Comparative Study of Different Decision Tree Classification Techniques

Mr. R. M. Huddar[#1]

[#]*Dept. Of Computer studies,CSIBER, Kolhapur*

Abstract- In today's world due to availability of huge storage devices and technologies at lower cost it becomes possible to store huge amount of data at lower cost. In banking sector due to digitization huge amount of data is generating, to store this huge amount of data is not problem but to extract useful information from this data is challenging task. There are various data mining tools available for extraction and prediction of information extracted from this data. One of the technologies to classify data and represent it is usage of decision tree. This paper is focused on comparison of various decision tree classification algorithms using WEKA tool.

Keywords- digitization, data mining, extraction, prediction, classification, decision tree, marketing

I. INTRODUCTION

The development of Information technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. Data mining technologies are used in various areas like in healthcare for diagnosis and prediction of disease based on symptoms, product design, customers buying behavior based on previous history and customer attributes, In banking data mining is used for different purposes. One of the most widely used areas of data mining for the banking industry is marketing. The banks marketing department can analyze customer's data for to find out potential customers for different types of products. Data mining is widely used for risk management in the banking industry. Bank professionals has to know whether the customers they are dealing with reliable or not based on historical data available with them and decide whether to approve demanded loan amount or not. Another important data mining application in banking is in fraud detection. Data mining helps to analyze day by day transactions of customers and find out the fraudulent actions and report them. In today's competitive market customer retention is an important than that of acquisition. One of the important applications of data mining in banking is customer relationship management. Data mining helps for customer acquisition, increase customer value and retention of existing customers.

In today's competitive world to stay in the business and make profit apart from regular banking, Banks introduced new services and products which will attract customers to their banks. Due to technological changes and involvement of digital banking, internet applications in the banking services throughout the world, the traditional face-to-face customer contacts are being replaced by electronic points of contact to reduce the time and cost of processing an application for various products [1].

II. DATA MINING

Data mining is a technology used for to extract and predict useful information from huge amount of generated data. Various types of data mining techniques are clustering, classification and association. Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other groups. There are various types of clustering algorithm used like Centroid-based clustering, Hierarchical clustering, Distribution based clustering, Density based clustering.

Association is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

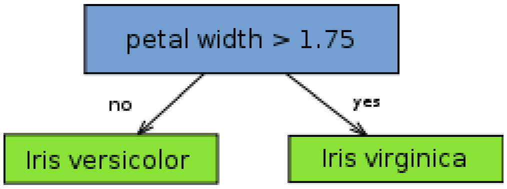
Classification consists of predicting a certain outcome based on a given input. Classification is a supervised machine learning algorithm; it consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except the goal attribute. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how good the algorithm is.

Various classification techniques are decision tree, Neural network and Naive Bayes. A decision tree is a flow chart like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test and each leaf node represents a class label, the paths from root to leaf represent classification rules [4].

III. DECISION TREE ALGORITHMS

Data mining applications has got rich focus due to its significance of classification algorithms, in this study author generated and compared various decision tree algorithms. For the study of generation and analysis of decision tree WEKA toll is used. For Classification WEKA tool provided four different decision tree algorithms.

A) DecisionStump - A decision stump is a machine learning model consisting of a one-level decision tree, that is, it is a decision tree with one root node which is immediately connected to the leaf nodes. A decision stump makes a prediction based on the value of just a single input feature. In DecisionStump depending on the type of the input feature, several variations are possible.



B) J48 – J48 is a simple decision tree learning algorithm developed by Ross Quinlan. The basic idea of J48 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node, in order to select the attribute which is most useful for classifying a given sets. A statistical property called information gain is defined to measure the worth of the attribute. Given a data table that contains attributes and class of the attributes, we can measure homogeneity (or heterogeneity) of the table based on the classes. If a table is pure or homogenous, it contains only a single class. If a data table contains several classes, then it says that the table is impure or heterogeneous. To measure the degree of impurity or entropy,

Entropy = ∑-Pjlog2Pj

Entropy of a pure table (consist of single class) is zero because the probability is 1 and log (1) = 0. Entropy reaches maximum value when all classes in the table have equal probability. To work out the information gain for A relative to S, it first need to calculate the entropy of S. To determine the best attribute for a particular node in the tree, information gain is applied. The information gain, Gain (S, A) of an attribute A, relative to the collection of examples S,

Gain(S,A) =Entropy(S)

After computing information gain of all input attributes, attribute having highest gain is used as root node and process is repeated for all the remaining attributes, Repeat this process until entropy of node reaches to null.

C) Random tree - A random tree is a tree constructed randomly from a set of possible trees having K random features at each node. “At random” in this context means that in the set of trees each tree has an equal chance of being sampled. Or we can say that trees have a “uniform” distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models.

D) Random Forest - Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating.

IV. *WEKA*

WEKA is a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization [6]. For our purpose the classification tools were used. There was no preprocessing of the data. WEKA has four different modes to work in.

- Simple CLI; provides a simple command-line interface that allows direct execution of WEKA commands.
- Explorer; an environment for exploring data with WEKA.
- Experimenter; an environment for performing experiments and conduction of statistical tests between learning schemes.
- Knowledge Flow; presents a “data-flow” inspired interface to WEKA. The user can select WEKA components from a tool bar, place them on a layout canvas and connect them together in order to form a “knowledge flow” for processing and analyzing data.

V. *PROPOSED SYSTEM*

In this section we concentrate on performance of DecisionStump, J48, Random Tree and Random Forest decision tree algorithms. The objective of this comparison is creating baseline which will be useful for the classification scenarios. It will also help in the selection of appropriate model.

Dataset

In this study for classification problems, we took these datasets from the UCI Machine Learning repository [7]. The dataset consist of 45211 instances, 16 + output attribute. Details about attributes is shown by Table1

TABLE I

ATTRIBUTES OF DTASET

a. Personal information attributes

Attribute Name	**Description**	**Data Type**
Age	Age of person	Numeric
Job	Type of job	Categorical
Marital	Marital status	Categorical
Education	Education level	Categorical
Default	Has credit in default	Binary
Balance	Average yearly balance in Euro	Numeric
Housing	Has housing loan	Binary
Loan	Has personal loan	Binary

b. Last contact of the current campaign

Attribute Name	**Description**	**Data Type**
Contact	Contact communication type	Categorical
Day	Last contact day of month	Numeric
Month	Last contact month of year	Categorical
Duration	Last contact duration in seconds	Numeric

c. Other attributes

Attribute Name	Description	Data Type
Campaign	No. Of contacts performed during this campaign and for this client	Numeric
Pdays	No. Of days that passed by after the client was last contacted from a previous campaign	Numeric
Previous	No. Of contacts performed before this campaign and for this client	Numeric
Poutcome	Outcome of previous marketing campaign	Categorical

Output variable Y – has the client subscribed a term deposit? Binary (Yes or No)

Figure1 shows distribution of clients based on various attributes.

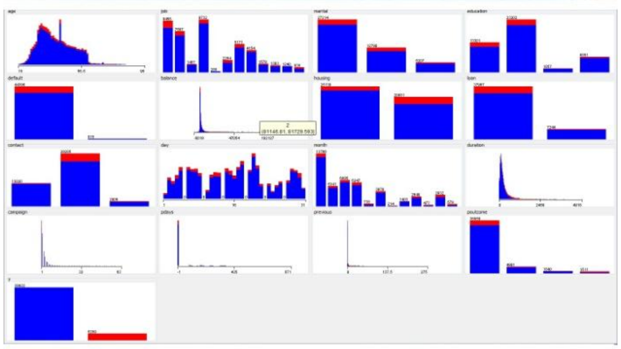


Figure 1: Dataset distribution based on attributes

Y is the target variable in the study, Class "No" clients were not subscribed for term deposit whereas Class "Yes" clients were subscribed for term deposit. Figure2 shows that out of 45211 clients 39922 not opted Term deposit whereas 5289 opted term deposit.

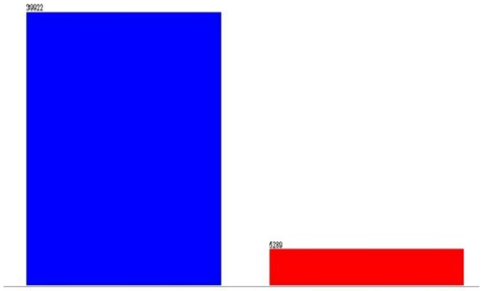


Figure 2: The distribution of target variable- Y

VI. *RESULTS*

In this study various decision tree algorithms are applied and analyzed used validation method. Performance measures for decision tree algorithms mentioned in Table II and III.

TABLE II

FINAL STATISTICS OF DECISION TREE

Decision tree	TP rate	FP rate	Precision	Recall	F-measure	Roc area	Class	Time taken (secs)
Decision Stump	1	1	0.833	1	0.938	0.699	NO	0.54
	0	0	0	0	0	0.699	YES	
J-48	0.959	0.519	0.933	0.959	0.946	0.843	NO	7.68
	0.481	0.041	0.609	0.481	0.537	0.843	YES	
Random Tree	0.929	0.553	0.927	0.929	0.928	0.699	NO	1.21
	0.447	0.071	0.456	0.447	0.452	0.699	YES	

TABLE III

CONFUSION MATRIX FOR ALL DECISION TREE

Decision tree	Mean Absolute error	A	b	Outcome
DecisionStump	0.1825	39922	0	a=no
		5289	0	b=yes
J-48	0.1269	38289	1633	a=no
		2747	2542	b=yes
Random Forest	0.1278	37104	2818	a=no
		2924	2365	b=yes

Some important terminologies of Result are

- N – Total number of classified instances
- True Positive(TP) – correctly predicted of positive classes.
- True Negative(TN)- Correctly predicted of negative classes.
- True Negative (FP) – wrongly predicted as positive classes.
- True Negative (FN) – total wrongly predicted as negative classes.
- False Positive rate (FPR)- negatives in correctly classified/total negatives.
- True Positive rate (TPR) – positive correctly classified/total positives
- Accuracy(A) – It shows the proportion of the total number of instance predictions which are correctly predicted.
  A= (TP + TN)/N
- ROC curve – It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is plotted along the y axis and FPR is plotted along the x axis. Performance of each classifier represented as a point on the ROC.
- Precision (P) - It is a determine of exactness. It is the ration of the predicted positive cases that were correct to the total number of predicted positive cases

P=TP/(TP+FP)

- Recall (R) - Recall is determine of completeness. It is the proportion of positive cases that were correctly recognized to the total number of positive cases. It is also known as sensitivity or true positive rate (TPR).

R=TP / (TP + FN)

- F-Measure- The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

F-measure= (2xrecallxprecision)/(precision+recall)

VII. CONCLUSION

Results shows that for bank dataset Decision Stump classification algorithm takes minimum time to classify data but gives less accuracy. J48 classification algorithm gives results with less errors but it takes more time as compared to Decision Stump and Random tree algorithms. For bank dataset random tree algorithm have good accuracy with little increase in time used for classification. Maximum accuracy is given by J48, but time taken to build classification model is much higher than other classifiers.

REFERENCES

1. Vivek Bhambri "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011
2. Kazi Imran Moin, Dr. Qazi Baseer Ahmed "Use of Data Mining in Banking" International Journal of Engineering Research and Applications (IJERA) vol.2, Issue 2, March- April 2012
3. T.M. Mitchell, Machine Learning. McGraw-Hill, 1997.
4. Wikipedia
5. Data Mining Algorithms for Classification BSc Thesis Artificial Intelligence Author: Patrick Ozer Radboud University Nijmegen January 2008
6. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005
7. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.
8. Purva Sewaikar, Kamal Kant Verma Comparative study of various decision tree classification algorithm using WEKA, International Journal of Emerging Research in Management & Technology ISSN:2278-9359(volume-4,Issue-10)
9. Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood Random Forests and Decision Trees IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814
10. Kamath RS, Kamat RK "Modelling Fetal Morphologic Patterns through Cardiotocography data: A Random Forest based Approach" Research Journal of Pharmaceutical, Biological and Chemical Sciences ISSN: 0975-8585 Sept. Oct. 2016