Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Evaluating Credit Card Dataset for Discrimination Prevention Using PSO Algorithm

Ms.T.Sasirekha*, Ms.C.Anoorselvi

Department of CSE,

V.S.B Engineering College, India

Abstract— Data mining is an increasingly important technology for extracting useful knowledge hidden in large collections of data set. Extracting information from some datasets (Adult, Credit card dataset) will give some personal details of the customer to the extracter. When we go for automatic extraction of data from the dataset collection using classification rule it will classify the results according to our conditions. If the training data sets are conditioned in what regards discriminatory (sensitive) attributes like gender, race, religion, etc., discriminatory decisions may be sensitive. Discrimination will 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 that is the background knowledge from the results extracted which is having some strong bias values. In this paper, we tackle discrimination prevention in data mining and propose new techniques applicable for direct or indirect discrimination prevention individually or both combined. The Credit Card Datasets are taken for making decisions to grant/deny loans based on the details we extract from it. We propose an algorithm Called Particle Swarm Optimization Algorithm (PSO) to produce an optimized result for decision making process. New measures are used to measure the discrimination values. The Optimized results will increase the speed and reduces the searching time.

Keywords : Data Mining, Discrimination, Anti Discrimination, Association Rule Mining.

I. INTRODUCTION

The Main goal of this work is to prevent the decision making process from discrimination, increase the speed of the model, reduce the searching time for the datasets, to choose best person for granting loan without any discrimination. In general choosing a Person for Giving Loan will be selected based on their income. But, the person getting less income will be able to repay the loan on time. That is they have some other sources of income. Based on the income and some sensitive attributes we may discriminate the person. There are two types discrimination. They are direct and indirect discrimination. Direct discrimination occurs when we take the decision based on sensitive attributes (e.g. race, age, nationality, sex, religion). Indirect discrimination occurs when decisions are made based on non-sensitive attributes. Indirect discrimination will be based on background knowledge from the extracted information. The non-sensitive attributes will make serious bias on the decision. Direct discrimination will group some of the people in a group and didn't give loan. Indirect discrimination will make a group based on their background knowledge. one non sensitive attribute is address of the person. A person residing in rural area will be rejected without considering their income details. In this paper, we consider the credit card data set for granting or denying loan and to prevent from both direct and indirect discrimination. we analyze the previous measures of discrimination and reduce the searching time and improve the speed and gives best results on the available datasets. Finally, we compare the proposed system with existing system and show the improvement in performance.

II. PREVIEOUS WORK

Data mining is the easily understandable and efficient mining technology which gives effective knowledge for the extracter in their process. Mining can be done with some efficient algorithm to give better results. Despite the wide deployment of information systems based on data mining technology in decision making, the concept of antidiscrimination in data mining did not receive much attention until 2008. Some proposals are oriented to the discovery and measure of discrimination. Others deal with the preventing the details from discrimination. the existing data transformation methods (i.e., rule protection and rule generalization (rg)) are based on measures for both direct and indirect discrimination and can deal with several discriminatory items. Also, we provide some measures for measuring the utility of the data in the dataset. Hence, our approach of discrimination prevention is broader than in previous work. Existing system introduced the use of rule protection in a different way for indirect discrimination prevention and we gave some preliminary experimental results. In this project, we present a unified approach to prevent from direct and indirect discrimination prevention, with protection algorithms and all possible data transformation methods based on rule and/ or rule generalization that could be applied for direct or indirect discrimination prevention. It specifies the different features of each method. A frequent classification rule is a classification rule with support and confidence greater than

respective specified low threshold bounds. Support is a measure of how frequently the item occurs in the dataset, whereas confidence is a measure of the derived rule's strength.

III. PROPOSED WORK

The two tasks undertaken in this paper provide the basis for the design of an information technology framework that is capable to direct and indirect information. The first task identifies and extracts informative sentences on credit card account holder, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between sensitive and non sensitive.

Particle Swarm Optimization Algorithm (pso) is used to solve the optimization problems**.** The particles fly through the problem space by following the current best particles. The algorithm is initialized with a group of random particles (solutions) and then searches for optima by updating their operations. in every updation, each particle is modified by following two "best" values.The first one is the best solution (fitness) it has achieved so far. This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the optimum value, obtained so far by any particle in the group. This best value is called as global best and it is denoted as gbest. When a particle takes part of the group as its similar characteristics neighbors, the best value is called as local best and is denoted as lbest. We will use this algorithm to provide an optimized decision.

IV. IMPLEMENTATION

A. A Proposal For Direct And Indirect Discrimination Prevention

In this section, we present our approach, with the data transformation methods that can be used for direct and/or indirect discrimination prevention.

A.1 The Approach

Our approach for direct and indirect discrimination prevention can be described in terms of two phases:

Discrimination measurement

Direct and indirect discrimination discovery includes identifying discriminatory rules and redlining rules. To this end, first, based on predetermined discriminatory items in db, frequent classification rules in FR are divided in two groups: pd and pnd rules. Second, direct discrimination is measured by identifying discriminatory rules among the pd rules using a direct discrimination measure (elift) and a discriminatory threshold. Third, indirect discrimination is measured by identifying redlining rules among the pnd rules combined with background knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold .Let MR be the database of direct discriminatory rules obtained with the above process. In addition, let RR be the database of redlining rules and their respective indirect discriminatory rules obtained with the above process.one of these measures is the

extended lift (elift):

Let A,B ->C be a classification rule such that Conf(B -> C) > 0. The extended lift of the rule is,

elift(A,B -> C) =conf(A,B >C)/conf(B->C)

The idea here is to evaluate the discrimination of a rule as the gain of confidence due to the presence of the discriminatory items (i.e., A) in the premise of the rule.

Data transformation

Transform 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.

B. System Modules

Discrimination Prevention model is to enhance the banking field and dataset without discrimination and it is subdivided in to five major modules. They are

1) User Interface Design
2) Preprocessing
3) Discrimination Measure
4) Decision Tree Learning
5) PSO Algorithm

1) User Interface Design

The goal of user interface design is to make the user's interaction as simple and effective as possible, in terms of providing user goals—which is often called user-centered design. User interface design facilitates finishing the task at hand without drawing unnecessary attention to it will be called as good user interface design. Graphic design may be utilized to support its usability. The design process must consider technical functionality and visual elements (e.g., mental model) to create a system that is not only operational but also usable and adaptable to changing user needs.

2) Preprocessing

Data pre-processing is an often neglected but important step in the data mining process. The redundant information present or noisy and unstable data, then knowledge discovery during the training phase is harder. Data formation and filtering steps can take considerable amount of processing time. Data pre-processing will have these following steps. They are cleaning, normalization, transformation, Integration, feature extraction and selection, etc. In the cleaning process we will clean the not applicable values in the dataset. The resultant dataset will contain only the relevant and effective values of their attributes. The product of data pre-processing is the final training set.

3)Discrimination Measure

Direct and indirect discrimination discovery includes identifying discriminatory rules and refusing rules. To this end, first, based on already decided discriminatory items in DB, frequent classification rules in frequent classification (FR) are divided in two groups: potentially discriminatory PD and potentially nondiscriminatory PND rules. Second, direct discrimination is measured by identifying discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold. Third, indirect discrimination is measured by identifying redlining rules among the PND rules combined with the hidden knowledge extracted from the dataset, using an indirect discriminatory measure (elb), and a discriminatory threshold. Let direct _-discriminatory MR be the database of direct discriminatory rules obtained with the above process. In addition, let RR be the database of redlining rules and their respective indirect discriminatory rules obtained.

4) Decision Tree Learning

Decision tree learning, used in data mining is as a predicting or guessing model which maps observations about an item to conclusions about the item's target value. More explanatory names for such tree models are classification trees or regression trees. In the classification or regression tree structures, leaves represent the labels of a class and branches represent conjunctions of features that lead to those class labels. It is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in fig2. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

5)Pso Algorithm

PSO learned from the scenario and used it to solve the optimization problems. In PSO, each single solution is a "bird" in the search space. We call it "particle". All of particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. In every iteration, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called gbest. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called lbest.
The pseudocode of the procedure is as follows:

```
For each particle
   Initialize particle
END

Do
   For each particle
      Calculate fitness value
      If the fitness value is better than the best fitness value (pBest) in history
         set current value as the new pBest
   End

 Choose the particle with the best fitness value of all the particles as the gBest
   For each particle
      Calculate particle velocity according equation (a)
      Update particle position according equation (b)
   End
```

The entire process of the algorithm is shown in the fig1. First, all the particles available in the search space is initialized. Second, fitness of each and every particle is calculated based on their process. Third, compare the best fitness value with the particles fitness value. Then fix the pbest value and calculate the gbest value based on the particles available in the group. The particle which is available in the entire group will have some best fitness value. The fitness value which is best among their local neighbors is calculated as lbest. The best value will be changed on each and every iteration.

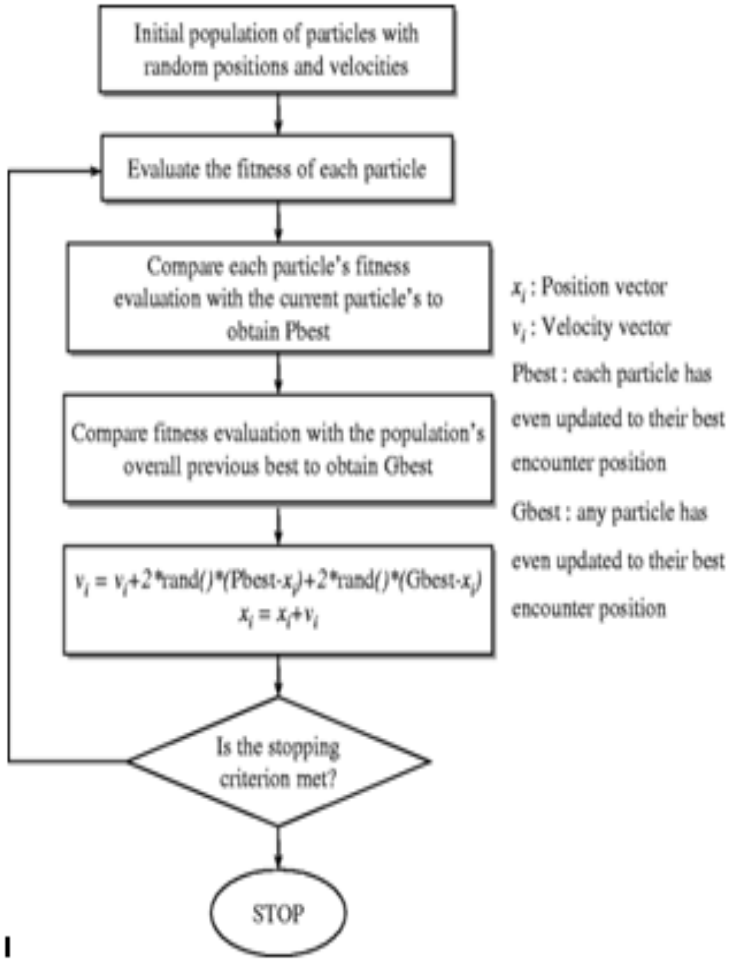

Fig 1 : PSO Algorithm Steps

IV. DATASET

German credit data set: We used the German Credit data set. This data set consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. This is a well-known real-life data set, containing both numerical and categorical attributes. The class attribute in the German Credit data set takes values representing good or bad classification of the bank account holders.

V. CONCLUSION

Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than 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. The purpose of this paper was 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.

REFERENCES

[1] S. Hajian, J. Domingo-Ferrer "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining",Knowledge and Data Engineering,vol 25 ,N0 .7, july 2013

[2] 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.

[3] 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.

[4] F.Kamiran & T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf.Belgium and The Netherlands, 2010.

[5] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[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

