International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-10)

Research Article | **October 2015**

Comparative Study of Various Decision Tree Classification Algorithm Using WEKA

Purva Sewaiwar, Kamal Kant Verma
Uttrakhand Technical University,
Dun, Uttrakhand, India

Abstract-

This paper is focused on comparison of various decision tree classification algorithms using WEKA tool. Data mining tools such as classification, clustering, association and neural network solve large amount of problem. These are all open source tools, we directly communicate with each tool or by java code. In this paper we discuss on classification technique of data mining. In classification, various techniques are present such as bayes, functions, lazy, rules and tree etc. . Decision tree is one of the most frequently used classification algorithm. Decision tree classification with Waikato Environment for Knowledge Analysis (WEKA) is the simplest way to mining information from huge database. This work shows the process of WEKA analysis of file converts, step by step process of weka execution, selection of attributes to be mined and comparison with Knowledge Extraction of Evolutionary Learning . I took database [1] and execute in weka software. The conclusion of the paper shows the comparison among all type of decision tree algorithms by weka tool.

Keywords – Data mining, Classification Algorithm, Decision tree, J48, Random forest, Random tree, LMT, WEKA 3.7.

I. INTRODUCTION

Data mining is a collection of techniques to glean information from data and turn into meaningful trends and rules to improve your understanding. The basic principles of data mining is to analyze the data from different direction, categorize it and finally to summarize it .Today we are living in digital world where data increasing day by day, to get any information from mountain of database is not only difficult but mind blogging also. To deal with this huge data we need data mining techniques. Data mining [2] define as the process of analysing, searching data in order to find contained, but prospective information. data mining is used to find the hidden information prototype and relationship between the large data set which is very useful in decision creation. The advantages of data mining are ;analysis routinely, results of analysis is objective, accuracy of data is constant. Data mining also known as knowledge discovery in database (KDD), mainly data mining follows these steps; Data cleaning, Data integration, Data selection, data transformation, data mining, pattern evolution, knowledge evolution data reduction. Data mining having various numbers of techniques which have own speciality, such as clustering, data processing, pattern recognition, association, visualization etc.

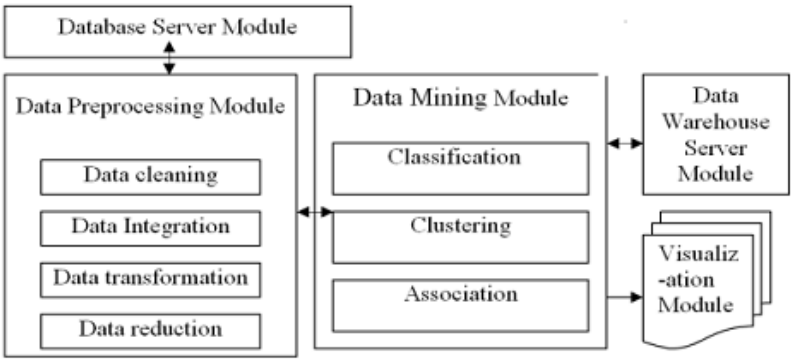

Fig.1. Data mining techniques .[3]

II. CLASSIFICATION

Classification is possibly the most frequently used data mining technique. Classification [4] is the process of finding a set of models that describe and differentiate data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown. There are many algorithms that can be used for classification, such as decision trees, neural networks, logistic regression, etc. In this work we are using decision tree algorithm for classification.

The Classification process involves following steps:

- Create training data set.
- Identify class attribute and classes.

- Identify useful attributes for classification (Relevance analysis).
- Learn a model using training examples in Training set.
- Use the model to classify the unknown data samples.

III. DECISION TREE

Decision trees [5] are a way of representing a sequence of rules that lead to a class or value. Decision Tree is a flowchart like tree structure.

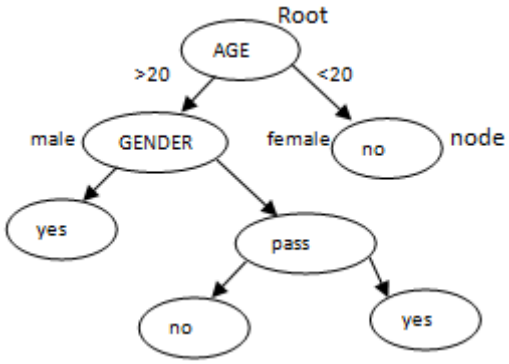

Fig.2. Decision tree.

The decision tree consists of three fundamentals, root node, internal node and leaf node. Top most fundamental is the root node. Leaf node is the terminal fundamental of the structure and the nodes in between is called the internal node. Each internal node denotes test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. Various decision tree algorithms are used in classification like ID3,AD Tree, REP, J48, FT Tree, LAD Tree, decision stamp, LMT, random forest, random tree etc. In this work following trees take for comparison-

A. ***J48***- A predictive machine-learning model which decide the target value of a new sample based on different attribute values of the available data is J48 decision tree [6] . the different attributes denote by the internal nodes of a decision tree, the branches between the nodes tell us the possible values that these attributes can have in the experimental samples, while the terminal nodes tell us the final value of the dependent variable.

B. ***LMT***- A classification model with an associated supervised training algorithm that combines logistic prediction and decision tree learning is logistic model tree (LMT)[7] . Logistic model trees use a decision tree that has linear regression models at its leaves to provide a section wise linear regression model.

C. ***Random Forest***- Random forests[8] are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean prediction of the individual trees. Random forests correct for decision trees' habit of over fitting to their training set. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase and some loss of interpretability, but generally greatly boosts the presentation of the final model.

D. ***Random tree***- A random tree [9] is a collection of tree predictors that is called forest. It can deal with both classification and regression problems. The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the classifier response is the average of the responses over all the trees in the forest. All the trees are trained with the same parameters but on different training sets.

IV. WEKA

WEKA[10] is a data mining software developed by the University of waikato in New Zealand that apparatus data mining algorithms using the JAVA language. Weka is a milestone in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption. Weka is a bird name of newzealand. WEKA is a modern feature for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data pre processing tools to researchers. The algorithms are directly to a database. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules; It also includes visualization tools. WEKA would not only afford a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. WEKA is open source software issued under General Public License [11]. The data file normally used by Weka is in ARFF file for-mat, which consists of special tags to indicate different things in the data file foremost: attribute names, attribute types, and attribute values and the data. For working of WEKA we not need

the deep knowledge of data mining that's reason it is very popular data mining tool. Weka also provides the graphical user interface of the user and provides many facilities .The GUI Chooser consists of four buttons—one for each of the four major Weka applications.
The buttons can be used to start the following applications:

- **Explorer** : It is the main interface in Weka. It has a set of panels, each of which can be used to perform a certain task .Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.
- **Experimente**r: An environment for performing experiments and conducting statistical tests between learning schemes.
- **Knowledge Flow**: This environment supports essentially the same functions as the Explorer but with a drag – and drop interface. One advantage is that it supports incremental learning.
- **Simple CLI**: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

Fig. 3. WEKA tool front view.

A. Execution in weka- Execution on weka is a step by step process. First is data loading, Data can be loaded from various sources, including files, URLs and databases. WEKA has the capacity to read in ".csv" format. firstly we take excel datasheet from real world , the first row contains the attribute names (separated by commas) followed by each data row with attribute values listed in the same order (also separated by commas), convert in .csv file format. Than go to the explore button on weka and save this .csv file. once data loaded into WEKA, the data set automatically saved into ARFF format.

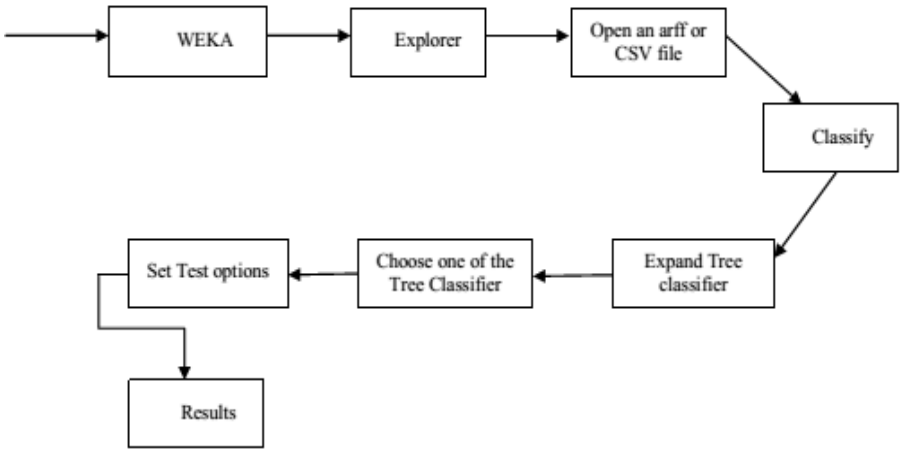

Fig. 4. Execution in weka tool

- Choosing The Data From File, After data is loaded, WEKA will recognize the attributes and during the scan of the data will compute some basic statistics on each attribute. the list of recognized attributes, while the top panels indicate the names of the base relation (table) and the current working relation. left panel will show the basic statistics on that attribute Click on any attribute . For categorical attributes, the frequency for each attribute value is shown; while for continuous attributes we can obtain min, max, mean, standard deviation, etc.
- Prepare the Data to Be Mined, Selecting Attributes From sample data file, each record is individually identified by attribute and using the Attribute filter in WEKA. In the "Filters" panel, click on the filter button (to the left of the "Add" button). This will show a popup window with a list available filters. Scroll down the list and select weka.filters.AttributeFilter"
- After setting filters, go to the classification button and click on it .This will show a popup window with a list of classification algorithm, expand decision tree on this and select the tree which one u want to experiment.

Before implementation of database some important terminology 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) – positives correctly classified/total positives.

- ***Accuracy (A)***: It shows the proportion of the total number of instance predictions which are correctly predicted

$$A = \frac{TP + TN}{N}$$

- ***Receiver Operating Characteristic (ROC) Curve***: It is a graphical approach for displaying the trade off 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 curve.
- ***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.

$$P = \frac{TP}{TP + FP}$$

- ***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).

$$R = \frac{TP}{TP + FN}$$

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

$$F - measure = \frac{2 \times recall \times precision}{precision + recall}$$

V. RESULTS

The cross validation method used to analysis for the datasets. Various performance measures for all the datasets mentioned in Table I, II, III. Comparative analysis of various decision tree classification, simulation results as follows -

Table 1. Final statistic of decision tree

Decision tree	TP Rate	FT Rate	precision	recall	f-measure	Roc curve area	CLASS	Time taken (sec)
J48	1	0	1	1	1	1	Y	0.14
	1	0	1	1	1	1	N	
Random forest	0.838	0.014	0.969	0.838	0.899	0.964	Y	0.07
	0.986	0.016	0.924	0.924	0.954	0.962	N	
Random tree	0.838	0.014	0.969	0.838	0.899	0.976	Y	0.01
	0.986	0.162	0.924	0.986	0.954	0.971	N	
LMT	1	0.014	0.974	1	0.987	1	Y	6.9
	0.986	0	1	0.986	0.993	0.99	N	
Decision stump	1	O	1	1	1	1	Y	0.18
	1	0	1	1	1	1	N	

Table. 2. Confusion matrix for all decision tree

Decision tree	Mean absolute error	a	b	Parametric variable	outcome
J48	0	37	0	la	YES
		0	74	lb	NO
LMT	0.0433	37	0	la	YES
		1	73	lb	NO
Random forest	0.2242	35	2	la	YES
		0	74	lb	NO
Random tree	0.3216	11	26	la	YES
		1	73	lb	NO
Decision stump	0	37	0	la	YES
		0	74	lb	NO

Table 3. compression of weight avg. for decision tree

Decision tree	TP Rate	FP rate	precision	recall	f-measure	ROC area
J48	1	0	1	1	1	1
LMT	0.991	0.005	0.991	0.991	0.991	0.993
Random forest	0.982	0.036	0.982	0.982	0.982	0.998
Random tree	0.757	0.473	0.797	0.757	0.712	0.702
Decision stump	1	0	1	1	1	1

VI. CONCLUSION

Results shows that Decision stump classification algorithm takes minimum time to classify data but gives less accuracy. J48 have quite good accuracy with a little increase in time used for classification. maximum accuracy given by LMT, but time taken to build classification model is much higher than other classifiers or we can say maximum in all the classifiers in most of cases. Rest of the models also lies in between the best and worst ones .In this paper Decision Tree classification algorithms are analysing and justification method to explain the results. The specific approaches for classification are characterized, we developed the WEKA method is based on choosing the file and selecting attributes to convert .csv file to flat file and discussed features of WEKA performance. Our work extends to utilize the implementation of different dataset. Each decision tree classify the data correctly and incorrectly instance. We can use these decision tree algorithms in medical , banking , stock market and various area.

ACKNOWLEDGEMENT

I would like to express my deepest thanks to all those who provided me the possibility to complete this paper. A special gratefulness give to my guide, Mr. K.K. Verma,(assistant prof.,)whose contribution in stimulating suggestions and encouragement, helped me to coordinate my work especially in writing this paper.Furthermore I would also like to acknowledge with much appreciation the crucial role of my family & friends, who gave the full effort in achieving the goal. I have to gratitude the guidance given by Mr. Satendra Kumar & Shushir Sangal to permission to use all the necessary equipment to complete the task . Last but not least, many thanks go to the god to giving me strength and courage to complete this paper.

REFERENCE

[1] Database
[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
[3] Swati singal ,monika jain:a study on weka tool for data preprocessing, classisfication and clustring ‘’intreranation journal of innovation technology and exploring enginnering.2013
[4] King, M., A., and Elder, J., F., Evaluation of Fourteen Desktop Data Mining Tools, in Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics, 1998.
[5] An Implementation of ID3: Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, and Australia.
[6] Wikipedia contributors, “C4.5_algorithm,” Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 28-Jan-2015.
[7] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees,” Mach. Learn., vol. 59, no. 1–2, pp. 161–205, 2005.
[8] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[9] Wikipedia contributors, “Random_tree,” Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 13-Jul-2014.
[10] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, “Weka,” in Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 1305 –1314.
[11] Pallavi, SunilaGodara “A Comparative Performance Analysis of Clustering Algorithms”International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3, pp. 441-445

