Research Journal of Applied Sciences 10 (10): 550-554, 2015
ISSN: 1815-932X

Modern Boost Decision Tree Algorithm: A Novel and Effective Discrimination Prevention Technique Using Data Mining Technique

P. Baskaran and K. Arulanandam
Department of Computer Science, Manonmanium Sundaranar University, Tirunelveli, India
Department of Computer Applications, G.T.M Govt. Arts and Science College,
Gudiyatam, Taimlnadu, India

Abstract: The data mining is the vital point of data combination for business intelligence. Now a day, there has been emerging trends in database to discover useful patterns and/or correlations among attributes called data mining. Here in this research presents the data mining techniques like classification, clustering and associations analysis which is include algorithms of decision tree (like C4.5), rule set classifier, kNN and Naive Bayes, clustering algorithms (like k-Means and EM) machine learning (like SVM), association analysis (like Apriori). These algorithms are applied on a data warehouse for extracting useful information from a big database. All algorithms contain their description, impact and review of algorithms. Here, it is also show the comparison between the classifiers by accuracy which shows ruleset classifier have higher accuracy when implement in MATLAB Software using EXCEL datasheet. These algorithms useful to find out the discrimination of employee, employee performance of industries like banking, insurance, medical, information technology sector, etc. and also age and sex discrimination.

Key words: Discrimination, employee, online questionnaire, algorithms, data

INTRODUCTION

Discrimination involves the group's initial reaction that influencing the individual's actual behavior towards the group, restricting members of one group from privileges that are available to another group, leading to the rejection of the individual orentities based on logical decision making. Discrimination based on age, religion, gender, caste, disability, employment, language, color and nationality. There are several decision-making tasks which made them to discrimination, e.g., loan granting, education loan granting, education and health insurances. Given a set of information items on a customer, an automated system agrees whether the customer is to be recommended for a credit or a certain type of life insurance. Automating such decision making reduces the workload of the employee of banks and insurance companies, among other organizations. Age discrimination is discrimination that depends on views and values which used to justify discrimination and subordination based on someone's age (Arulanandam and Baskaran, 2014). Ageism and sexism defines that it directed towards old people or adolescents and children. Disability discrimination is the process of individuals which treats as the standard of usual living that ends in public and private places, education and social work that are built to endure best people, thereby rejecting those with various disadvantages. Denying someone job opportunities or disallowing one from applying for particular jobs is often considered as employment discrimination for such a rejection is not related to the requirements protected characteristics may include age, disability, ethnicity, weight, religion, gender, gender identity, height, nationality, gender orientation and skin color but skin color is not considered in India (Baskaran and Arulanandam, 2015). In the beginning, automating decisions may give a sense of fairness but the decision rule does not learn itself by personal preferences. The classification rules are actually learned by the system model based on historical data or questionnaire data. If the original data are in herently prejudiced against a particular community or groups, the learned model may also show the negative impacton it. For example, in a certain loan granting organization, foreign people are rejected for loan for the years. If this biased historical dataset in database is used as training data to learn classification rules for an automated loan granting system, educational loan granting, pension granting, the learned rules also suffered from biased behavir toward foreign people and also the system may consider that the foreign is a legitimate criterion for loan rejection.

Corresponding Author: P. Baskaran, Department of Computer Science, Manonmanium Sundaranar University, Tirunelveli, India

Literature review: The researchers collect the sampling for classification with no discrimination by preferential sampling classification with no discrimination by preferential sampling is an expected solution for the discrimination issue. It gives guaranteeing results with both steady and unstable classifiers. It decreases the security level by keeping up a high precision level. It gives comparable execution to "massaging" however without changing the dataset and dependably beats the "reweighing" plan.

In integrating induction and deduction for finding evidence of discrimination researchers by Pedreschi presented a reference model for the examination and expose of discrimination in socially-sensitive choices taken by DSS. The methodology comprises first of extracting frequent classification rules and after that of examining them on the premise of quantitative measures of discrimination and their measurable significance. The key legitimate ideas of protected-by-law groups, direct discrimination, indirect discrimination, honest to goodness occupational prerequisite, affirmative activities and partiality are formalized as explanations over the set of concentrated runs and perhaps, extra foundation information.

Ruggieri present the issue of finding discrimination through data mining in a dataset of recorded choice records, taken by people or via programmed frameworks for data model. They formalize the techniques of direct and indirect discrimination revelation by displaying protected by law groups and connections where discrimination happens in a classification based extraction. Essentially, classification rules extracted from the dataset permit for revealing connections of unlawful discrimination where the level of load over protected-by-law groups is formalized by an augmentation of the lift measure of a classification rules.

In classification without discrimination suggest the thought of discrimination is non-minor and poses ethical furthermore legitimate issues and also hindrances in common sense applications. CND furnishes us with a basic yet influential beginning stage for the arrangement of the discrimination issue. CND classifies the future information (both discriminatory and non-discriminatory and direct and indirect discrimination) with least discrimination and high exactness. It likewise addresses the issue of redlining.

Arulanandam and Baskaran (2014) introduce a group of pre-processing discrimination prevention methods and specify the different features of each approach and how these approaches deal with direct or indirect discrimination. A presentation of metrics used to evaluate the performance of those approaches is also given. Finally, we conclude our study by enumerating interesting future directions in the research body. In direct discrimination, the extract system can be in a straight line mined in search of discriminatory contexts. In indirect discrimination, the mining method needs some background knowledge as an additional contribution, e.g., online student data that combine with the extracted system might permit for presentation contexts of discriminatory decisions.

Baskaran and Arulanandam (2015) study focuses on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. It should also be noted that classifiers of a particular group also did not perform with similar accuracies. Overall, the results indicate that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study. So in this study recommend that users should try their data set on a set of classifiers and choose the best one. Here discussed few data mining algorithms which are used to perform data analysis tasks in different fields. The researchers proposed modern boost decision tree classifier algorithms has higher accuracy that other classifiers. This algorithms employed in fraud detection, intrusion detection, BPO Industry, Finance and Health for extraction of useful information (Baskaran and Arulanandam, 2015).

Hajian and Domingo-Ferrer discusses the discrimination prevention in data mining and propose new techniques applicable for direct or indirect discrimination prevention individually or both at the same time. The researchers discuss how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (non-discriminatory) classification rules. It is also propose new metrics to evaluate the utility of the proposed approaches and here compare these approaches. The experimental evaluations demonstrate that the proposed techniques are effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality (Hajian and Domingo-Ferrer, 2013; Hajian *et al*., 2011).

Calders and Verwer (2010) investigate how to modify the naive bayes classifier in order to perform classification that is restricted to be independent with respect to a given sensitive attribute. Such independency restrictions occur naturally when the decision process leading to the labels in the data-set was biased, e.g., due to gender or racial discrimination.

Hajian *et al.* (2011) discusses how to clean training datasets and outsourced datasets in such a way that legitimate classification rules can still be extracted but indirectly discriminating rules cannot.

MATERIALS AND METHODS

This is a descriptive research that studies various aspects of designation of male and female work performance, discrimination done in incentives granted and impact of level of discrimination on organizational productivity and employee performance. Data regarding employee performance was collected from well structured online questionnaire in which employees fill the questionnaire and filled questionnaire received by me. So, the sample size was 620 including of 380 males, 240 females and 0 transgender. The male and female statistics were then separately studied and compared with the date and performance statistics to draw various inferences and get various output.

Existing system: In existing system, the initial idea of using rule protection and rule generalization for direct discrimination prevention, here it is no experimental results. Here, it is introduced the use of rule protection in a different way for indirect discrimination prevention and it is some preliminary experimental results. In this research, it is 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. Here, it is specify the different features of each method. Since, methods in our earlier published and working papers could only deal with either direct or indirect discrimination, the methods described in this research are new.

Proposed system: In our proposed system easy and fast model is invented. The following is our proposed model (Fig. 1).

Methods: Formative research was conducted in early 2015 and consisted of interviews with policy makers, a focus group with health personnel and a review of national labor and gender policies to assess interest and the extent to which various forms of violence were recognized in India. This formative research informed the development of data collection instruments; the identification of avenues of data analysis and the generation of culturally appropriate descriptions of workplace violence and gender discrimination, age discrimination and color discrimination including associated behaviors.

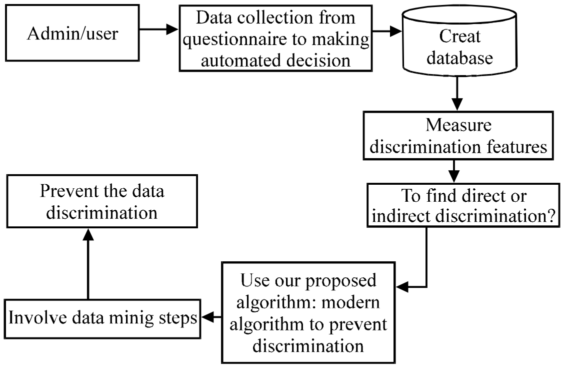
Admin/user
Data collection from questionnaire to making automated decision
Creat database
Measure discrimination features
To find direct or indirect discrimination?
Use our proposed algorithm: modern algorithm to prevent discrimination
Involve data minig steps
Prevent the data discrimination

Fig. 1: Proposed algorithm design

Data collection for the study took place in January 2015 and combined qualitative and quantitative approaches to determine the prevalence of workplace violence and its forms, victims and perpetrators; identify contributing factors to workplace violence including gender-related factors; describe victims' reactions and significances and describe any existing workplace violence policy and programs that could be strengthened or extended to address the issue. The study made use of six data collection tools: a health workers survey, facility manager and key informant interviews, patient focus groups and a facility risk assessment inventory (NB: this article draws only from a subset of IT sector, housewife and government servant in addition to information collected following dissemination of the study report). The forms of violence studied were verbal abuse, bullying, physical attack and sexual harassment. Exploration of the influence of gender on workplace violence focused on individual, organizational, facility-specific and societal factors contributing to workplace violence. The IT sector and government workers survey included open and closed-ended questions covering forms of gender discrimination not measured in previous studies of workplace violence (e.g., workers' self report on equal access to jobs, training and career advancement; equal treatment of men and women; pregnancy and family responsibility discrimination; the "glass ceiling" or vertical segregation; task segregation and perceptions of women and men at work).

After the study report results were disseminated, the researchers conducted new analyses of health worker survey data to better understand the perpetrator/victim dyad, documented the recommendations made by the Rwandan study stakeholder institutions and reviewed the content of three versions of the national code regulating labor in Rwanda to identify any policy impact the study may have had.

Sampling: The health workers survey was carried out in fifteen of Rwanda's 30 administrative districts which were selected at random. Within each district, three health facilities were then selected at random. The facility sample

Table 1: Instances of perpetration by type of workplace violence and sex of victim

	Verbal abuse experience by		Bullying experienced by		Sexual harassment		Physical attack experienced by		
Perpetrator identified as	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Total
Male	28	23	71	51	0.0	91	51	51	221
Female	59	54	17	35	86	0.0	25	49	216
Female and male	13	33	12	24	14	9	24	0.0	104
Total (100%)	76	164	65	126	23	38	18	31	541

included referral hospitals, district hospitals, health centers, clinics and public health units or health posts, each of which were managed either by the government (public) or by non-governmental organizations authorized by the government (accredited facilities) or in some cases, by the private sector. The government and private workers sample consisted of those who were in the randomly selected facilities on the day data collectors arrived at the targeted sites. Wherever possible, female and male workers were selected to reflect the proportion of men and women believed to be in the population of India (i.e., six female and four male, yielding a stratified sample). A total of 620 responses were surveyed; 240 were women and 380 men. Of the total number, 566 were from urban and 54 from rural sites (Fig. 2, 3 and Table 1).

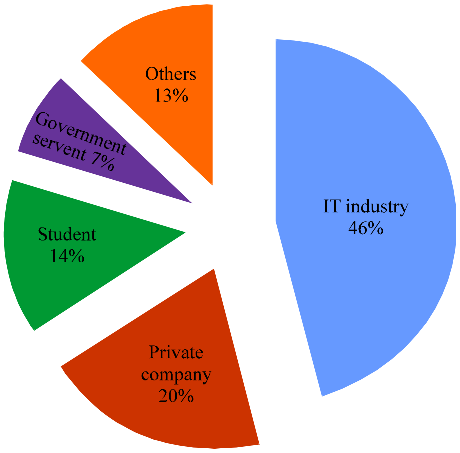

Fig. 2: Distribution of cadres in the sample

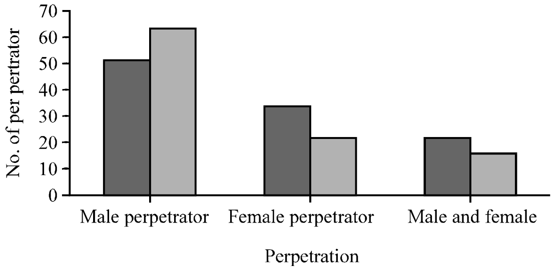

Fig. 3: Patterns of perpetration and victimization for bullying

Data analysis: Qualitative data were analyzed for content and trends. Survey, interview and facility data were collected by online questionnaire in a database using Excel.

RESULTS AND DISCUSSION

Our proposed algorithm runs efficiently on large databases and has the capability of handling thousands of input variables. It generates the generalization error as the effective method for estimating missing data and maintains accuracy when large proportion of the data are missing. Our proposed that has been generated can be saved in order to make comparative study about the features of the attributes. To measure the effectiveness of the approach experiments have been conducted.

Meanwhile, decision trees are constructed in a top-down recursive divide-and-conquer manner and the compatibility of decision trees degrades because the output is limited to one attribute. Trees created from the numeric datasets seems to be more complex and also when the database is large the complexity of the tree increases. In comparison with the 16 algorithms the time complexity of decision trees increases exponentially with the tree height. Hence, shallow trees tend to have large number of leaves and high error rates.

As the tree size increases, training error decreases. However as the tree size increases, testing error decreases at first since, we expect the test data to be similar to the training data but at a certain point, the training algorithm starts training to the noise in the data, becoming less accurate on the testing data. At this point, we are no longer fitting the data and instead fitting the noise in the data. This is called over fitting to the data in which the tree is fitted to spurious data. As the tree grows in size, it will fit the training data perfectly and not be of practical use for other data such as the testing set.

In Fig. 4, accuracy of our proposed algorithms is better than the other sixteen data mining techniques. The classification error also is lesser than other sixteen data mining technique. Compare with J48, decision tree data mining technique is lesser accuracy and lesser classification error. Here in this research decision tree technique has same value of accuracy and classification error. The Kstar and simple logistics classifier has same accuracy.

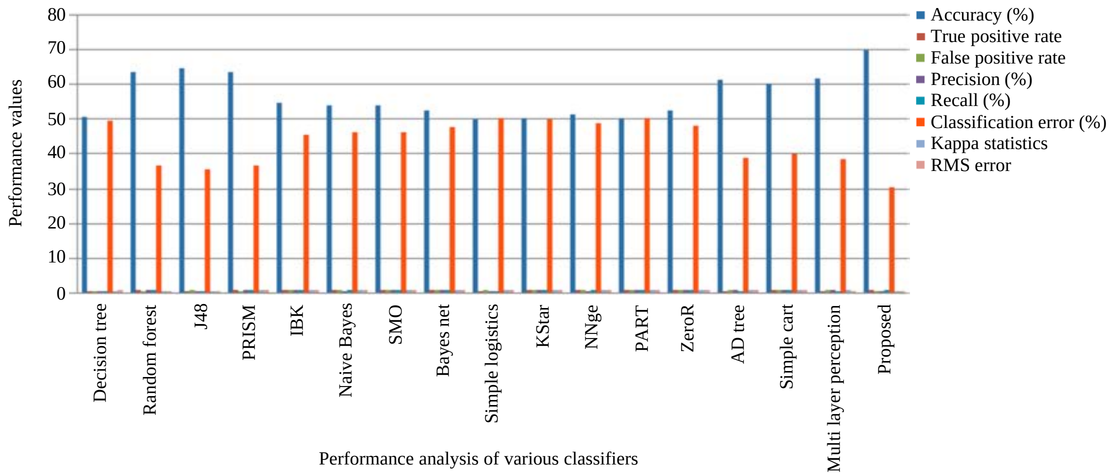

Fig. 4: Performance analysis of various classifiers

CONCLUSION

This study focuses on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. It should also be noted that classifiers of a particular group also did not perform with similar accuracies. Overall, the results indicate that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study. So, here it is recommend that users should try their data set on a set of classifiers and choose the best one, i.e., our proposed technique. Here in this research discussed few data mining algorithms which are used to perform data analysis tasks in different fields. Our proposed modern boost decision tree algorithms have higher accuracy that other classifiers. This algorithms employed in discrimination and prevention of IT sector, Government and Private employees.

REFERENCES

Arulanandam, K. and P. Baskaran, 2014. Discrimination invention and preclusion in online edification structure for information technology studies. Int. J. Hum. Comput. Interact., 6: 62-71.

Baskaran, P. and K. Arulanandam, 2015. Modern boost: An effective discrimination prevention using data mining technique. Aust. J. Basic Appl. Sci., 9: 297-305.

Calders, T. and S. Verwer, 2010. Three naive Bayes approaches for discrimination-free classification. Data Mining Knowl. Discovery, 21: 277-292.

Hajian, S. and J. Domingo-Ferrer, 2013. A methodology for direct and indirect discrimination prevention in data mining. Knowl. Data Eng., 25: 1445-1459.

Hajian, S., J. Domingo-Ferrer and A. Martinez-Balleste, 2011. Rule Protection for Indirect Discrimination Prevention in Data Mining. In: Modeling Decision for Artificial Intelligence. Torra, V., Y. Narakawa, J. Yin and J. Long (Eds.). Springer, Berlin Heidelberg, Germany, ISBN: 978-3-642-22588-8, pp: 211-222.