of 15 attributes (class attribute). The prediction task associated with the Adult data set is to determine based on census and demographic information about people. The data set contains both categorical and numerical attributes. Although the Age attribute in the Adult data set is numerical. And they filter the incomplete records from the adult dataset.

Discrimination measurements

In the discrimination measurement, direct and indirect discrimination discovery includes identifying alpha-discriminatory rules and redlining rules. Firstly, based on predetermined discriminatory items in DB, frequent classification rules in FR are divided in two groups: PD and PND rules. Secondly, direct discrimination is measured by Identifying alpha-discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold (alpha).

Third, indirect discrimination is measured by identifying redlining rules among the PND rules combined with back-ground knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold (alpha). Let MR be the database of direct alpha-discriminatory rules obtained with the above process. In addition, let RR be the database of redlining rules and their respective indirect alpha-discriminatory rules obtained with the above process. In data transformation, the transformation of the original data DB in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data.

Utility Calculations

This module is used to measure the success of the method in removing all evidence of direct and/or indirect discrimination from the original data set; on the other hand, have to measure the impact of the method in terms of information loss (i.e., data quality loss). To measure discrimination removal, four metrics were used: Direct discrimination prevention degree (DDPD), direct discrimination protection preservation (DDPP), indirect discrimination prevention degree (IDPD), indirect discrimination protection preservation (IDPP).

Rules Generation

This module is deal with the key problem of transforming data with minimum information loss to prevent at the same time both direct and indirect discrimination. Firstly, when direct and indirect discrimination could simultaneously occur, depends on whether the original data set (DB) contains discriminatory item sets or not. Two cases arise: Discriminatory item sets did not exist in the original database DB or have previously been removed from it due to privacy constraints or for preventing discrimination. However, if background knowledge from publicly available data (e.g., census data) is available, indirect discrimination remains possible. In addition to direct discrimination, indirect discrimination might occur because of background knowledge obtained from DB itself due to the existence of non-discriminatory items that are highly correlated with the sensitive (discriminatory) ones. Hence, in this case both direct and indirect discrimination could happen. To provide both direct rule protection (DRP) and indirect rule protection (IRP) at the same time, an important point is the relation between the data transformation methods. Any data transformation to eliminate direct alpha-discriminatory rules should not produce new redlining rules or prevent the existing ones from being removed. Also any data transformation to eliminate redlining rules should not produce new direct alpha-discriminatory rules or prevent the existing ones from being removed. And have to generate the rules for discrimination discovery.

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.

Algorithm:

4 Basic definitions:

Some basic definitions related to data mining [17].After that, we elaborate on measuring and discovering discrimination.

• **A data set** is a collection of data (records)and their attributes. Let DB be the original data set.• An **item** is an attribute along with its value, e.g., Race = black.• An **item set** is a collection of one or more items, e.g.,{ Foreign worker = Yes, City =NYC}.

• **A classification rule** is an expression X ->C, where C is a class item (a yes/no decision), and X is an item set containing
no class item, e.g., {**Foreign worker =Yes, City = NYC-> Hire = no}**. X is called the premise of the rule.

• **The support** of an item set, supp(X), is the fraction of records that contain the item set X. We say that a rule X -> C is completely supported by a record if both X and C appear in the record.

• **The confidence** of a classification rule, conf(X)->C, measures how often the class item C appears in records.

Steps:

1) Perform Preprocessing to clean the data sets.
1) Select data sets, store into database.
2) Create cluster of data and delete row which having garbage values.
3) Shows the data value which preprocess.
2) Identify sensitive and non sensitive data attributes.
i) Find the support of sensitive attributes.

```
Support = Occurrence / Total support
Call procedure support ()
 malecour();
femalecour();
d=(float)a/(float)c;
e=(float)b/(float)c;
write(d);
write(e);
 return(d);
```

Above step calculate the support of attribute gender.maleoccur() and femaleoccur() function find the occurrence of male and female from datasets. Then it calls function support to calculate support gender.

3) Calculate confidence of all sensitive and non sensitive attributes.

```
call procedure Confisex()
{
r1=(float)e/(float)d;//confidencce of male -> female
write("confisex);
return(r1);
}
```

Above procedure calculate the confidence of male and female. Same as calculate the confidence of other attributes.

4) calculate the utility of measures lift and ulift to check the data quality of direct and indirect discrimination.

```
E20=(float)e18/e17;//support sensitive attributes
E21=(float)e19/e16;//support non sensitive attributes
  e23= e20+e21;  //bn7
  e31=e24+e25+e26+e27+e28+e29+e30;// q3=d+e+q+r+s+d1+e1+q1[k];
  e32=e16+e17+e18+e19;//bn4
  e39=e38+e37;//bn8
 e40=(float)e17/e18;//bn9
 e45=e41+e42+e43+e44;
 elift=(el1+el2+e13+e14+e15)/e23;//elift calculation to measure quality of data
```

Above step calculate the measures and according to it generates results.

III. CONCLUSION

The determination of this paper was to progress a new pre-processing discrimination prevention methodology including different data transformation methods that is used to prevent direct discrimination, indirect discrimination or both of them. To attain this objective, firstly, they measure discrimination and identify categories and groups of individuals that have been directly or indirectly discriminated in the decision-making processes; secondly, have to transform data in a proper way to remove all those discriminatory biases. Finally, discrimination-free data models can be produced from the transformed data set without seriously damaging data quality. Proposed algorithm used to measure the quality and help to prevent direct and indirect discrimination.

REFERENCES

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

[2]. T. Calders and S. Verwer, “Three Naive Bayes Approaches forDiscrimination-Free Classification, “Data Mining and KnowledgeDiscovery, 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:2006:204:0023:0036: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 CyberSecurity (CICS ’11),pp. 47-54, 2011.

[6]. S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, “Rule Protection for Indirect Discrimination Prevention in DataMining,”Proc. Eighth Int’l Conf. Modeling Decisions for ArtificialIntelligence (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 Discriminationby Preferential Sampling,” Proc. 19th Machine Learning Conf.Belgium and The Netherlands,2010.

[9]. F. Kamiran, T. Calders, and M. Pechenizkiy, “DiscriminationAware Decision Tree Learning,”Proc. IEEE Int’l Conf. Data Mining(ICDM ’10),pp. 869-874, 2010.

[10]. R. Kohavi and B. Becker, “UCI Repository of Machine LearningDatabases,” http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[11]. D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, “UCIRepository of Machine Learning Databases,” http://archive.ics.uci.edu/ml, 1998.

[12]. D. Pedreschi, S. Ruggieri, and F. Turini, “Discrimination-AwareData Mining,”Proc. 14th ACM Int’l Conf. Knowledge Discovery andData Mining (KDD ’08),pp. 560-568, 2008.

[13]. D. Pedreschi, S. Ruggieri, and F. Turini, “Measuring Discrimina-tion in Socially-Sensitive Decision Records,” Proc. Ninth SIAMData Mining Conf. (SDM ’09),pp. 581-592, 2009.

[14]. D. Pedreschi, S. Ruggieri, and F. Turini, “Integrating Inductionand Deduction for Finding Evidence of Discrimination,”Proc. 12thACM Int’l Conf. Artificial Intelligence and Law (ICAIL ’09),pp. 157-166, 2009.

[15]. S. Ruggieri, D. Pedreschi, and F. Turini, “Data Mining forDiscrimination Discovery,”ACM Trans. Knowledge Discovery fromData,vol. 4, no. 2, article 9, 2010.

[16]. 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.

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