Extracting Precise Data by Preventing Discrimination in Data Mining

R. Shagana, C. Nancy Nightingale

Abstract: ***The concept of classification is one of the most popular data mining tasks. The result of classification depends critically on data quality. There are some negative social perceptions about data mining which include potential privacy invasion and discrimination. Discrimination refers to the data set which contain unwanted data items. Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on non sensitive attributes. This project discusses how to clean training datasets and outsourced datasets in such a way that legitimate classification rules can still be extracted.***

Keywords: Data Mining, Redlining rules, Discrimination, Rule generalization, Rule protection

I. INTRODUCTION

Modern computers have made it so that every field of study is generating data at an unprecedented rate. Computers can process data in ways and speeds humans could never achieve. Data mining is the entire process of applying a computer-based methodology for developing knowledge from data. Data mining is an iterative process in which progress is defined by discovery, through either manual or automatic methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined ideas about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. discrimination is the prejudicial treatment of an individual based on their membership in a certain group or category. It involves denying to members of one group opportunities that are available to other groups.

Services in the information society allow for automatic and routine collection of large amounts of data. Those data are often used to train association/classification rules in view of making automated decisions, like loan granting/denial, insurance premium computation, personnel selection, etc. At first sight, automating decisions may give a sense of fairness classification rules do not guide themselves by personal preferences. One of the challenge in automated decision making is discrimination. Discrimination can be either direct or indirect. Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or unintentionally could generate discriminatory decisions. Indirect discrimination could happen because of the availability of some background knowledge, for example, that a certain zip code corresponds to a deteriorating area or an area with mostly black population.

Revised Manuscript Received on 30 March 2014.
* Correspondence Author

Ms. R.Shagana, Master of Computer Science and Engineering, IFET College of Engineering, Villupuram, India.
Ms. C.Nancy Nightingale, Master of Information Technology, IFET College of Engineering, Villupuram, India.

II. APPROACHES TO PREVENT DISCRIMINATION

The discovery of discriminatory decisions was first proposed by Pedreschi. The 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. Three approaches are conceivable in order to prevent discrimination namely preprocessing, In-processing, post processing.

A. Preprocessing

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.

B. Inprocessing

Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. For example, an alternative approach to cleaning the discrimination from the Original data set is proposed in whereby the non discriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy.

C. Post Processing

Modify the resulting data mining models, instead of cleaning the original data set or changing the data mining algorithms. For example, the post processing is used in confidence-altering approach for classification rules inferred by the CPAR algorithm.

III. EXISTING SYSTEM

The existing system make use of classification rules to make automated decision making in filed such as loan granting/denial, etc . The classification rule use the training data set to make decision which leads to problem called discrimination. Discrimination refers to the thing which consist of unwanted items in the resulting data set.

Discrimination problem occurs because of attributes that may be either sensitive or insensitive. Direct discrimination occur when the decision are based on sensitive attribute and indirect discrimination occur when decision made on insensitive attributes. In order to avoid discrimination several method has been proposed one of the method is discrimination aware decision tree learning.

A. Discriminatory And Nondiscriminatory Classification Rule

Let DI be the set of predetermined discriminatory items in database(DB), frequent classification rules(FR) into one of the following two classes,

1. A classification rule (X →C) is said to be potentially discriminatory (PD) when (X =A,B) with a non empty discriminatory item set and B is a discriminatory item set.
2. A Classification rule (X→C) is said to be potentially non discriminatory (PND) when (X = D,B) is an item set which is non discriminatory.

B. Direct and indirect discriminatory measure

The technique used to measure direct discrimination is extended lift (elift).
if A,B→C is a classification rule such that conf(B→C) >0. The extended elift
of rule is

$$\text{elift (A, B} \rightarrow \text{C)} = \frac{\text{conf (A, B} \rightarrow \text{C)}}{\text{Conf (B} \rightarrow \text{C)}}$$

The idea of using elift is to evaluate the discrimination of rule as the gain of confidence due to the presence of the discriminatory items in the premise of the rule.In addition to elift, two other measures slift and olift are used to measure the direct discrimination.
The indirect discrimination can be measured using the redlining rules

1. A PND Classification rule r:D, B →C is a redlining rule if it could yield discriminatory result by the combination of available background knowledge from the rule r1: A, B→D and r2= D,B→A, where A is discriminatory item set.
2. A PND Classification rule r:D, B →C is a nonredlining rule if it cannot yield discriminatory result by the combination of available background knowledge from the rule r1: A, B→D and r2= D,B→A, where A is discriminatory item set

C.Limitation Of Using Preprocessed Technique To Find Discrimination

Discrimination prevention methods based on preprocessing presents some limitation.

1. They attempt to detect discrimination in the original data only for one discriminatory item and based on a single measure.
2. They only consider direct discrimination.
3. They do not include any measure to evaluate how much discrimination has been removed and how much information loss has been incurred.

IV. PROPOSED SYSTEM

The proposed system make use of unsupervised algorithm that use two classifiers for duplicate detection namely weighted component similarity summing, support vector machine classifier. In Unsupervised Learning no formal training data is used to make decision Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. Unsupervised learning may be divided into two types of problems, data clustering and feature extraction. Data clustering, or exploratory data analysis, aims to unravel the structure of the provided data set. Feature extraction, on the other hand, often seeks to reduce the dimensionality of the data so as to provide a more ‘compact’ representation of the data set.

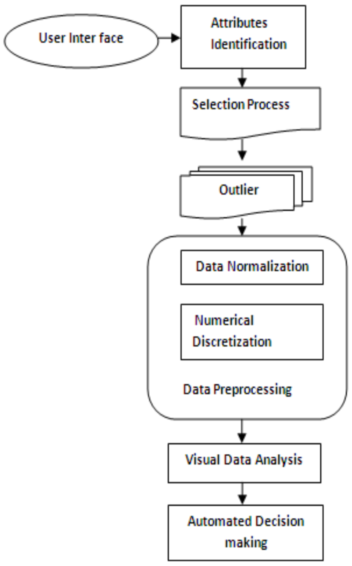

Fig1 System Architecture of Existing system

A. Weighted Component Similarity Summing Classifier

Classifier is the main algorithm of UDD system in identifying duplicates. Inputs to this classifier are the similarity vectors of record pairs from potential duplicates and non-duplicate sets. This classifier tries to find out the duplicates from non-duplicate and potential duplicate datasets. The output from this classifier is a duplicate dataset identified from the potential duplicates and non-duplicate sets.

B. Support Vector Machine Classifier

Classifier is a tool used to classify data. SVM uses a two-step process, training and classification. In the training step, labeled data is supplied to the classifier, labeling each record as either positive or negative. During the classification step, when the system is supplied with data to be classified, it classifies the record as either being positive or negative based on the training data.

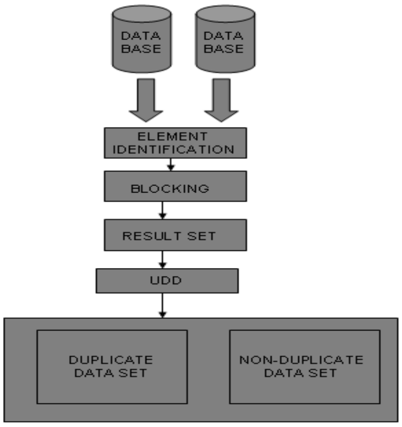

Fig 2. System Architecture of proposed system

C. Steps Involved In Proposed System

1. Analyze the problems in data-set
2. Preprocessing the data-set
3. Duplicity Detection using alliance rules
4. Detection of errors using q grams

Analyze the problems in data-set

A data warehouse is constructed by amalgamating and merging data from different sources. In this module we discuss the problems and errors related to the 'name' field in the data warehouse. A name field in a data warehouse faces the problem of duplicity due to semantic heterogeneity among the various sources of data. A name field is specific in its usage. The importance of name field explains the need of cleaning it in the data warehouse.

Preprocessing the data-set

In preprocessing the strings in the name field are converted into a numerical value which is stored in a file for ready reference. Two different data marts are considered such that a name from one data mart say DM1 has to checked and matched for duplicity with all the names in another data mart say DM2. The data sets taken for both DM1 and DM2 consist of name fields which are converted to numerical integer values. The converted integer values are stored in a file called Score and are called scores of the name.

Duplicity Detection using alliance rules

The algorithm for detecting duplicity in name filed of data warehouse is as follows

1. Take a name from DM1.
2. Determine the no of words in the name. Let it be denoted by N.
3. Calculate N+1 scores for the name each corresponding to a a word present in name. The (N+1)th score is score of the initials of name.
4. In Data Mart 2 (DM2), Cluster the names which have same value of N.
5. Calculate N+1 scores for all names in this cluster of DM2.
6. Match the last name scores Sn of Name in DM1 & Sn of each name in DM2 cluster and cluster all those names in DM2 that have same score value for Sn and further decrease the size of cluster
7. Now match S1 of name from DM1 & new DM2 cluster & Store the Further reduced cluster in file

Detection of errors using q grams

The score matching finally helps in detecting the duplicates. Till now Sn has been matched and using S1 the further algorithm will be followed for concluding the duplicity detection.

Case 1:-Perfect Match

a) Single entry match: In case of single entry there is no duplicity.

b) Multiple entry matches: In this case, match other scores of the name which are S2, S3 up to Sn-1.

Case 2:-No match

If S1 score does not match then match the score of initials

a) Same person

b) Different person with same initials

Case 3:- None of score matches from Sn to Sn+1 This can be due to two reasons:

a) That entry does not exist.

b) Entry exists with some errors in name

V. CONCLUSION

The concept of classification is one of the most popular data mining tasks. It is result of classification depends critically on data quality. Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. The conclusion of our project is challenging problem of discriminations. The purpose is, preferential sampling classifies the future data without Discrimination and High accuracy. It also addresses the problem of redlining. The final result is to find a good trade off between discrimination removal and the quality of the resulting training data sets and data mining models.

REFERENCES

[1]. J. Karlsteen, " Automation of meta data updates in a time critical environment", 2006.

[2]. P. Poonniah , Data Warehousing Fundamentals – A comprehensive guide IT professionals ,Ist ed., 81-265-0919-8, Glorious Printers: New Delhi , India, 2006.

[3]. B.Palace, " Data mining ", [Online],1996,http://www.anderson.ucla.edu/faculty/jason.frand/

[4]. C.Kelley, "Best Uses of data warehouse "http://www.itworld.com.[Online],2003, http://www.itworld.com/nl/db_mgr/,2003[5]. T. Redman,"The Impact of Poor Data Quality on the Typical Enterprise",Communications of the ACM, Vol. 41. 8. 02, 1998.

[6]. E.Rahm, H.H Do," Data Cleaning: Problems and Current Approaches", University of Leipzig,Germany.

[7]. A.Marcus,J.I.Maletic,"Automated Identification of Errors in Data Sets,TR-CS-00-02,University of Memphis,2002.

[8]. A.Marcus,J.I.Maletic,"Utilizing Association Rules For the Identification of Errors in Data",TR-CS-00-04,University of Memphis,2004.

[9]. K.Orr,"Data Quality and Systems Theory" Communications of the ACM, Vol. 41. 9, 2, 1998.

[10]. M.A.Hernandez, S.J.Stolfo, "Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem", 1998.