



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Classification of data using New Enhanced Decision Tree Algorithm (NEDTA)

Hardeep Kaur¹ Harpreet Kaur²
Department of Computer Science and Engineering
SBBSIET, Jalandhar, Punjab, India

Abstract: Data mining is method of maintaining a large amount of data stored in the database. Decision tree is a technique of data mining which classify the data and produces valuable results. These results are used in analysis and future prediction. The prime objective of this research work is to present an enhanced decision tree algorithm that classifies the data more efficiently and effectively than existing decision tree classifiers. We apply existing decision tree classifiers ID3, J48, NBTree on a large amount of data. Then the efficiency and performance of existing algorithms is examined and compared with new enhanced decision tree algorithm (NEDTA). Our enhanced decision tree algorithm produces better results as compared to other decision tree algorithms.

Keywords: Data Mining; Decision Tree; ID3; J48; NBTree

I. Introduction

Data mining is a knowledge discovery process in which analysis of the data is done. The analysis is based on historical activities stored in very large repositories and results are used to obtain useful information. Data mining is method to find the hidden patterns in a large amount of data. There are various applications of data mining such as banking, insurance, medicine, real estate etc. Data mining concept is applied in insurance and banking field [1] for fraud detection, identification of loyal customers, sales promotion and enhanced research. The process of data mining is iterative and also known as Knowledge Discovery process. It consists of following phases:

1. Problem Definition: This phase consists of data mining experts, business experts and domain experts, who understand the problem, define objectives.
2. Data Exploration: In this phase data is explored and metadata is defined by domain experts.
3. Data Preparation: A data model is formed in this phase from collected data.
4. Modeling: Data mining functions from data model are selected and applied on data.
5. Evaluation: The results obtained by modeling are evaluated. If the results are not according to expectations then model is rebuild until required results are obtained.
6. Deployment: In deployment process, the mining results are deployed into applications

There are various data mining techniques are available such as Clustering, Classification, Association. Clustering is the process of grouping the similar objects into one class. Therefore multiple classes are formed which comprises similar objects. Association is the process in which association rules are created. These association rules analyze unrelated data and produces association between them.

In this paper, Section I describes introduction of data mining and the process of knowledge discovery. Section II describes about classification and their techniques. Section III gives information about decision tree classification technique. In this section some decision tree algorithms are explained. In Section IV the objectives of research work are discussed. In Section V Weka data mining tool is discussed briefly. Weka tool is used in research work. Section VI describes about data used in research work. The proposed work is implemented in Section VII. The results are evaluated and compared in section VIII. The comparison of algorithms is based on execution time, accuracy and error rate (Mean absolute error (MSE), Root mean squared error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE)). In section IX a new enhanced decision tree algorithm (NEDTA) is proposed. In section X the results of NEDTA are evaluated and compared with ID3, J48 and NBTree. Section XI contains conclusion. Section XII contains references that are used in this research work.

II. Classification

Classification is data mining technique which classifies data with the help of certain classification rules and valuable results are formed. There are two areas of classification: Decision Tree Induction and Neural Induction.

Decision tree induction is a classification method in which a set of rules are applied recursively on a dataset and a tree is generated.

III. Decision Tree

A decision tree consists of nodes [2]. Each node represents some information. Decision tree learning is started from root node and discrete values are produced at each node by testing the values of attribute. These discrete values acts as target function. Then by using target function, value of attribute for next node is evaluated. This process is repeated for each new node. The learned tree is represented by if-then rules.

Decision tree algorithms [3] such as ID3, C4.5, J48 [4] NBTree can be applied on large amount of data and valuable predictions can be produced. These predictions evaluate future behavior of problem. Decision tree are preferred because they can evaluate information more accurately than other methods.

In this research work following decision tree algorithms are used:

- (1) **ID3:** ID3 means Iterative Dichotomiser 3. It is a decision tree algorithm which is developed by Ross Quinlan. The steps of ID3 algorithm are as following:
 - (a) ID3 is a greedy algorithm in which the tree created from top to bottom.
 - (b) At each node, the appropriate attribute is selected which best classifies the data. Data is in the form of training examples.
 - (c) The above process is repeated until the complete tree is generated or until all the attributes used.
- (2) **J48:** J48 is the open source Java implementation [5] of C4.5 decision tree algorithm in Weka data mining tool. Following are the steps of J48 algorithm:
 - (a) This algorithm uses basic algorithm which create trees by using recursive top down divide and conquer approach.
 - (b) First of all, the training examples are at the root node.
 - (c) Test attribute is selected based on some measures such as information gain, entropy etc.
 - (d) Examples are divided repeatedly by using test attribute.
 - (e) The process continued until no sample leaf is leaf.
- (3) **NBTree:** NBTree(Naive Bayesian tree) consists of [6] naïve Bayesian classification and decision tree learning. An NBTree classification sorts the example to a leaf and then assigns a class label by applying a naïve bayes on that leaf. The steps of NBTree algorithm are:
 - (a) At each leaf node of a tree, a naïve bayes is applied.
 - (b) By using naïve bayes for each leaf node, the instances are classified.
 - (c) As the tree grows, for each leaf a naïve bayes is constructed.
 - (d) This process repeated until no example is left.

IV. Objectives of Research Work

The objectives of this research work are as following:

- (1) To apply Decision tree algorithms ID3, J48 and NBTree on banking dataset.
- (2) Evaluation of results produced.
- (3) Comparative analysis of results using parameters accuracy, execution time and error rate for ID3, J48 and NBTree.
- (4) To build a new enhanced method for classification of data.

V. Tool Used

In this research work, an open source tool named Weka is used. Weka is free open source data mining software which is based on a Java data mining library. Weka consists of various machine learning algorithms for different data mining applications. The algorithms are directly applied to dataset and results are generated in the form of tree. Weka contains various classifiers for classification [7], clustering, association, regression, pre-processing and visualization. Weka is also used for development of new machine learning schemes.

VI. Data Set Used

In our research work, we have used banking dataset [8]. The main focus of this research is performance and evaluation of decision tree algorithms. There are many decision tree algorithms in data mining but we focus mainly on ID3, J48 and NBTree. The data set contains 5264 rows and 13 columns.

VII. Implementation of Work

To implement the objectives of research work, Firstly, we have applied the ID3, J48 and NBtree algorithms on banking [9] dataset using data mining tool Weka 3.4 Figure 1 shows implementation of ID3 decision tree

algorithm on Weka data mining tool. Figure 2 shows implementation of J48 on Weka data mining tool. Figure 4 shows implementation of NBTree decision tree algorithm on Weka data mining tool

Figure 1: ID3 Algorithm

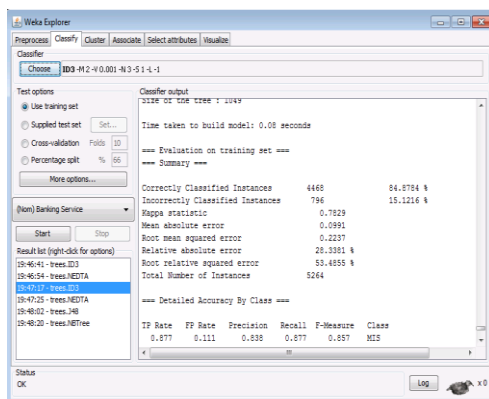


Figure 2: J48 Algorithm

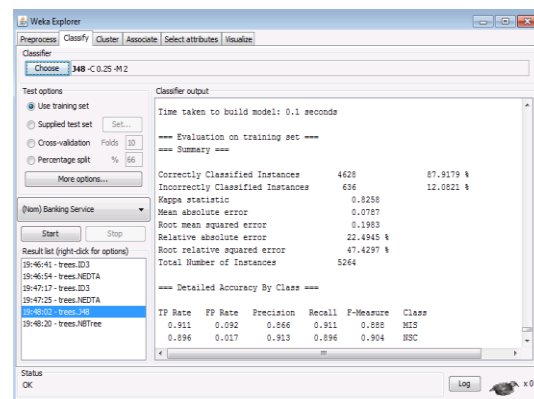


Figure 3: J48 Algorithm Visualization Tree in Weka Explorer

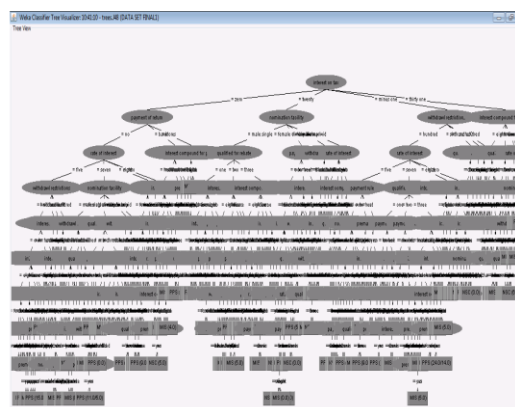
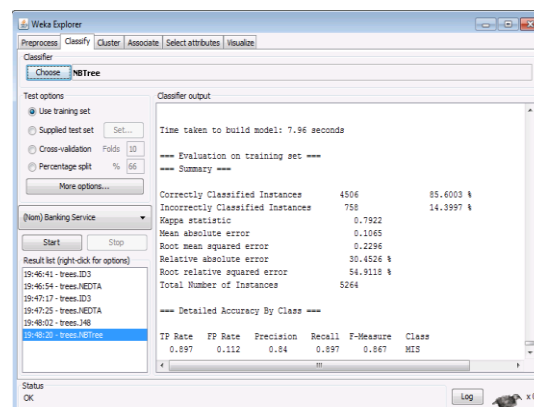


Figure 4: NBTree Algorithm



VIII. Comparative Analysis of ID3, J48 and NBTree

The experiments have been conducted and different decision tree algorithms are applied on banking dataset in Weka Framework. The results of Decision tree classifiers ID3, J48 and NBTree are compared. In our experiment, parameters such as error rate, execution time and accuracy are evaluated and compared. Table I shows the accuracy of decision tree classifiers ID3, J48 and NBTree. Table II shows the performance of different decision tree classifiers. The table shows execution time of various classifiers.

Table I: Classifier Accuracy

Algorithm	Correctly Classified instances	Incorrectly Classified Instances
ID3	84.8784 %	15.1216 %
J48	87.9179 %	12.0821 %
NBTree	85.6003 %	14.3997 %

Table II: Performance of Classifiers

Algorithm	Time Taken to build model (in seconds)
ID3	0.08
J48	0.1
NBTree	7.96

Table III shows the error rate of decision tree classifiers. Error rate is shown as Mean absolute error (MSE), Root mean squared error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE)

Table III: Error rate of Classifiers

Algorithm	MSE	RMSE	RAE	RRSE
ID3	0.0991	0.2237	28.3881 %	53.4855 %
J48	0.0787	0.1983	22.4945 %	47.4297 %
NBTree	0.1065	0.2296	30.4526 %	54.9118 %

Following graphs shows accuracy, error rate and execution comparison of ID3, J48 and NBTree algorithms.

Figure 5: This graph shows accuracy comparison of ID3, J48 and NBTree algorithms

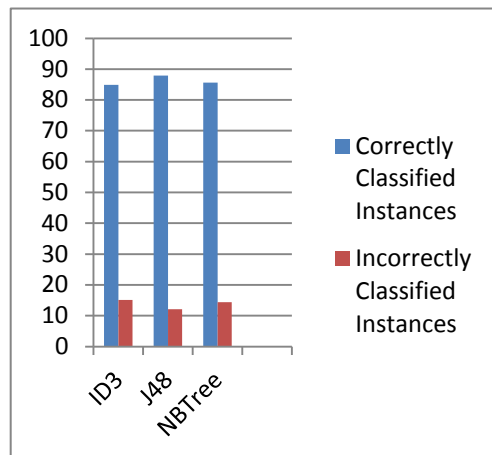


Figure 6: This graph shows Execution Time of ID3, J48 and NBTree algorithms

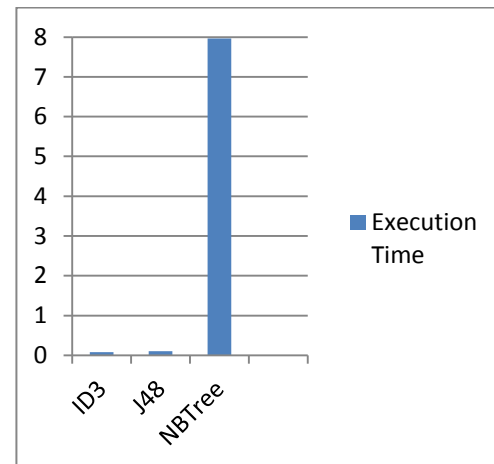
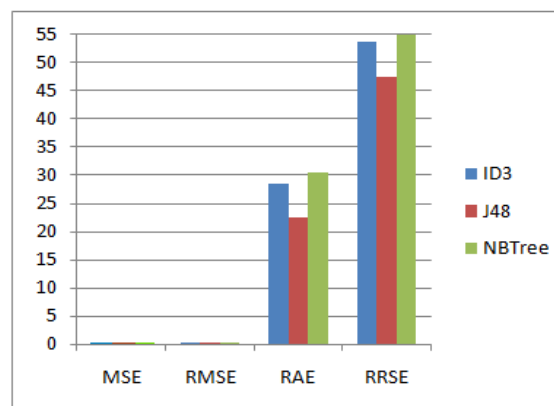


Figure 7: This graph shows Error Rate (MSE, RMSE, RAE and RRSE) comparison of ID3, J48 and NBTree algorithms



XI. New Enhanced Decision Tree Algorithm

We have proposed a new decision tree algorithm which classifies a large amount of data. Existing decision tree algorithms have some drawbacks. But our enhanced algorithm produces better results as compared to ID3, J48 and NBTree [10]. Following are the steps of proposed algorithm:

1. A decision tree DT built from the training examples, with a collection S of m source leaf nodes and a collection D of n destination leaf nodes.
2. A pre specified constant k ($k \leq m$), where m is the total number of source leaf nodes,
3. Construct the branches according to different values of attribute P_i so that the samples are partitioned accordingly.
4. If samples in a certain value are all of the same class, then generate a leaf node and is labeled with that class.
5. Otherwise use the same process repeated recursively to form a decision tree for the samples at each partition.

X. Results

NEDTA is applied on banking dataset and results are compared with ID3, J48 and NBTree algorithms. Figure 9 shows the evaluation of NEDTA on Weka data mining tool. NEDTA produces better results as compared to ID3, J48 and NBTree in terms of execution time, accuracy and error rate.

Figure 8: Implementation of New Enhanced Decision Tree Algorithm (NEDTA) on Weka Explorer

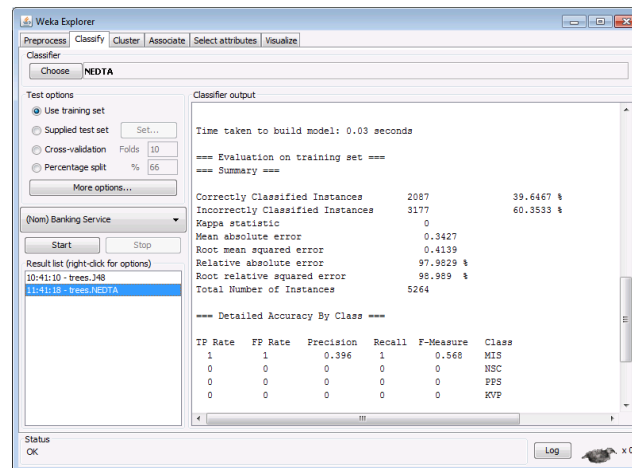


Table IV shows accuracy comparison of ID3, J48 and NBTree with NEDTA. The percentage of correctly classified instances of our algorithm is better than other algorithms. Table V shows the performance comparison of ID3, J48 and NBTree with NEDTA. The execution time of our algorithm is better than other algorithms. Table VI shows error rate comparison of ID3, J48 and NBTree with NEDTA.

Table IV: Comparison of accuracy of NEDTA with ID3, J48 and NBTree

Algorithm	Correctly Classified instances	Incorrectly Classified Instances
ID3	84.8784 %	15.1216 %
J48	87.9179 %	12.0821 %
NBTree	85.6003 %	14.3997 %
NEDTA	88.3549 %	11.6451 %

Table V: Performance comparison of NEDTA with ID3, J48 and NBTree

Algorithm	Time Taken to build model (in seconds)
ID3	0.08
J48	0.1
NBTree	7.96
NEDTA	0.06

Table VI: Comparison of error rate of NEDTA with ID3, J48 and NBTree

Algorithm	MSE	RMSE	RAE	RRSE
ID3	0.0991	0.2237	28.3881 %	53.4855 %
J48	0.0787	0.1983	22.4945 %	47.4297 %
NBTree	0.1065	0.2296	30.4526 %	54.9118 %
NEDTA	0.0675	0.1838	19.3113 %	43.9458 %

Figure 9: This graph shows accuracy comparison of ID3, J48 and NBTree algorithms with NEDTA

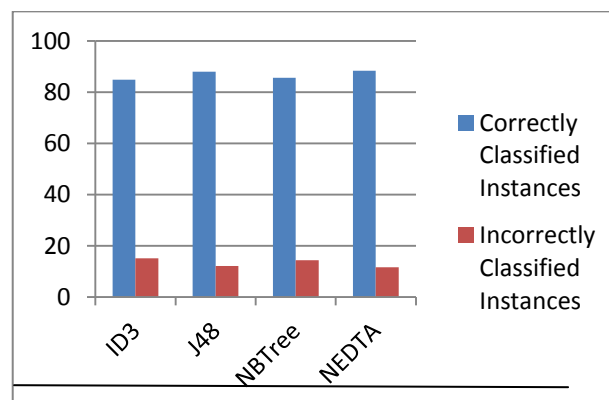


Figure 10: This graph shows performance comparison of ID3, J48 and NBTree algorithms with NEDTA

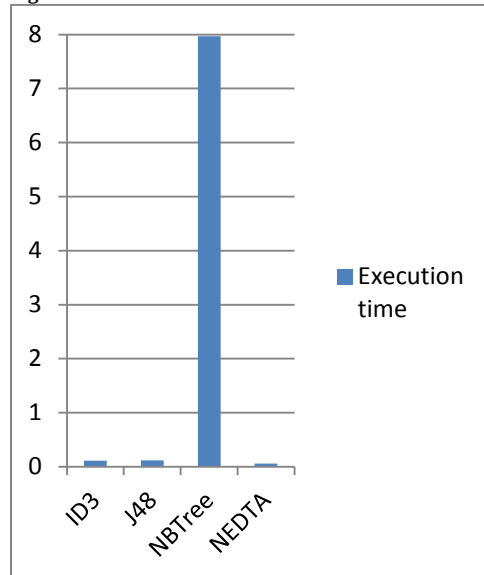
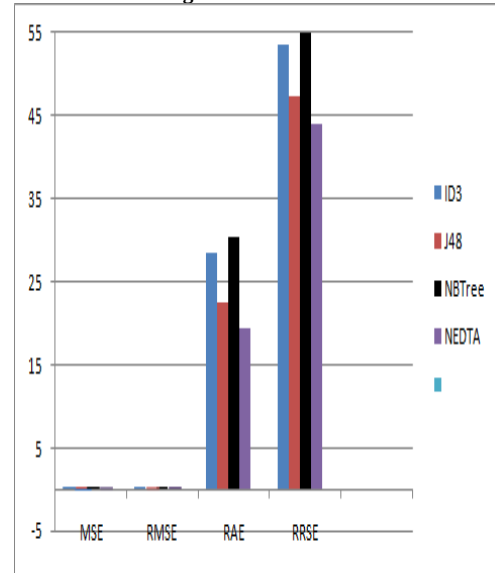


Figure 11: This graph shows Error rate (MSE, RMSE, RAE and RRSE) comparison of ID3, J48 and NBTree algorithms with NEDTA



XI. Conclusion

Data mining plays an important role in knowledge discovery. There are various decision tree algorithms which are used to classify a larger amount of data. Each algorithm has different performance for different data set. While classifying a large amount of data, the performance of some algorithms decreases. Our algorithm removes this problem. The results of NEDTA show better performance in terms of execution time, error rate and accuracy than other algorithms.

References

- [1] Kazi Imran Moin and Dr. Qazi Baseer Ahmed, "Use of Data Mining in Baking", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 2, pp.738-742, 2012.
- [2] J. R. Quinlan, 'Introduction of decision tree', Journal of Machine learning.
- [3] Mrs. Swati .V. Kulkarni, "Mining knowledge using Decision Tree Algorithm", International Journal of Scientific & Engineering Research, Volume 2, Issue 5.
- [4] Youvrajsinh Chauhan, Jignesh Vania, "J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient", International Journal of Computer Trends and Technology (IJCTT), volume 5 number, 2013.
- [5] Bangsuk Jantawan and Cheng-Fa Tsai, "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment", "International Journal of Computer Science and Information Security, Vol. 11, No. 10, October 2013
- [6] Yumin Zhao, Zhendong Niu_ and Xueping Peng, "Research on Data Mining Technologies for Complicated Attributes Relationship in Digital Library Collections", "Applied Mathematics & Information Sciences, An International Journal", Appl. Math. Inf. Sci. 8, No. 3, 1173-1178 (2014)
- [7] Aman Kumar Sharma and Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSSE), Vol. 3 No. 5, pp. 1890-1895, 2011.
- [8] Pardeep Kumar, Nitin, Vivek Kumar Sehgal and Durg Singh Chauhan, "A BENCHMARK TO SELECT DATA MINING BASED CLASSIFICATION ALGORITHMS FOR BUSINESS INTELLIGENCE AND DECISION SUPPORT SYSTEMS", International Journal of Data Mining & Knowledge Management Process (IJDMP), Vol.2, No.5, pp. 25-42, 2012.
- [9] Vivek Bhambri, "Role of Data Mining in Banking Sector", International Indexed & Referred Research Journal, Vol. III, ISSUE-33, pp. 70-71, 2012.
- [10] Milija Suknovic, Boris Delibasic, Milos Jovanovic, Milan Vukicevic, Dragana Becejski-Vujaklija and Zoran Obradovic, "Reusable components in decision tree induction algorithms", Springer, 2011.

Acknowledgment

I express my sincere gratitude to Er. Harpreet Kaur, Assistant Professor in department of computer science engineering at SBBSIET, Jalandhar, Punjab for her stimulating guidance, continuous encouragement and supervision.