ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Modern Boost: An Effective Discrimination Prevention Using Data Mining Technique

[1]P.Baskaran and [2]Dr. K. Arulanandam

[1]*Research Scholar, Dept of Computerr Science, Manonmanium Sundaranar University, Tirunelveli*
[2]*Research Supervisor,Professor, Dept. of Computer Applications, G.T.M Govt. Arts and Science College, Gudiyatam, Taimlnadu, India*

ARTICLE INFO

Article history:
Received 16 April 2015
Accepted 12 June 2015
Available online 1 July 2015

Keywords:
Decision Tree, Rule set Classifier, kNN, Naïve Bayes, k-Means, EM, SVM, Apriori

ABSTRACT

The Data Mining is the essential 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. This paper presents the data mining techniques like Classification, Clustering and Associations Analysis which include algorithms of Decision Tree (like C4.5), Rule set Classifier,kNN and Naïve Bayes ,Clustering algorithms(like k-Means and EM)Machine Learning (Like SVM),Association Analysis(like Apriori). These algorithms are applied on data warehouse for extracting useful information. All algorithms contain their description, impact and review of algorithms. We also show the comparison between the classifiers by accuracy which shows ruleset classifier have higher accuracy when implement in MATLAB. These algorithms useful in increasing sales and performance of industries like banking, insurance, medical etc and also detect fraud and intrusion for assistance of society. Since the four metrics like 1).Direct discrimination prevention degree (DDPD) 2).Direct discrimination protection reservation(DDPP) 3).Indirect discrimination prevention degree (IDPD) and 4).Indirect discrimination protection preservation(IDPP) measures are used to evaluate the success of the proposed method in direct and indirect discrimination prevention, ideally their value should be 100 percent. To measure data quality, we use two metrics proposed in the literature as information loss measures in the context of rule hiding for privacy-preserving data mining (PPDM).

To Cite This Article: P. Baskaran and Dr. K. Arulanandam., Modern Boost: An Effective Discrimination Prevention Using Data Mining Technique. ***Aust. J. Basic & Appl. Sci.,*** *9(20): 297-305, 2015*

INTRODUCTION

As the data sizes accumulated from various fields are exponentially increasing, data mining techniques that extract information from large amount of data have become popular in commercial and scientific domains, including marketing, customer relationship management, quality management. We studied various articles regarding performance evaluation of Data Mining algorithms on various tools, some of them are described here, Abdullah compared various classifiers with different data mining tools & found WEKA as best tool, Mahendra Tiwari & Yashpal Singh evaluated performance of 4 clustering algorithms on different datasets in WEKA with 2 test modes. Some people worked on use of classification algorithms in WEKA for datasets from specific areas such as Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R compared different data mining classification techniques to predict length of stay for an inpatient in hospital.

Generally arff datasets have 2 types of attributes nominal & numeric. There is need to find suitable classifiers for datasets with different type of class (either nominal or numeric), so we focused on evaluating performance of different classifiers in WEKA on datasets with numeric & nominal class attribute. During the evaluation, the input datasets and the number of classifier used are varied to measure the performance of Data Mining algorithm. Datasets are varied with mainly type of class attribute either nominal or numeric. We present the results for performance of different classifiers based on characteristics such as accuracy, time taken to build model identify their characteristics in acclaimed Data Mining tool- WEKA.

Classification maps data into predefined classes often referred as supervised learning because classes are determined before examining data. A classification algorithm is to use a training data set to build a model such that the model can be used to assign unclassified records in to one of the defined classes. A test set is used to determine the accuracy of the model. Usually, the given dataset is divided in to training and test sets, with training set used to build the model and test set used to validate it. There

Corresponding Author: P. Baskaran, Research scholar, Dept of Computerr Science, Manonmanium Sundaranar University, Tirunelveli,
E-mail: bas_vlr@yahoo.co.in

are various classifiers are an efficient and scalable variation of Decision tree classification. The Decision tree model is built by recursively splitting the training dataset based on an optimal criterion until all records belonging to each of the partitions bear the same class label. Among many trees are particularly suited For data mining, since they are built relatively fast compared to other methods, obtaining similar or often better accuracy.

Bayesian classifiers are statistical based on Bayes' theorem, they predict the probability that a record belongs to a particular class. A simple Bayesian classifier, called Naïve Bayesian classifier is comparable in performance to decision tree and exhibits high accuracy and speed when applied to large databases. Rule-based classification algorithms generate if-then rules to perform classification. PART, OneR & ZeroR of Rule, IBK, and KStar of Lazy learners, SMO of Function are also used in evaluation process.

MATERIALS AND METHODS

We have used the popular, open-source data mining tool Weka (version 3.6.6) for this analysis. Three different data sets have been used and the performance of a comprehensive set of classification algorithms (classifiers) has been analyzed. The analysis has been performed on a Windows 7 Enterprise system with Intel Dual Core CPU, 3GHz Processor and 4.00 GB RAM. The data sets have been chosen such that they differ in size, mainly in terms of the number of attributes.

A. Data set:

The first data set is a BPO Employee data used in our earlier study. The data set contains 9 attributes apart from the class attribute with 500 instances.

B. Classifiers Used:

A total of 14 classification algorithms have been used in this comparative study. The classifiers in Weka have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers, etc. A good mix of algorithms have been chosen from these groups that include Bayes Net & Naive Bayes (from Bayes), Multilayer Perceptron, Simple Logistics & SMO (from functions), IBk & KStar (from Lazy), NNge, PART & ZeroR (from Rules) and ADTree, J48, Random Forest & Simple Cart (from Trees). The following sections explain a brief about each of these algorithms.

1. SMO:

Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all the missing the values and transforms nominal attributes into binary ones. A single hidden layer neural network uses exactly the same form of model as an SVM.

2. IBk

IBk is a k-nearest-neighbor classifier that uses the same distance metric. k-NN is a type of instance based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. In this algorithm an object is classified by a majority vote of its neighbors.

3. Bayes Net:

Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. A Bayesian network is an annotated Directed Acyclic Graph (DAG) that encodes a joint probability distribution.

4. Naive Bayesian:

Naive Bayesian classifier is developed on bayes conditional probability rule used for performing classification tasks, assuming attributes as statistically independent; the word Naive means attributes of the data set are considered as independent and strong of each other.

5. Simple Logistics:

It is a classifier used for building linear logistic regression models. LogitBoost with simple regression functions are base learners used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection.

6. KStar (K):*

Aha, Kibler & Albert describe three instance-based learners of increasing sophistication. IB1 is an implementation of a nearest neighbor algorithm with a specific distance function. IB3 is a further extension to improve tolerance to noisy data. Instances that have a sufficiently bad classification history are forgotten and only instances that have a good classification history are used for classification. Aha described IB4 and IB5, which handle irrelevant and novel attributes.

7. Multilayer Perceptron:

Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perception network with three layers.

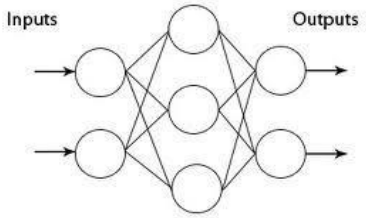

Fig. 1: Illustration of a perception network with three layers.

8. ***PART:***

PART uses the separate-and-conquer strategy, where it builds a rule in that manner and removes the instances it covers, and continues creating rules recursively for the remaining instances. Where C4.5 and RIPPER does global optimization to produce accurate rule sets, this added simplicity is the main advantage of PART.

9. ZeroR:

ZeroR is the simplest classification method which depends on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

10. NNge:

Instance-based learners are "lazy" in the sense that they perform little work when learning from the data set, but expend more effort classifying new examples. The simplest method, nearest neighbor, performs no work at all when learning. NNge does not attempt to out-perform all other machine learning classifiers. Rather, it examines generalized exemplars as a method of improving the classification performance of instance-based learners.

11. ADTree:

Alternating Decision Tree is one of the classification methods used in Machine learning which consists of decision nodes and prediction nodes. An instance is classified by an ADTree for which all decision nodes are true and summing any prediction nodes that are traversed. This makes it different from basic classification tree models that follow only one path through the tree.

12. J48:

The J48 algorithm is a WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning.

13. Simple Cart:

CART is a recursive and the gradual refinement algorithm of building a decision tree, to predict the classification situation of new samples of known input variable value. Breiman et. al., 1984 provided this algorithm and is based on Classification and Regression Trees (CART). In our study, we have applied all the above classifiers on the 3 different cancer data sets and the results have been analyzed.

14. Random Forest:

Random forest is an ensemble classifier which consists of many decision trees and gives the class as outputs i.e., the mode of the class's output by individual trees. Random Forests give many classification trees without pruning.

3. Modules Description:

1) Automated Data Collection:

Data mining is an increasingly important technology for extracting useful knowledge hidden in large collections of data. The problems outlined above can be eliminated when production data is collected automatically. When production data is collected automatically as it happens, you can be assured that it is timely, accurate, and unbiased. Until recently, automatically collecting production data was a costly and unreliable proposition.

There are, however, negative social perceptions about data mining, among which potential privacy invasion and potential discrimination. The latter consists of unfairly treating people on the basis of their belonging to a specific group. Automated data collection and data mining techniques such as classification rule mining have paved the way to making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training data sets are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc., discriminatory decisions may ensue. For this reason, antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining.

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.

4) Measure The Different Types Of Discrimination:

To construct the automated data collection database consist of discrimination rules. To measure the discrimination it has two types

1. Direct discrimination
2. Indirect discrimination

Negative social perceptions about data mining, among which potential privacy invasion and potential discrimination 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 which are strongly correlated with biased sensitive ones.

5) Direct Discrimination Measure:

Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Translated the qualitative statements in existing laws, regulations, and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures of the degree of discrimination of a PD rule. One of these measures is the extended lift (elift).

The purpose of direct discrimination discovery is to identify α discriminatory rules. In fact, α discriminatory rules indicate biased rules that are directly inferred from discriminatory items (e.g., Foreign worker). We call these rules direct α discriminatory rules. In addition to elift, two other measures slift and olift were proposed indirect discrimination measure Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or unintentionally could generate discriminatory decisions. Redlining by financial institutions (refusing to grant mortgages or insurances in urban areas they consider as deteriorating) is an archetypal example of indirect discrimination, although certainly not the only one. With a slight abuse of language for the sake of compactness, in this paper indirect discrimination will also be referred to as redlining and rules causing indirect discrimination will be called redlining rules.

Indirect discrimination could happen because of the availability of some background knowledge (rules), for example, that a certain zip code corresponds to a deteriorating area or an area with mostly black population. The background knowledge might be accessible from publicly available data (e.g., census data) or might be obtained from the original data set itself because of the existence of nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set.

The purpose of indirect discrimination discovery is to identify redlining rules. In fact, redlining rules indicate biased rules that are indirectly inferred from nondiscriminatory items.

6) Discrimination Prevention Based On The Measurement:

To measure the discrimination prevention it has two types

1. Direct Discrimination Prevention
2. Indirect Discrimination Prevention

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

1. Discrimination measurement.
2. Data transformation

7) Preventing Direct And In Direct Discrimination:

Discriminatory item sets (i.e., A) 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 fact, in this case, only PND rules are extracted from DB so only indirect discrimination could happen.

At least one discriminatory item set (i.e., A) is not removed from the original database (DB). So it is clear that PD rules could be extracted from DB and direct discrimination could happen. However, in addition to direct discrimination, indirect discrimination might occur because of background knowledge obtained from DB itself due to the existence of nondiscriminatory 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 α 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 α discriminatory rules or prevent the existing ones from being removed.

8) Direct And Indirect Prevention Algorithm:

Construct the above data transformation method and to implementing the prevention algorithm. This algorithm used to prevent simultaneously direct and indirect discrimination at the same time. The algorithm starts with redlining rules and discriminatory rules. Algorithm based on the rule protection and rule generalization methods. If some rules can be extracted from DB as both direct and indirect α discriminatory rules, it means that there is overlap between MR and RR, in such case, data transformation is performed until both the direct and the indirect rule protection requirements are satisfied. This is possible because, the same data transformation method (Method 2 consisting of changing the class item) can provide both DRP and IRP. However, if there is no overlap between MR and RR, the data transformation is performed according to Method 2 for IRP, until the indirect discrimination prevention requirement is satisfied for each indirect α discriminatory rule ensuing from each redlining rule in RR.

9) ***Measure The Computational Cost And Prevention Degree:***

The final stage is computed the computational cost and utility measurement. To measure discrimination removal, four metrics were used:

1. Direct discrimination prevention degree (DDPD). This measure quantifies the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed data set. DDPD can be defined as DDPD= □□ −|□□′| / |□□| where MR is the database of α-discriminatory rules from DB and MR′ is the database of α- discriminatory rules extracted from the transformed data set DB′.

2. Direct discrimination protection preservation (DDPP):

This measure quantifies the percentage of the α-protective rules in the original data set that remain α-protective in the transformed data set. It is defined as □□□□= |□□|∩|□□′| / |□□| where □□ is the database of α-protective rules extracted from the original data set □□ and □□′ is the database of α-protective rules extracted from the transformed data set □□′

3. Indirect discrimination prevention degree (IDPD):

This measure quantifies the percentage of redlining rules that are no longer redlining in the transformed data set.

4. Indirect discrimination protection preservation (IDPP):

This measure quantifies the percentage of nonredlining rules in the original data set that remain nonredlining in the transformed data set.

Since the above measures are used to evaluate the success of the proposed method in direct and indirect discrimination prevention, ideally their value should be 100 percent. To measure data quality, we use two metrics proposed in the literature as information loss measures in the context of rule hiding for privacy-preserving data mining (PPDM).

10. Proposed system:

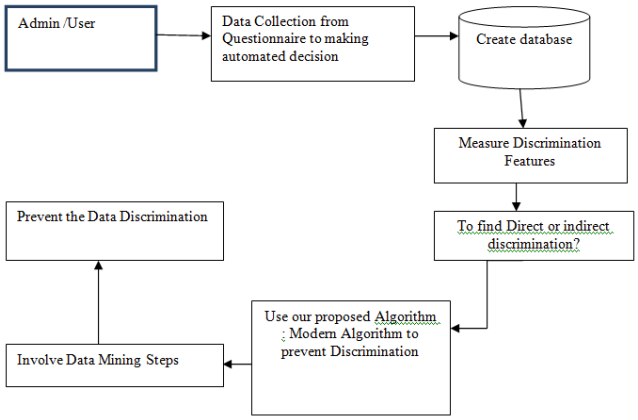

Fig. Proposed Algorithm Design.

11. Performance measures:

In this approach, the classification accuracy rates for the datasets were measured. For example, in the classification problem with two-classes, positive and negative, a single prediction has four possibility. The True Positive rate (TP) and True Negative rate (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative. A False Negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

Table 1: Confusion Table.

Prediction		Disease	
		+	-
Test	+	True Positive (TP)	False Negative (FP)
	-	False Negative (FN)	True Negative (TN)

1. Accuracy - It refers to the total number of records that are correctly classified by the classifier.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \text{ -------} \quad (1)$$

2. Classification error - This refers to the misclassified datasets from the correctly classified records.

3. True Positive Rate (TP): It corresponds to the number of positive examples that have been correctly predicted by the classification model.

4. False Positive Rate (FP): It corresponds to the number of negative examples that have been wrongly predicted by the classification model.
5. Kappa Statistics - A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
6. Precision - is the fraction of retrieved instances that are relevant.

$$Precision = \frac{TP}{TP+FP} \text{----------} \quad \text{--(2)}$$

7. Recall - is the fraction of relevant instances that are retrieved.

$$Recall = \frac{TP}{TP+FN} \text{----------} \quad (3)$$

8. Root-Mean-Squared-error - It is a statistical measure of the magnitude of a varying quantity. It can be calculated for a series of discrete values, or for a continuously varying function. Since the class label prediction is of multi-class, the result on the test set will be displayed as a two-dimensional confusion matrix with a row and a column for each class. Each matrix element shows the number of test cases for which the actual class is the row and the predicted class is the column. Finally, the error rate is one minus this. ROC curves depict the performance of a classifier without regard to class distribution or error costs. They plot the number of positives included in the sample on the vertical axis, expressed as a percentage of the total number of positives, against the number of negatives included in the sample, expressed as a percentage of the total number of negatives, on the horizontal axis. Information retrieval researches define parameters called recall and precision.

$$\text{recall} = \frac{\text{number of documents retrived that are relevant}}{\text{Total number of documents that are relevant}}$$

$$\text{Precision} = \frac{\text{number of documents retrived that are relevant}}{\text{number of documents retrived that are retrived}}$$

F-measure is another information retrieval measure that is calculated from TP, FP, FN or recall or precision values

$$\text{f} - \text{measure} = \frac{2 * \text{recall} * \text{Precision}}{\text{recall} + \text{Precision}}$$

$$\text{f} - \text{measure} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}$$

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.

Table 2: Performance Analysis of Various Classifiers.

Classifier	Accuracy(%)	True Positive Rate	False Positive Rate	Precision(%)	Recall(%)	Classification Error (%)	Kappa Statistics	RMS Error
Decision Tree	50.68	0.507	0.230	0.478	0.507	49.32	0.211	0.404
Random Forest	63.34	0.633	0.254	0.570	0.633	36.66	0.354	0.313
J48	64.45	0.500	0.544	0.521	0.500	35.55	0.344	0.310
PRISM	63.45	0.750	0.350	0.825	0.750	36.55	0.635	0.718
IBK	54.50	0.871	0.594	0.571	0.871	45.50	0.484	0.539
Naïve Bayes	53.75	0.571	0.594	0.528	0.571	46.25	0.484	0.539
SMO	54.00	0.643	0.465	0.629	0.643	46.00	0.589	0.632
Bayes Net	52.50	0.681	0.502	0.625	0.681	47.50	0.585	0.536
Simple Logisitics	49.80	0.547	0.450	0.520	0.547	50.20	0.580	0.598
KStar	50.25	0.564	0.459	0.561	0.564	49.75	0.480	0.654
NNge	51.20	0.655	0.500	0.540	0.655	48.80	0.490	0.655
PART	49.99	0.652	0.550	0.650	0.652	50.01	0.500	0.654
ZeroR	52.25	0.584	0.546	0.643	0.584	47.75	0.480	0.680
AD Tree	61.18	0.500	0.640	0.684	0.500	38.82	0.465	0.500
Simple Cart	60.16	0.600	0.490	0.682	0.600	39.84	0.470	0.654
Multi Layer Perception	61.58	0.546	0.500	0.855	0.546	38.42	0.495	0.356
Proposed(Modern Boost)	68.80	0.650	0.545	0.420	0.650	31.20	0.301	0.212

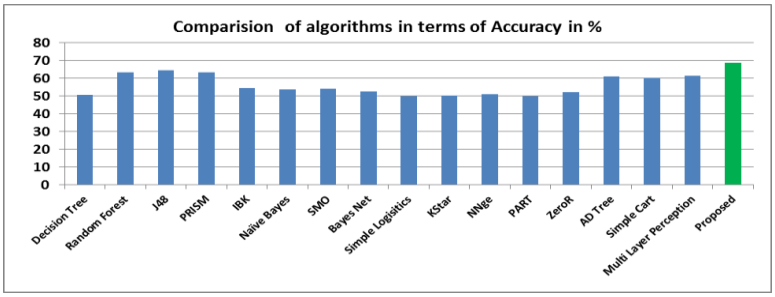

Fig. 2: Comparison of accuracy in between the seventeen algorithms.

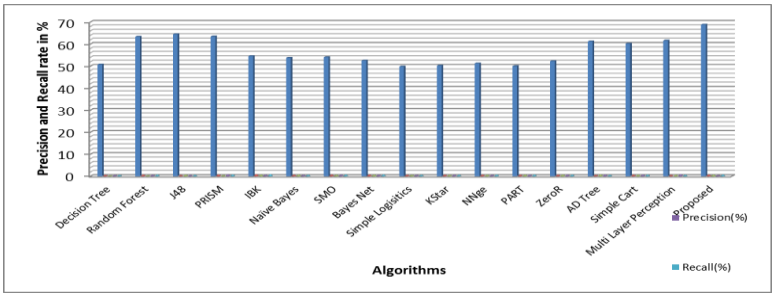

Fig. 3: Comparison of Precision and Recall values in between the seventeen algorithms.

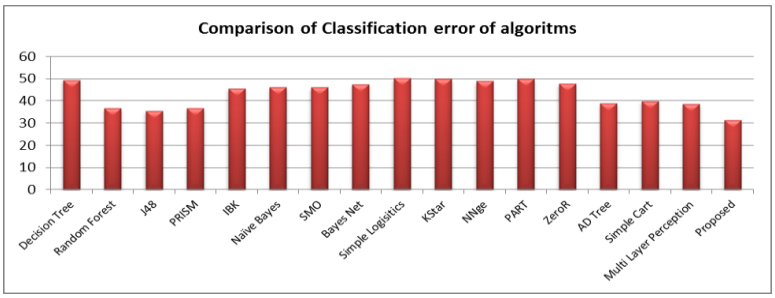

Fig. 4: Comparison of Classification error in between the seventeen algorithms.

The performance obtained using Proposed Modernt Boost Decision Tree classifier was found to be higher than the results obtained by our et.al as described in the Table 2 which depicts that Proposed Modern Boost Decision Tree algorithm performs better than that Decision tree.

In weka Proposed Modern Boost Decision Tree classifier algorithm shows only the number of instances is correctly clustered and incorrectly clustered. From this we are able to know that which algorithm is best. Table show the results that the Proposed Modern Boost Decision Tree classifier correctly clustered more instances than other algorithm.

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, we recommend that users should try their data set on a set of classifiers and choose the best one. Here we discussed few data mining algorithms which are used to perform data analysis tasks in different fields. Our 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.

Future Work:

We would like to develop web based software for performance evaluation of various classifiers (including our proposed) where the users can just submit their data set and evaluate the results on the fly.

VI. *Future Work:*

We would like to develop web based software for performance evaluation of various classifiers (including our proposed)where the users can just

submit their data set and evaluate the results on the fly.

REFERENCE

Ponniah, P., 2007. ―Data Warehousing Fundamentals- ―A comprehensive guide for IT professionals‖, Ist ed.,second reprint , ISBN-81-265-0919-8, Glorious Printers: New Delhi India.

An Introduction to Data Mining, Review: http://www.thearling.com/text/dmwhite/dmwhite.htm

A Tutorial on Clustering Algorithms, Review http://home.dei.polimi.it/matteucc/Clustering/tutorial_html

Naive Bayes Classifier Review: http://www.statsoft.com/textbook/naive-bayes-classifier/

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2005. ―An Introduction to Data Mining‖, ISBN : 0321321367. Addison-Wesley.

XindongWu, Vipin Kumar, 2008. Top 10 algorithms in data mining‖ Knowl Inf Syst., 14: 1–37. DOI 10.1007/s10115-007-0114-2

Murthy, S., S. Salzberg, 1995. Lookahead and pathology in decision tree induction,in C. S. Mellish, ed., `Proceedings of the 14th International Joint Conference on Articial Intelligence', Morgan Kaufmann, pp: 1025-1031.

Jiawei Han, Micheline Kamber, 2006. Data Mining:Concepts and Techniques, Second Edition, ISBN 13: 978-1-55860-901-3, Elsevier.

Utgo, P.E., 1997. `Decision tree induction based on efficient tree restructuring', Machine Learning, 29-5.

Esposito, F., D. Malerba, G. Semeraro, 1995. Simplifying decision trees by prun-ing and grafting: New results, in N. Lavrac & S. Wrobel, eds, `Machine Learning:ECML-95 (Proc. European Conf. on Machine Learning, Lecture Notes in Articial Intelligence 914, Springer Verlag, Berlin, Heidelberg, New York, pp: 287-290.

Shafer, J., R. Agrawal, M. Mehta, 1996. Sprint: a scalable prallel classier for data mining, in `Proceedings of the 22nd International Conference on Very Large Databases (VLDB)'.

Freitas, A.A., S.H. Lavington, 1998. Mining Very Large Databases with Parallel Processing, Kluwer Academic Publishers.

Kearns, M. Y. Mansour, 1998. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization, in J. Shavlik, ed., `Machine Learning: Proceedings of the Fifteenth International Conference', Morgan Kaufmann Publishers,Inc., pp: 269-277.

Friedman, J., R. Kohavi, Y. Yun, 1996. Lazy decision trees, in `Proceedings of the Thirteenth National Conference on Articial Intelligence', AAAI Press and the MIT Press, pp. 717-724.

Quinlan, J.R., R.L. Rivest, 1989. `Inferring decision trees using the minimum description length principle', Information and Computation, 80: 227-248.

Mehta, M., J. Rissanen, R. Agrawal, 1995. MDL-based decision tree pruning,in U. M. Fayyad & R. Uthurusamy, eds, `Proceedings of the first international conference on knowledge discovery and data mining', AAAI Press, pp: 216-221.

Wallace, C., J. Patrick, 1993. `Coding decision trees', Machine Learning, 11: 7-22.

Biao Qin, Yuni Xia, 2009. ―A Rule-Based Classification Algorithm for Uncertain Data‖ IEEE International Conference on Data Engineering, pp: 1633-1640.

Jiuyong Li., Construct robust rule sets for classification, SIGKDD ’02 Edmonton, Alberta, Canada.

Yang, Y., 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In SIGIR-94.

Fix, E. and J.L. Hodges, Jr., 1951. "Discriminatory analysis, nonparametric discrimination: consistency properties," U.S. Air Force Sch. Aviation Medicine, Randolf Field, Tex., Project 21-49-004, Contract AF 41(128)-31, Rep. 4.

Cost, S. and S. Salzberg, 1993. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning,10(1): 57–78.

Eui-Hong (Sam) Han, Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification.

Patrick, E.A. and F.P. Fischer, III, 1970. "A generalized k-nearest neighbor rule," Inform. Contr., 16: 128-152.

Cohen, W.W. and H. Hirsh, 1998. Joins that generalize: Text classification using WHIRL. In Proc. of the Fourth Int’l Conference on Knowledge Discovery and Data Mining.

Dennis, l., 1972. Wilson, Asymptotic properties of nearest neighbor rules using edited data‖ IEEE transactions on systems, man, and cybernetics, Smc-2-3.

Langley, P., W. Iba, K. Thompson, 1992. An analysis of Bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence (pp: 223–228). San Jose, CA: AAAI Press.

Langley, P., S. Sage, 1994. Induction of selective Bayesian classifiers. In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (pp: 399–406). Seattle, WA: Morgan Kaufmann.

Clark, P., T. Niblett, 1989. The CN2 induction algorithm. Machine Learning, 3: 261–283.

Cestnik, B., 1990. Estimating probabilities: A crucial task in machine learning. Proceedings of the Ninth European Conference on Artificial Intelligence. Stockholm, Sweden: Pitman.

Pazzani, M., J. Muramatsu, D. Billsus, 1996. Syskill&Webert: Identifying interesting web sites. Proceedings of the Thirteenth National Conference

on Artificial Intelligence (pp: 54–61). Portland, OR: AAAI Press.

John, G., P. Langley, 1995. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (pp: 338–345) . Montr´eal, Canada: Morgan Kaufmann.

Kubat, M., D. Flotzinger, G. Pfurtscheller, 1993. Discovering patterns in EEG-Signals: Comparative study of a few methods. Proceedings of the Eighth European Conference on Machine Learning (pp: 366–371). Vienna, Austria: Springer-Verlag.

Langley, P., 1993. Induction of recursive Bayesian classifiers. Proceedings of the Eighth European Conference on Machine Learning (pp: 153–164). Vienna, Austria: Springer-Verlag.

San, O.M., 2004. An alternative extension of the k-means algorithm for clustering categorical data‖ Int. J. Appl. Math. Comput. Sci., 14(2): 241–247.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations.—Proc. 5-th Symp.Mathematical Statistics and Probability, Berkelely, CA, 1: 281–297.

Ralambondrainy H., 1995. A conceptual version of the kmeans algorithm. — Pattern Recogn. Lett., 15(11): 1147–1157.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. — Data Mining Knowl. Discov., 2(2): 283–304.

Ngai, W.K., B. Kao, C.K. Chui, R. Cheng, M. Chau and K.Y. Yip, 2006. ―Efficient clustering of uncertain data,‖ in IEEE International Conference on Data Mining (ICDM), 436–445.

Chau, M., R. Cheng, B. Kao and J. Ng, 2006. ―Data with uncertainty mining: An example in clustering location data,‖ in Proc. of the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD).

A.C., 2007. On density based transforms for uncertain data mining,‖ in Proceedings of IEEE 23rd International Conference on Data Engineering, 866–875.

A.C. and Y.P.S., 2008.A framework for clustering uncertain data streams,‖ in Proceedings of IEEE 24rd International Conference on Data Engineering, 150–159.

Y. Xia and B. Xi, ―Conceptual clustering categorical data with uncertainty,‖ in IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007, pp. 329–336.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39:1–38.

McLachlan, G.J. and T. Krishnan, 1997. The EM Algorithm and Extensions. Wiley, New York.

Neal, R.M. and G.E. Hinton, 1998. A view of the EM algorithm that justifies incremental,sparse, and other variants. In Jordan, M.I., editor, Learning in Graphical Models, 355–368. Kluwer, Dordrecht.

Bradley, P.S., U.M. Fayyad and C.A. Reina, 1998. Scaling EM (expectation maximization) clustering to large databases. Technical Report No. MSR-TR-98- 35 (revised February, 1999), Microsoft Research, Seattle.

Moore, A.W., 1999. Very fast EM-based mixture model clustering using multiresolution kd-trees. In Kearns, M.S., Solla, S.A., and Cohn, D.A., editors, Advances in Neural Information Processing Systems, 11: 543–549. MIT Press, MA.

Boser, B.E., I.M. Guyon and V.N. Vapnik, 1992. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144{152, Pittsburgh, PA. ACM Press.

Chang,. C.C. and C.J. Lin, 2001. LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Vapnik, V., S. Golowich and A. Smola, 1997. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems, 9: 281–287, Cambridge, MA, MIT Press.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C.,

Rakesh Agrawal Ramakrishnan Srikan, 1994. Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference Santiago, Chile.

Yu, Z. and H. Wong, 2006. Mining uncertain data in low-dimensional subspace,‖ in International Conference on Pattern Recognition (ICPR), 748–751.

Chui, C., B. Kao and E. Hung, 2007. Mining frequent itemsets from uncertain data,‖ in Proc. of the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD), 47–58.

Houtsma, M. and A. Swami, 1993. Set-oriented mining of association rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

Othman, B., Md. Fauzi and T.M.S. Yau, 2007. "Comparison of different classification techniques using WEKA for breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. Springer Berlin Heidelberg.

Rohit, A., 2012. "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA." Int. Journal of Computer Applications, 54.13.

Delen, D., G. Walker and A. Kadam, 2005. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine, 34.2: 113-128.

Mark, H., 2009. "The WEKA data mining software: an update." ACM SIGKDD Explorations Newsletter, 11.1: 10-18.