An Analysis on the Performance of Enhanced Decision Tree Algorithm (EDTA) and Comparison with C4.5 Algorithm

A Dissertation Proposal submitted by

Sunakshi Sharma
(Registration No. 11310338)

To

Department of CSE/IT

In partial fulfilment of the requirements for the award of the Degree of

Master of Technology in Computer Science Engineering

Under the guidance of

Mrs. Alpana Vijay Rajoriya
Assistant Professor (CSE/IT)

Lovely Professional University, Punjab

May, 2015

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	: Sunakshi Sharma	Registration No	: 11310338
Batch	: 2013-2015	Roll No	: RK2306A06
Session	: 2014-2015	Parent Section	: K2306

Details of Supervisor:

Name	: Alpana Vijay Rajoriya	Designation	: Assistant Professor
UID	: 17447	Qualification	: M.Tech
		Research Exp.	: 1 year

Specialization Area: Database (pick from list of provided specialization areas by DAA)

Proposed Topics:-

[✓ 1. An analysis on performance of enhanced decision tree algorithm using a data set.]

2. An enhancement on decision tree algorithm.
3. An improvement on decision tree algorithm.

17447

Signature of supervisor

PAC Remarks: Topic 1 approved. Paper expected.

11057

APPROVAL OF PAC CHAIRPERSON: Signature: 11041 Date: 24/11/14

*Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to supervisor.

ABSTRACT

There are several techniques that are used in data mining, each one having advantages but also disadvantages. To find out which one is most appropriate for our case, when we want to use our databases in a decision-make process we need to have information about our data business and data mining techniques. Alternatively we can try them all and find out which one is the best in our case. The dataset used in this research is based on mobile environment obtained from WirelessMon software. This report is based on the findings maximum use of mobile service. The results in this report are based on data from mobile service related. There are various algorithms used for creating decision trees such as ID3, CART and C4.5. Along with this we can create our own algorithm for making decision tree which could be used for prediction and analysis. Every algorithm depends on some or the other splitting criteria and the way how pruning is done. The goal of this research is to look at one particular self-developed decision tree algorithm called enhanced decision tree algorithm (EDTA) and compare its results with the existing c4.5 algorithm. The results would be compared in the terms of accuracy that how accurate are the results produced by both the algorithms. The tendency to build accurate decision tree will be same for these algorithms when used with any other dataset.

CERTIFICATE

This is to certify that **Sunakshi Sharma** has completed M.Tech dissertation proposal titled **An Analysis on the Performance of Enhanced Decision Tree Algorithm (EDTA) and Comparison with C4.5 Algorithm** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature of Advisor

Name: Alpana Vijay Rajoriya

UID: 17447

ACKNOWLEDGEMENT

I would like to present my deepest gratitude to **Mrs. Alpana Vijay Rajoriya** for her guidance, advice, understanding and supervision throughout the development of this dissertation study. I would like to thank to the **Project Approval Committee members** for their valuable comments and discussions. I would also like to thank to **Lovely Professional University** for the support on academic studies and letting me involve in this study.

DECLARATION

I hereby declare that the dissertation proposal entitled **An Analysis on the Performance of Enhanced Decision Tree Algorithm (EDTA) and Comparison with C4.5 Algorithm** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 02 May 2015

Investigator: Sunakshi Sharma

Registration No. 11310338

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Remote resources such as computers, databases, files etc. along with people like analysts, professionals, end users are often involved in the complex process of analysis of data. This analysis is in an omnipresent manner and is very important for applications which deals in finance, process control, defence and many more domains. The ability to analyse large data amount is the demand of these applications. Decision tree a data mining technique which are CART,ID3 and C4.5 as are scalable and fast and are for data streams monitoring from omnipresent devices such as computers, palmtops etc.

1.1 Databases

The collection of information in such a format that a computer programme is able to select required data entry quickly is known as databases. The information in a database is organized in such a manner that it can be easily managed, updated and accessed. Computer databases contain aggregations of data files or records likes, product inventors, employee details, costume profiles and other transactions. There are various databases approaches namely traditional databases, relational databases and an object-oriented programming databases. Along with this distributed databases is also very common, it is the database which can be replicated or dispersed among various points in a network. A system that enables to access, organize and select data in a database in known as database management system. Many database management systems types are available from small system which run on personal computers or huge systems that run on mainframes.

1.2 Data Warehouse

The concept of data extracted from various databases, operational systems and other resources for use as historical snapshots for schedule reporting and ad-hoc queries is known as data warehouse. Current and historical data which is used for creating reports for certain decision of higher management such as annual comparisons is stored in data warehouse. Basically in computing a system designed for reporting and data analysis is data warehouse. Basically two types of approaches are used to integrate various heterogeneous databases namely

1. Query driven approach
2. Update driven approach.

The process of building and using data warehouse is known as data warehousing. Various processes involved in data warehousing is data cleaning, data consolidations and data cleaning.

1.3 Difference between Data warehouse and databases.

Both data warehouse and databases contains data and information in the form of tables and along with this both have indexes, views, keys, joins etc. but yet data warehouse house is different from databases. The basic difference between both of them is that data warehouse is optimized to answer questions regarding analysis which is critical and on the other hand data bases are optimized and designed for record keeping. Database is application oriented whereas data warehouse is subject oriented. Other difference is between the processing systems. Data warehouse is designed as an online analytical processing system (OLAP) which contains read-only data which can be analysed more efficiently. On the other hand databases are designed as online transaction processing system (OLTP) where record of every transaction is made. Database is used for day-to-day operations whereas data warehouse is used for long term operation and decision support. The users for both database and data warehouse are different. Database is used by clerks, DBA, database professionals whereas data warehouse has common users as knowledge workers examples analysts, executive and manager. Apart from all the differences even the size of both makes a huge difference, database has size of 100 MB to GB. Lastly, the design of database is Entity-Relationship Diagram based and view of data in database is flat relational whereas the design of the data warehouse is done using star/snowflake schema and the view of data in data warehouse is multidimensional.

1.4 Data Mining

An analytical process which is developed to examine data in form of patterns which are consistent is known as data mining. From ages only the physical excerption of patterns from data is going. The data collection, storage and manipulations has increased by the accretion and prevalence computer technology. Due to the grown size as well as complexity of datasets the direct manual analysis has amplified with indirect and automatic processing of data. These are various methods as clustering, neural networks, other genetic algorithms, vector support machines and decision trees which are applied to data with aim of not hiding the patterns which are hidden. Data mining is sometimes turns as knowledge discovery and its tools are here to predict behaviours and future trends making proactive business and knowledge driven decisions. These sophisticated data analysis tools used by data mining are to discover not the known, patterns and relationships of them in large datasets. Various models as statistical and mathematical along other machine learning methods can be involved as tools of data mining. These are basically the algorithms which improves their performance automatically by experience such as neural networks and decision trees.

There are basically three stages of data mining process.

1. Initial Exploration.

2. Model building and validation.

3. Final deployment.

Variety of parameter are used by data mining applications to examine the data. Hence data mining can be performed on data which is represented on any form like quantitative, textual or even multimedia form. These various parameters includes

1. Associations rules where connecting of event is there that is one connected with other example the events of purchasing bread and butter.

2. Path or sequence analysis pattern where one event is leaded by another event, such as child brother and diapers which are purchased that time.

3. Classification-In which new patterns are identified such as the coincidence made by purchasing plastic sheet and also cello tapes.

4. Clustering-Where various clusters are made of unknown puts to know them better such as geographical locations and perfumes.

5. Forecasting- Which is discovering of patterns for making reasonable predictions for the future activities such as people may buy automatic washing machines in rainy season.

1.5 Techniques of Data Mining

The three main techniques of data mining are

1. Clustering- Clustering is the process in which the objects of similar kind are grouped together into various classes which are termed as clusters. Various customer groups are discovered by the cluster analysis and the characteristics of each group is also analysed. This is the common technique used for market analysis.

2. Classification- The technique involved in predicting certain outcome based on any given input is known as classification. This approach involves certain processes of mining which are made to discover rules which are used to define the sub processes of the technique which bare model building and predicting. In this terms are belonged to class or particular subset of data.

3. Association Analysis- The analysis which shows the association rules discovery giving value conditions of attribute which are constantly occurred in a given dataset. This analysis is very much popular in transaction data analysis and market basket.

1.6 Decision trees

There are variety of algorithms being used in classification technique. One if these is the decision tree approach. To represent both the regression models and classifiers decision tree in the state of predicative model is used. Decision tree basically us the hierarchal model of decisions and their consequences. The structure of decision tree includes branch, root node and leaf node. Attributes test is denoted on each interval node, the test outcome is denoted by branch and class labels are shown by leaf node. The topmost node is the root node of the tree. The tree learning is done by dividing the source into set which are generally based on a test of attribute value. The top down approach of decision tree sets an example of greedy algorithm. Apart from this bottom-up approach is also common these days.

Definition of decision trees can also be on the basis of combination of computational and mathematical techniques for getting the categorisation, description and generalisation of a given dataset.

There are mainly two types of data trees used in data mining.

1. Classification tree analysis- It is done when the class to which data depends in the predicted outcome.

2. Regression tree analysis-It is done when a real number can be taken as the predicted outcome example (The cost of a building)

To refer both of these procedures the term classification and regression tree CART analysis is used. Trees used for both regression and classification are same at some perspective but along with this they have differences too such as procedures which are used to determine the split point.

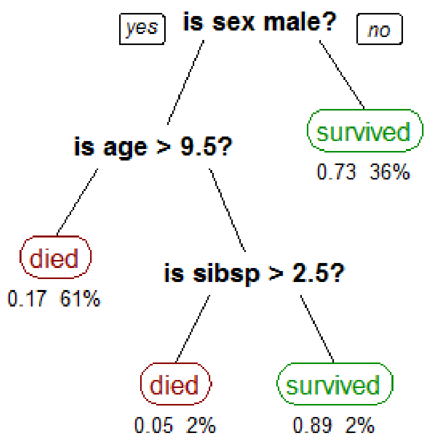


Figure 1.1 Example of decision tree

Many decision tree algorithms are as follows:

- ID3 Algorithm – The beginning of the algorithm is done by simply setting the root node to the original set. Basically the strategy of this algorithm is selecting the attribute with highest information gain and consider that attribute for split. The probability of occurrence is related to the information amount which is collaborated with the attribute value. After the selection of attribute the entropy is measured which is basically the amount of information. The uncertainty amount and randomness of an attribute is measured using entropy. If all the data belongs to same class then the entropy will be zero.

- C4.5 Algorithm – C4.5 is basically the ID3 algorithm's extension. It creates tree of any depth. The decision tree made by c4.5 is the result of recursively data partitioning of data present in the dataset. C4.5 uses depth first strategy for its decision. This algorithm takes into account all the tests which are feasible for splitting the dataset and out of them it selects the one that gives most suitable information gain. C4.5 considers both discrete and continuous attributes. One test with all the number of outcomes of distinct values is examined for discrete attributes and binary test is considered with distinct value of attributes for continuous attribute. The training data which belongs to the node is being sorted for continuous attribute for gathering the entropy gain efficiently for the binary tests. C4.5 performs post-pruning that is after the tree creation it goes back again through the tree for removing extra branches which are not useful by replacing them with leaf node.

- CART – It is a decision tree technique which is non-parametric and produces either regression trees or classification trees. If the dependent variable is numeric then regression tree is produced and if the dependent variable is categorical then classification tree is produced. The collection of rules which are based on variables of dataset are used to create the decision tree. The best split is selected by the rules based on value variables. After this the node is split into two which is further continued to all the child nodes. When no further gain is detected by CART then the splitting ends.

- CHAID – CHAID algorithm only uses nominal data. Along with this it also uses ordinal categorical data. If the predicates are continuous then for using them in this algorithm they are first converted to ordinal categorical data. Splitting attribute is selected by comparing the p-value which is adjusted with an associated variable. Multi-way splits are used by this algorithm. This algorithm is based on bonferroni testing. Along with prediction and classification CHAID can also be used for detecting the amount of interaction between the variables. The major advantage of this algorithm is that its output is easy to interpret and is highly able to visualise.
- MARS – In this algorithm flexible models are made using piecewise linear regression. This algorithm is based on the concept of spline knot, which is where one model of local regression is given to other model and there the point of intersection of spline is created. The search of knots are generalised using basis functions, these functions are basically the set of various functions used to present the information laying in various variables. This model creates the basis functions in pairs. Firstly the tree is over-fitted and then it is pruned for the solution which is optimal. An upper limit is described on the maximum number of basis functions.

1.7 Pruning in Decision Trees

Pruning is a technique used for reducing the size of the tree by removing those tree sections which do not provide any useful information for classifying the instances. By the process of pruning the complexity by which the accuracy of the prediction is increased. Pruning can be done by two types basically top down pruning and bottom up pruning. In top down pruning the nodes will be traversed starting from the top that is from the root itself, and in the bottom up pruning the traversing is started from the child node.

There are two forms of pruning used for error and cost of the tree.

- Reduced error pruning – this form of pruning is the simple pruning. In this type of pruning replacement of each node is done by its immediate upper class. If after doing so there is no change in the accuracy of the prediction then the change is kept. This type of pruning is good for speed and simplicity.

- Cost complexity pruning – A tree series is generated from $D_{0.}$ . . D_M where D_M is the root and D_0 is the initial tree. In this the creation of tree is done by removing the sub trees and replace it with the chosen value leaf, which is done by any algorithm. From the tree series the best tree is selected by calculating the accuracy by the training data set and with the help of cross-validation.

There are two types in which the tree pruning can be done

- Pre-pruning – In this type of pruning the tree is pruned while it is building. It looks for the unwanted nodes and remove them while building of tree only.
- Post-pruning – In this type of pruning first the tree is constructed and then the tree is traversed again for pruning. It is more beneficial as it gives the liberty to the tree to classify the dataset and then remove the unwanted nodes.

Chapter 2

REVIEW OF LITERATURE

Qiang Yang *et al* (2007)**:** Qiang Yang have discovered an unique algorithm that suggest actions which changes customers undesired status to a desired status such as from attractors to loyal. They have achieved this by maximizing the net profit which is expected and this is their main objective function that is the expected net profit. The main purpose for designing the algorithm was that certain technique which were applied to various problems of industry such as customer relationship management requires human experts for post processing the manually generated knowledge. And many of the post processing technique do not suggests actions directly which would lead to profit inverse rather these are limited to production of visualization results and interesting ranking. The approach used in this integrator decision making and data mining tightly by developing the decision making problems directly on top of results of data mining in the post processing step. They have conducted the factual tests on both UCI benchmark data and a realistic insurance application. As the data set used by them in form UCI. The final objective of their algorithm in research is to maximize the profit while reducing the cost. [1]

Mingquan Ye *et al* (2013)**:** In this paper an innovative multi-level rough set model (MLRs) which is based on attribute value taxonomies and a program of full sub tree generalization is presented. The researchers have compared the results of MLRs with that of the Pawlak's rough set model. Along with this another different concept of cut reduction which is based on MLRs is introduced. According to researchers a cut reduction has the ability to reduce the multi-level decision table which is more abstract, the reduction is done with the classification ability which is same on the decision table which is raw. The main focus of the researchers is to enhance the simple-level Pawlak's rough data set model on a concept of multi-level rough set model(MLRs).The cut reduction in MLRs evaluation has n-hard problem and for computing the cur reduction a CRTDR algorithm is presented. The experimented results of the research proved the powerfulness of the methods proposed by the researchers. Further they have researched on the extension of the proposed model for discovering multi-level decision rules and how other rough sets can be extended in association with attribute value taxonomies. [2]

Gilad Katz *et al* (2014)**:** Gilad Katz have developed a method named confDtree (confidence-based decision tree) which can be used for the three drawbacks of decision tress. According to the researchers there are these problems which effects decision trees which are performance reduction while dealing with the small training set; criteria of decision tree is very solid and exact; and that a single uncharacteristic attribute sometimes results in derailing of the process of classification. ConfDtree is a post processing method which has the liberty to classify the instances outliers of decision tress in a better way. The researcher stated that the predictive performance of decision trees is increasing steadily and powerfully. The average improvement calculated for minor, in equal or multi-level class dataset is from 5% tom 9%.When reported in the performance of AUC. For making the method able to select appropriate algorithm for particular dataset along with maintaining the gained benefits which are introduced by using confidence intervals, it is important that the method has the facility to integrate with every algorithm of decision tree. There are mainly two drawbacks of the proposed algorithm: firstly the algorithm makes a small increase in the computer that cost used for classification of new instance and secondly the reduction of the comprehensibility of the model is done by also. [3]

Jasna Soldic-Aleksic *et al* (2012)**:** Jasna Soldic-Aleksic have provided the results of the application which is a combination of two models of data mining namely Kohonen Self-Organizing map (SOM) and CHAID. The result provided was for the problem of clustering in the marketing sphere. Kohonen SOM model was used by the researcher for visualizing and making clusters of market data. Further the researcher used the results for the analysis using the CHAID algorithm. The combination of two models was used because according to the researcher CHAID model is an efficient interpreter of visuals for the cluster results depicted by SOM. This two phase technique can be used in studying various aspects of markets, as open survey of market will help to get the output according to customer needs because customer needs can be studied easily by using this approach of two different models which are combined together. [4]

Ji Dan *et al* (2010)**:** Ji Dan have developed an incorporated algorithm of data mining named as CA. This algorithm improves the initial methods of C4.5 and LURE. The algorithm uses the principle component analysis (PLA), parallel processing and grid partition to get reduction of feature and scale for data sets which are large. The researchers have developed this algorithm for maize seed breeding and their experimental result proves that original methods are not that better as their approach. For this research they have assembled large amount of agricultural information data which is used for vast territory and diversity of crop resources. Due to agricultural distinctiveness such as crop resources complexity, consequences among thickness, climate, fertilize thickness and lack of useful tools researches used only small quantity of data. The main objective of their research is to help people in order to analyse and collect useful information for seed breeding .And according to them data mining development to agriculture is a new research point. [5]

Minas A. Karaolis *et al* (2010)**:** Minas A.Karaolis developed a system in data-mining which was used for the judgement of risk factors which related to the events of heart, aimed in reducing the coronary heart disk factors. Coronary Heart Disease nowadays is one of the main cause of death in many countries. The researchers used 528 cases which they assembled from the pathos district which is in cypress. They used the C4.5 algorithm of decision tree for the events by using five criteria of splitting. The necessary risk factors which were collected from the analysis were for MI for PCI (Percutaneous coronary intervention) and for CABG (coronary artery bypass graft) surgery. The right classifications percentages got were 66% for MI model, 75%for PCI model and 75% for CARB model. By their paper the researcher stated that for diagnosis of high and low risk factors, data mining techniques could be used. [6]

Hillol Kargupta *et al* (2001)**:** Hillol Kargupta presented a new Fourier analysis based technique is presented on this paper. This technique is used to enhance the connection between the mobile network and decision trees. In this paper the researcher found that the numeric functional representation of the function as decision trees has many advantages in Fourier basis. They observed that the representation of the function is easy to compile and is more profitable. According to the paper the approach described by the authors provides a new angle to look and understand about decision tree which is totally opposite from that of the original decision tree presentation which was used in various soft wares of data mining. The main purpose for this research was to develop an approach which could be used in various small screen mobile devices. They named this approach as touch screen and ticker-based approach and found that touch screen approach can be used in touch screen devices. The researchers approach can be used in touch screen devices. The researchers now are further working on the ticker approach to mine precision financial data online. This can be considered as the future scope for this research and further work can be easily done on this approach. [7]

Duong Van Hieu *et al* (2014)**:** In this paper Doung Van Hieu used decision tree techniques to predict behaviour of various facebook users by analysing the factors like internet, their age, gender, income education and other personal details. The main aim of this research was to calculate the amount of time which the users will spend on facebook in next year as compared to the time spent this year, and another was to calculate the impact level of facebook on users in next year as compared the impact of current year. The researcher concluded that people of age greater than 39 years will spend same time as are spending now, the people of age between 20 to 39 years will spend more time on facebook and the people of age less than 20years and they do not use facebook on mobile devices will spend less time. Along with this he concluded that the men with higher education higher impact level whereas the men with lower education will have same impact level and the female who have do not have child under 18 will get more impact level. For this research the data was collected by pew research centre. There was no new technique proposed rather the researcher used the existing technique WEKAJ48 classifier, which is generally an implementation of C4.5 algorithm. By this technique researchers got a decision tree having 74.79% accuracy. [8]

Chapter 3

PRESENT WORK

Present work is mainly targeted on a proposed technology which is typically based on the algorithms of data mining are used for decision trees induction which is an algorithm of classification approach

3.1 Problem Formulation

The proposed technology is very well suited for various reasons

1. Enhanced decision tree algorithm which will work on large scale high dimensional dataset- there is a problem of data mining in the classification of large datasets. There is no such algorithm stated that performs well in this problem. An algorithm can be made with certain split selection methods involved from the literature which includes algorithms like C4.5.
2. Enhancement in the efficiency of decision tree construction- various techniques are proposed which can help in the improvement of decision tree construction.
3. Analysis between the computation times- the computation time can be reduced by making alterations in the number of node and leaves. Lesser the no of nodes lesser will be the computation time of the algorithm.
4. Reducing present error rate- the errors rates produced by a predictive model can be reduced by the algorithms. Basically error rate is one minus accuracy.

3.2 Scope

This proposed technology will help the mobile service providers to analyse the amount of services present at any location and will tell that which services are more used by users. This will include all the services which are generally provided by the mobile networks and with this they will get to know that which service has less frequency at which location. By this analysis the service providers will be able to have required services at all various possible locations. The proposed technology will use an enhanced algorithm of decision trees which is an algorithm of classification approach and will perform a predictive analysis to analyse the services at various locations. This algorithm will predict that which service is used more. The prediction is done on various stations like airport, store, movie, restaurant, station. This analyses is done to predict that where the users are using more services and what kind of services are more wanted in these locations. The results for this can also be provided by other algorithms but the results by EDTA are probably more accurate, this is proved by the comparison of this algorithm with another one.

3.3 Objectives of the study

The main objectives for the research focuses on the analysis of data in mobile environment with the help of decision tree technique are as follows

1. To propose an idea of complex activities which study the continuously changing behaviour patterns of mobile users. The idea will predict more accurate results when compared to other algorithms.
2. To analyse different activities which may exhibit dependencies that affect user behaviours.
3. To propose new methods for analysis of the services and predict results.
4. To analyse users activities which will help service providers to provide high quality of services to user at right place and right time.

3.4 Research Methodology

Research methodology is the organized way to solve a research problem. It is a conceptual way which tells that how the research is done by the researcher**.** When we talk about research methodology, the logic behind the methods we use in our research study and also the reason why we are using this method and why we are not using others is discussed. Quantitative research methodology is used in our research to get good results.

The methodology of the research is easy to understand using a flow chat. Therefore a proper flow chart of our research methodology is provided

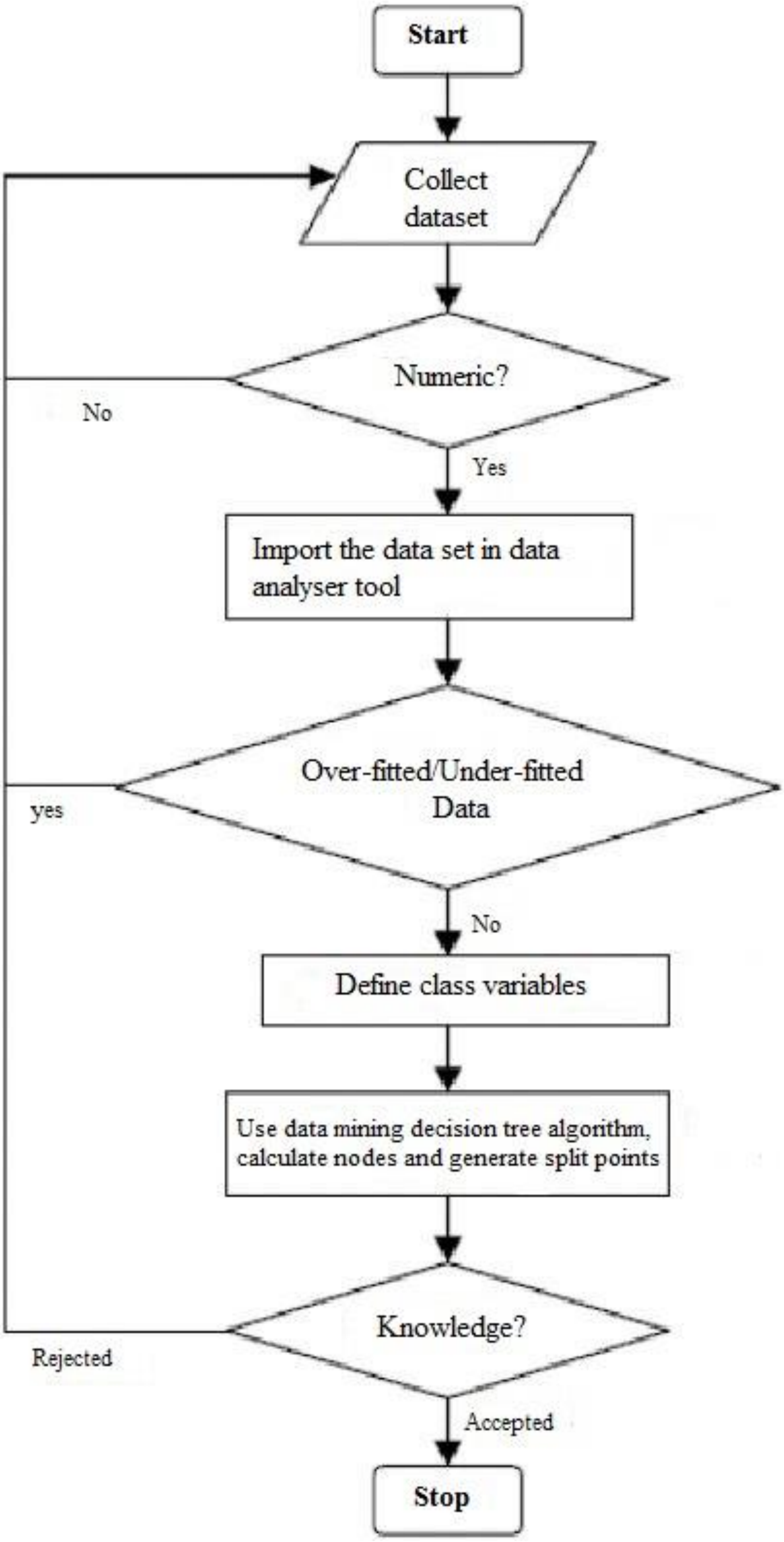

Start
Collect dataset
Numeric?
No
Yes
Import the data set in data analyser tool
Over-fitted/Under-fitted Data
yes
No
Define class variables
Use data mining decision tree algorithm, calculate nodes and generate split points
Knowledge?
Rejected
Accepted
Stop

Figure 3.1 Flow chart of research methodology

The basic strategy defined in this flow chart is as follows:

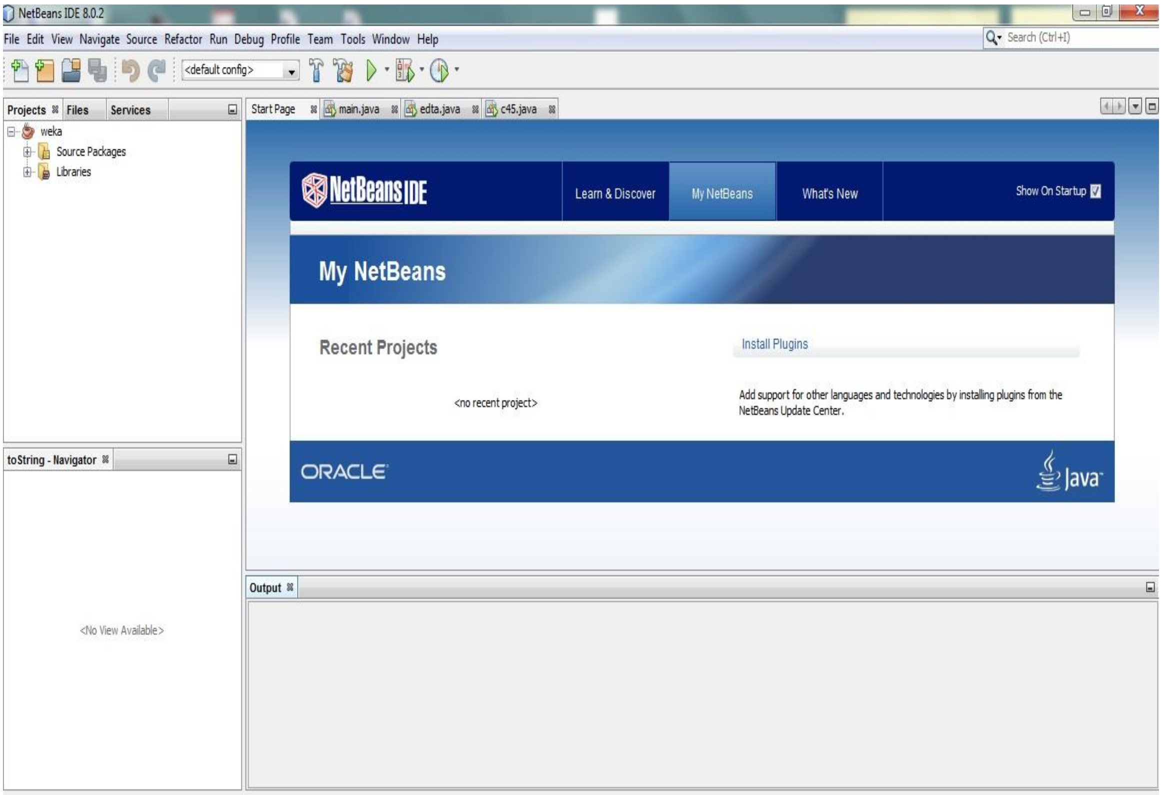
- Start with the collection of data set, which is collected by the software named as WirelessMon.
- The data set will include the attributes as strength, Authentication Type, Threshold and Frequency etc. of the mobile network.
- Check whether the data is numeric or not as the splitting of data in nodes will correspond to this only. If the data is not numeric then go back and collect the data set again but if the data is numeric then move further.
- Import dataset to the data analyser tool.
- Check whether the data is over-fitted/under-fitted. If the data is over-fitted/under-fitted then go back to the collect of dataset else move further.
- Define class variable. Then generate the tree based on that class variable by using a tree generation algorithm by specifying the number of nodes and children.
- Calculate the computation time, error rate and correctly classified instances of the C4.5 tree making algorithm.
- Now use the same dataset with EDTA to create a decision tree for prediction and calculate the computation time, error rate, correctly and incorrectly classified instances.
- Finally, computing both the results of the existing algorithm and the proposed technique (EDTA).

The proposed technology will use an enhanced algorithm of decision trees which is an algorithm of classification approach and will perform a predictive analysis to analyse the services at various locations. EDTA constructs many number of decision trees. This algorithm works for both categorical feature such as gender and continuous feature such as income. In decision path the categorical feature is chosen for once whereas the continuous feature can be chosen for more than once. As soon as the node is empty that is there is no attribute left to split then the tree constructions stops over there and even if the tree depth is exceeded by some defined limit the tree stops further construction.

The platform used for development of the algorithm is used as NetBeans. The dataset collected for this research is done by the software known as WirelessMon and the output of the research is taken in the tool known as WEKA. This research will compare the results with other algorithm.

Below is the brief description of the tools used in this research:-

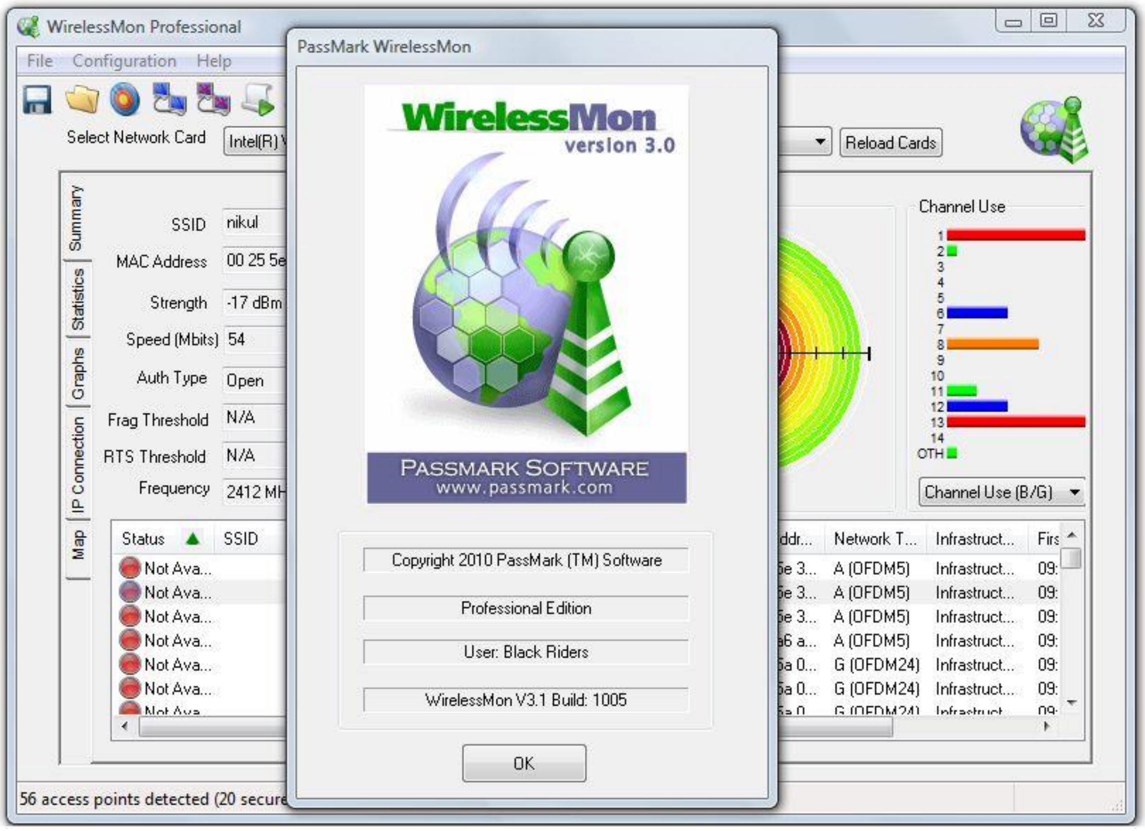
- NetBeans – NetBeans is a platform used for software development. There are some modular software components set, these components are known as modules. This software allows the development of the applications from modules. Third party developers can extend the NetBeans based applications including the applications of NetBeans IDE. It is basically used for java applications development but along with this other languages like PHP, HTML and C/C++ are also supported. It runs on maximum all the operating system platforms which have a compatible java virtual machine. The NetBeans version used in this research is NetBeans IDE 8.0.2.



Add support for other languages and technologies by installing plugins from the NetBeans Update Center.

Fig 3.2 NetBeans

- WirelessMon – WirelessMon is used for collecting the dataset for this research. It is a software tool developed by Pass Mark software. It basically helps in locating the interference sources of the network. It creates maps of signal strength of a particular area and wireless antenna is located which is used to check the coverage and range of Wi-Fi and signal levels of the connect Wi-Fi network and other networks which are in range. WirelessMon is also used for measuring speed, throughput and available data rates of the network. It runs on all windows operating system both for 32 bits and 64 bits and needs compatible 802.11wireless adaptor. The WirelessMon version used in this research is WirelessMon 3.0.



SSID nikul
MAC Address 00 25 5e
Strength -17 dBm
Speed (Mbits) 54
Auth Type Open
Frag Threshold N/A
RTS Threshold N/A
Frequency 2412 MH

Status	SSID
Not Ava...	
Not Ava...	
Not Ava...	
Not Ava...	
Not Ava...	
Not Ava...	
Not Ava...	

ddr...	Network T...	Infrastruct...	Firs
e 3...	A (OFDM5)	Infrastruct...	09:
e 3...	A (OFDM5)	Infrastruct...	09:
e 3...	A (OFDM5)	Infrastruct...	09:
6 a...	A (OFDM5)	Infrastruct...	09:
a 0...	G (OFDM24)	Infrastruct...	09:
a 0...	G (OFDM24)	Infrastruct...	09:
a 0	G (OFDM24)	Infrastruct	09:

Copyright 2010 PassMark (TM) Software

Professional Edition

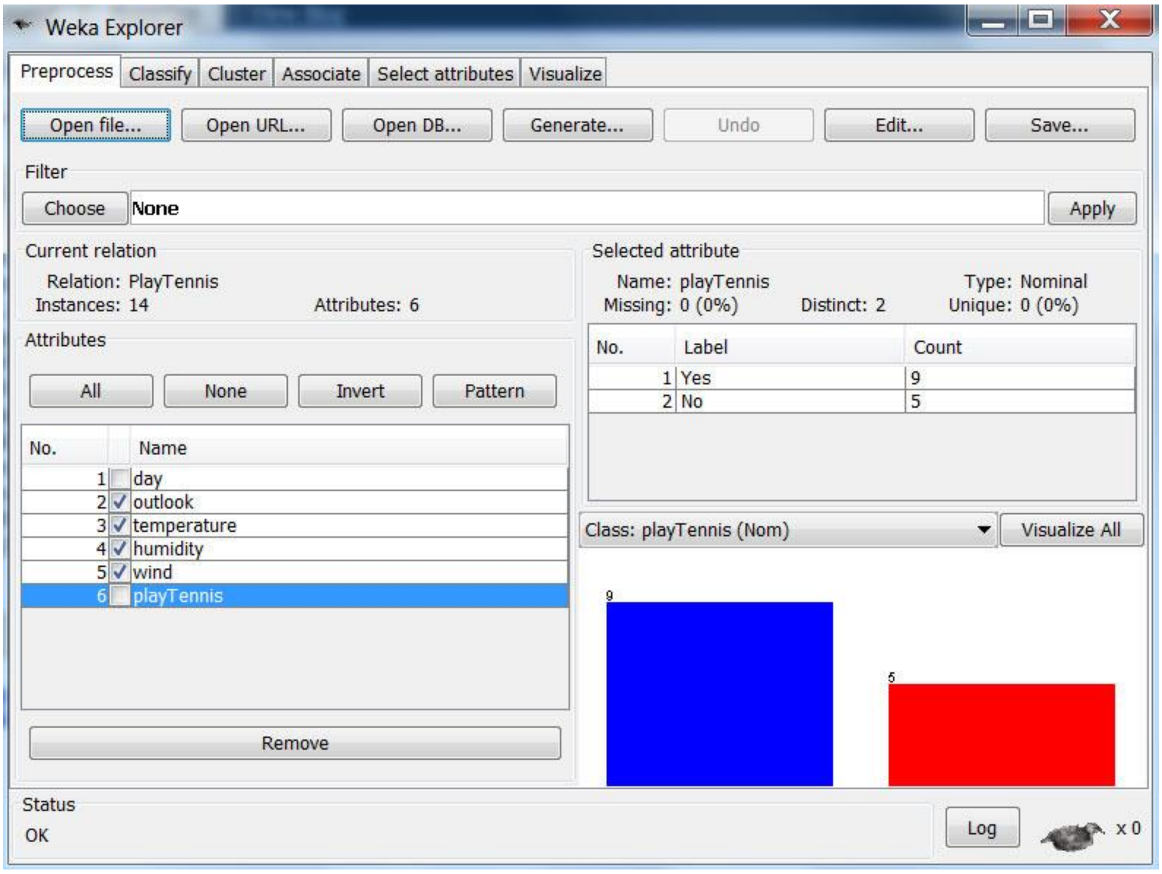
User: Black Riders

WirelessMon V3.1 Build: 1005

56 access points detected (20 secure

Fig 3.3 WirelessMon

- WEKA - WEKA is a data analysis and predictive modelling tool which includes clustering, classification, regression, pre-processing, and association rules. Java database connectivity is used by WEKA to provide SQL connectivity and the result is easily processed which is returned by query. WEKA have different panels, to use classification and regression algorithms classify panel is used, for association rule learners associate panel is used and for clustering techniques cluster panel is used. Due to its graphic user interface it is easy to use. The WEKA version used in this research is WEKA 3.4.



Filter: None

Current relation
Relation: PlayTennis
Instances: 14
Attributes: 6

Selected attribute
Name: playTennis
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	Yes	9
2	No	5

Attributes

No.		Name
1	□	day
2	☑	outlook
3	☑	temperature
4	☑	humidity
5	☑	wind
6	□	playTennis

Class: playTennis (Nom)

Status
OK

Fig 3.4 WEKA

C4.5 Algorithm

The pseudo code for the C4.5 is given below along with the input and output requirements

Input:

- Set s of variables either continuous or discrete attribute
- Each variables belongs to a class

Output:

- Decision tree

Steps:

- Base case is checked
- Highest information gain attribute is found
- Partition the values of set s into s1, s2… according to the highest information gain attribute value
- Steps are repeated for s1, s2…
- End

C4.5 is a pre-existing decision tree algorithm developed by Ross Quinlan. C4.5 is basically the ID3 algorithm's extension. It creates tree of any depth. The decision tree made by c4.5 is the result of recursively data partitioning of data present in the dataset. C4.5 uses depth first strategy for its decision. This algorithm takes into account all the tests which are feasible for splitting the dataset and out of them it selects the one that gives most suitable information gain. C4.5 considers both discrete and continuous attributes. One test with all the number of outcomes of distinct values is examined for discrete attributes and binary test is considered with distinct value of attributes for continuous attribute. The training data which belongs to the node is being sorted for continuous attribute for gathering the entropy gain efficiently for the binary tests. C4.5 performs post-pruning that is after the tree creation it goes back again through the tree for removing extra branches which are not useful by replacing them with leaf node.

Enhanced Decision Tree Algorithm (EDTA)

The pseudo code for EDTA is given below along with the input and output requirements

Input:

- Set s of variables either continuous or discrete attribute
- Each variables belongs to a class

Output:

- Decision tree

Steps:

- Training instances proportions at each branch
- Tree constructing class which considers k chosen attributes at each node.
- Calculate in the leaf the minimum instances weight
- Set instances of minimum number for each leaf
- End

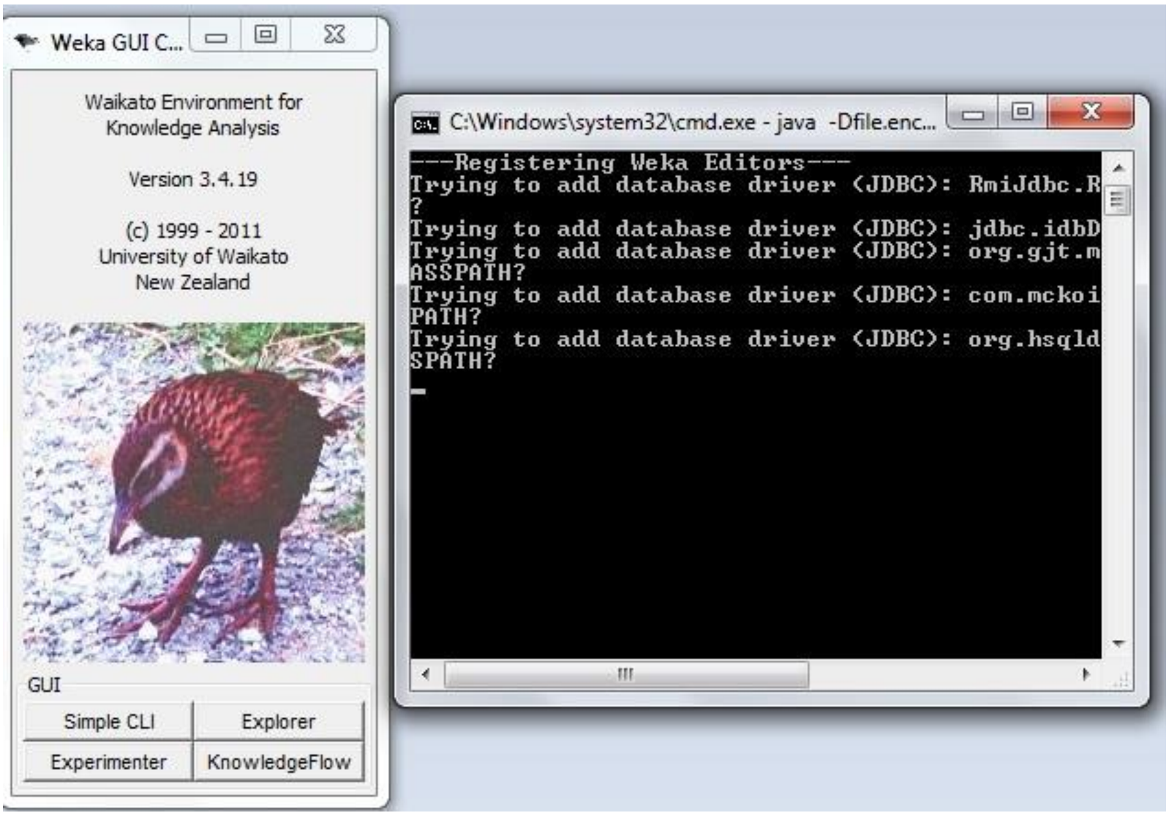
EDTA constructs many number of decision trees. This algorithm works for both categorical feature such as gender and continuous feature such as income. In decision path the categorical feature is chosen for once whereas the continuous feature can be chosen for more than once. As soon as the node is empty that is there is no attribute left to split then the tree constructions stops over there and even if the tree depth is exceeded by some defined limit the tree stops further construction. No pruning is performed in this decision tree algorithm but still the unwanted nodes are removed. Expansion of node is not that necessary, if any of its children do not have different class distribution from node, then the node expansion is removed and made it as leaf node. While building the tree recursively all the necessity checking is done when return of recursion is there.

.

Chapter 4

RESULTS AND DISCUSSIONS

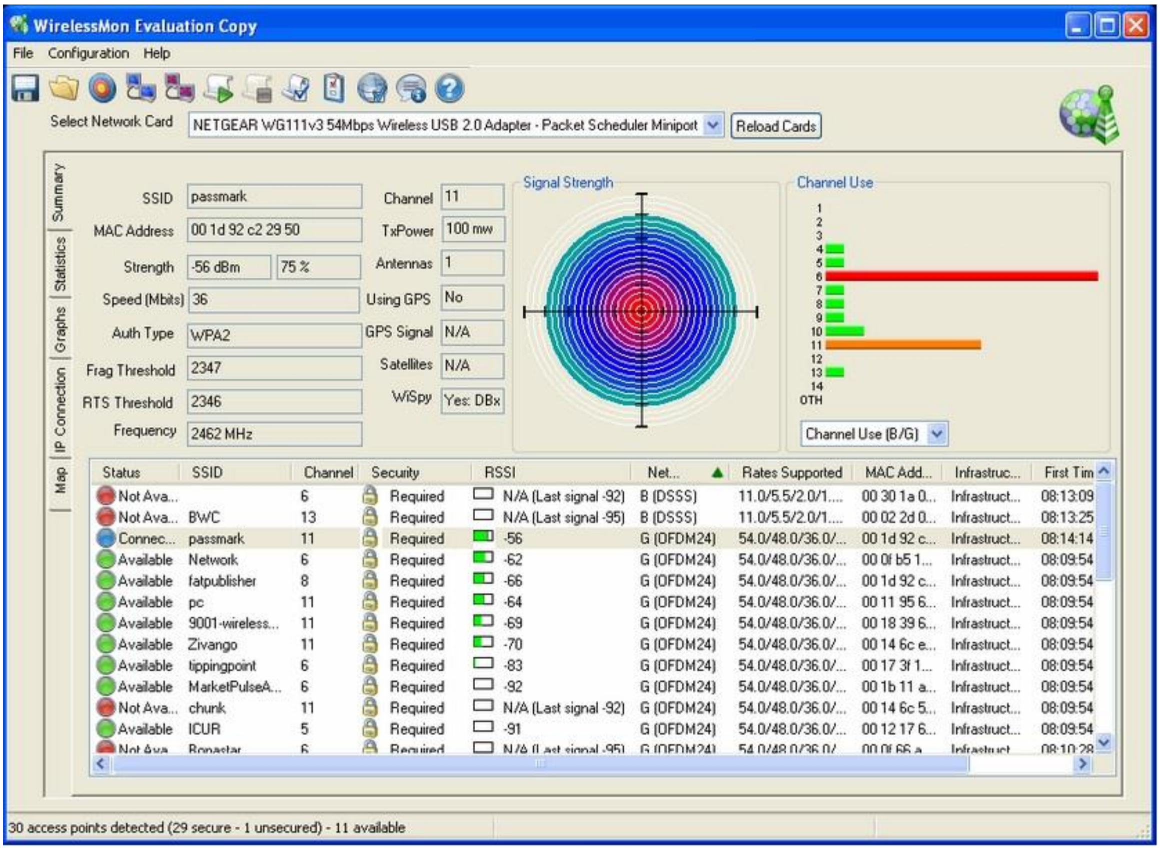
System evaluation is as important as development of the system and the results of this research is based on the evaluation of enhanced decision tree algorithm (EDTA) and its comparison with the pre-existing C4.5 algorithm. The platform used to implement and evaluate the algorithm is WEKA and the dataset used is generated by WirelessMon software. The attribute values displayed on the screen are written manually to an excel sheet. The code of this algorithm is written in java using NetBeans as the java platform.



```
---Registering Weka Editors---
Trying to add database driver (JDBC): RmiJdbc.R
?
Trying to add database driver (JDBC): jdbc.idbD
Trying to add database driver (JDBC): org.gjt.m
ASSPATH?
Trying to add database driver (JDBC): com.mckoi
PATH?
Trying to add database driver (JDBC): org.hsqld
SPATH?
_
```

Fig 4.1 WEKA interface

WirelessMon is used to collect the data set. Below is the interface of WirelessMon software and displayed values are the collected data services over a particular location. WirelessMon is also used for measuring speed, throughput and available data rates of the network. It runs on all windows operating system both for 32 bits and 64 bits and needs compatible 802.11wireless adaptor. The WirelessMon version used in this research is WirelessMon 2.1.



Select Network Card: NETGEAR WG111v3 54Mbps Wireless USB 2.0 Adapter - Packet Scheduler Miniport

SSID	passmark	Channel	11
MAC Address	00 1d 92 c2 29 50	TxPower	100 mw
Strength	-56 dBm 75 %	Antennas	1
Speed (Mbits)	36	Using GPS	No
Auth Type	WPA2	GPS Signal	N/A
Frag Threshold	2347	Satellites	N/A
RTS Threshold	2346	WiSpy	Yes: DBx
Frequency	2462 MHz		

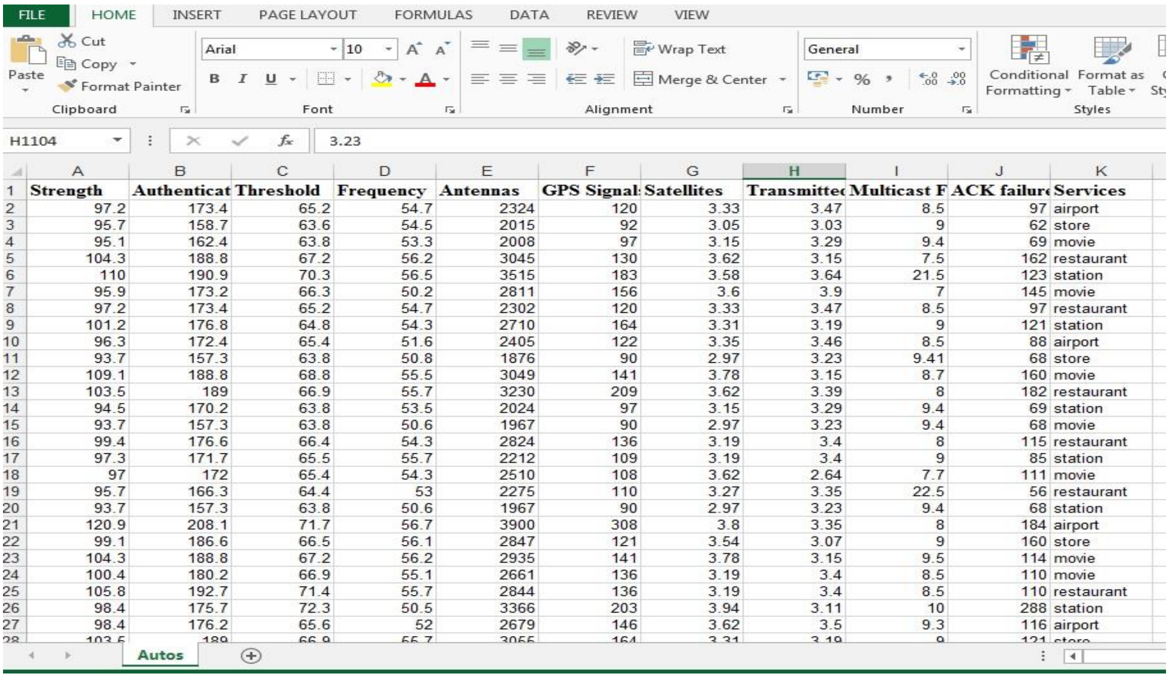
Status	SSID	Channel	Security	RSSI	Net... ▲	Rates Supported	MAC Add...	Infrastruc...	First Tim
Not Ava...		6	Required	N/A (Last signal -92)	B (DSSS)	11.0/5.5/2.0/1....	00 30 1a 0...	Infrastruct...	08:13:09
Not Ava...	BWC	13	Required	N/A (Last signal -95)	B (DSSS)	11.0/5.5/2.0/1....	00 02 2d 0...	Infrastruct...	08:13:25
Connec...	passmark	11	Required	-56	G (OFDM24)	54.0/48.0/36.0/...	00 1d 92 c...	Infrastruct...	08:14:14
Available	Network	6	Required	-62	G (OFDM24)	54.0/48.0/36.0/...	00 0f b5 1...	Infrastruct...	08:09:54
Available	fatpublisher	8	Required	-66	G (OFDM24)	54.0/48.0/36.0/...	00 1d 92 c...	Infrastruct...	08:09:54
Available	pc	11	Required	-64	G (OFDM24)	54.0/48.0/36.0/...	00 11 95 6...	Infrastruct...	08:09:54
Available	9001-wireless...	11	Required	-69	G (OFDM24)	54.0/48.0/36.0/...	00 18 39 6...	Infrastruct...	08:09:54
Available	Zivango	11	Required	-70	G (OFDM24)	54.0/48.0/36.0/...	00 14 6c e...	Infrastruct...	08:09:54
Available	tippingpoint	6	Required	-83	G (OFDM24)	54.0/48.0/36.0/...	00 17 3f 1...	Infrastruct...	08:09:54
Available	MarketPulseA...	6	Required	-92	G (OFDM24)	54.0/48.0/36.0/...	00 1b 11 a...	Infrastruct...	08:09:54
Not Ava...	chunk	11	Required	N/A (Last signal -92)	G (OFDM24)	54.0/48.0/36.0/...	00 14 6c 5...	Infrastruct...	08:09:54
Available	ICUR	5	Required	-91	G (OFDM24)	54.0/48.0/36.0/...	00 12 17 6...	Infrastruct...	08:09:54
Not Ava...	Ronastar	6	Required	N/A (Last signal -95)	G (OFDM24)	54.0/48.0/36.0/...	00 0f 66 a...	Infrastruct...	08:10:28

30 access points detected (29 secure - 1 unsecured) - 11 available

Fig 4.2 WirelessMon interface

There are in total 11 attributes in this dataset namely Strength, Authentication, Threshold, Frequency, antennas, GPS signals, satellites, Transmitted frame, Multicast frame, ACK failure count, services.

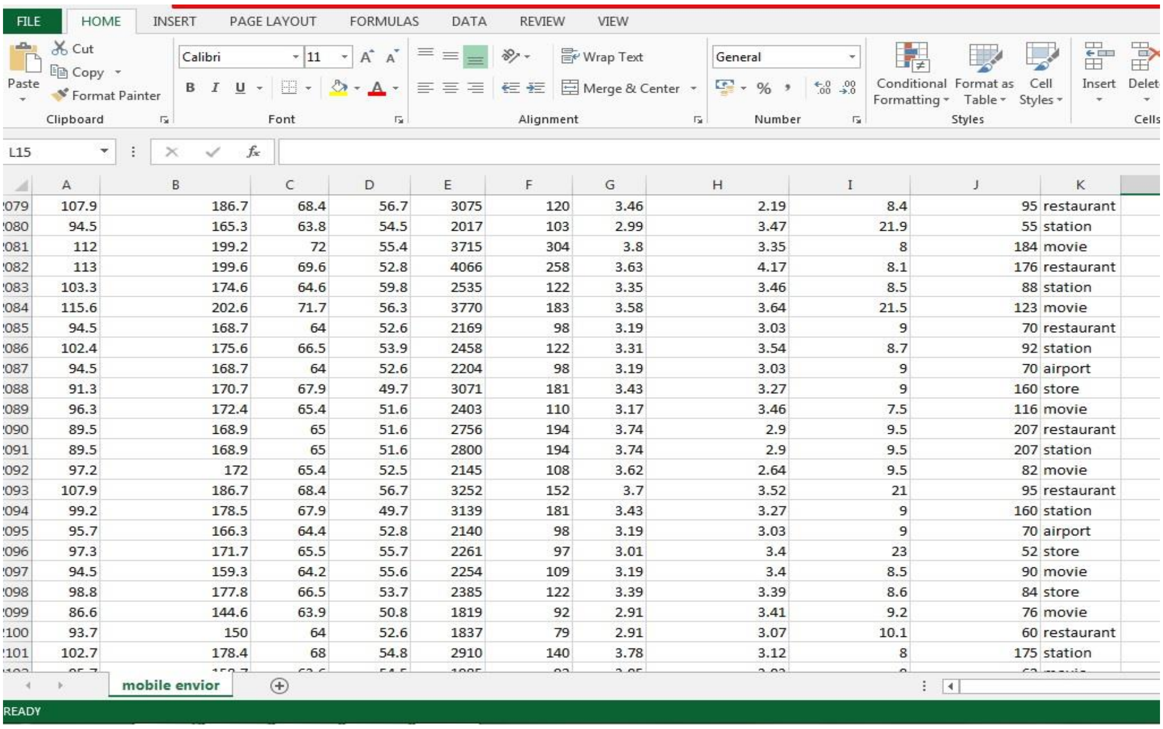
Dataset written to excel sheet and the format used is .csv as WEKA reads this file format



H1104 | 3.23

	A	B	C	D	E	F	G	H	I	J	K
1	Strength	Authenticat	Threshold	Frequency	Antennas	GPS Signal:	Satellites	Transmittec	Multicast F	ACK failure	Services
2	97.2	173.4	65.2	54.7	2324	120	3.33	3.47	8.5	97	airport
3	95.7	158.7	63.6	54.5	2015	92	3.05	3.03	9	62	store
4	95.1	162.4	63.8	53.3	2008	97	3.15	3.29	9.4	69	movie
5	104.3	188.8	67.2	56.2	3045	130	3.62	3.15	7.5	162	restaurant
6	110	190.9	70.3	56.5	3515	183	3.58	3.64	21.5	123	station
7	95.9	173.2	66.3	50.2	2811	156	3.6	3.9	7	145	movie
8	97.2	173.4	65.2	54.7	2302	120	3.33	3.47	8.5	97	restaurant
9	101.2	176.8	64.8	54.3	2710	164	3.31	3.19	9	121	station
10	96.3	172.4	65.4	51.6	2405	122	3.35	3.46	8.5	88	airport
11	93.7	157.3	63.8	50.8	1876	90	2.97	3.23	9.41	68	store
12	109.1	188.8	68.8	55.5	3049	141	3.78	3.15	8.7	160	movie
13	103.5	189	66.9	55.7	3230	209	3.62	3.39	8	182	restaurant
14	94.5	170.2	63.8	53.5	2024	97	3.15	3.29	9.4	69	station
15	93.7	157.3	63.8	50.6	1967	90	2.97	3.23	9.4	68	movie
16	99.4	176.6	66.4	54.3	2824	136	3.19	3.4	8	115	restaurant
17	97.3	171.7	65.5	55.7	2212	109	3.19	3.4	9	85	station
18	97	172	65.4	54.3	2510	108	3.62	2.64	7.7	111	movie
19	95.7	166.3	64.4	53	2275	110	3.27	3.35	22.5	56	restaurant
20	93.7	157.3	63.8	50.6	1967	90	2.97	3.23	9.4	68	station
21	120.9	208.1	71.7	56.7	3900	308	3.8	3.35	8	184	airport
22	99.1	186.6	66.5	56.1	2847	121	3.54	3.07	9	160	store
23	104.3	188.8	67.2	56.2	2935	141	3.78	3.15	9.5	114	movie
24	100.4	180.2	66.9	55.1	2661	136	3.19	3.4	8.5	110	movie
25	105.8	192.7	71.4	55.7	2844	136	3.19	3.4	8.5	110	restaurant
26	98.4	175.7	72.3	50.5	3366	203	3.94	3.11	10	288	station
27	98.4	176.2	65.6	52	2679	146	3.62	3.5	9.3	116	airport

Fig 4.3 Dataset (1)

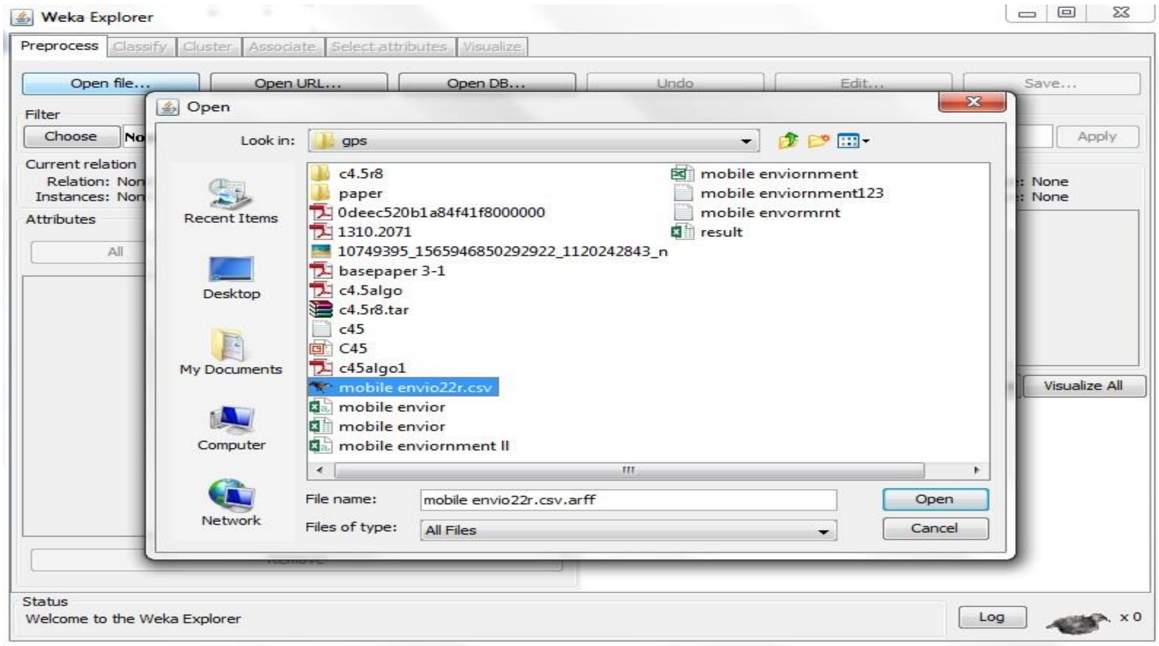


L15

	A	B	C	D	E	F	G	H	I	J	K
079	107.9	186.7	68.4	56.7	3075	120	3.46	2.19	8.4	95	restaurant
080	94.5	165.3	63.8	54.5	2017	103	2.99	3.47	21.9	55	station
081	112	199.2	72	55.4	3715	304	3.8	3.35	8	184	movie
082	113	199.6	69.6	52.8	4066	258	3.63	4.17	8.1	176	restaurant
083	103.3	174.6	64.6	59.8	2535	122	3.35	3.46	8.5	88	station
084	115.6	202.6	71.7	56.3	3770	183	3.58	3.64	21.5	123	movie
085	94.5	168.7	64	52.6	2169	98	3.19	3.03	9	70	restaurant
086	102.4	175.6	66.5	53.9	2458	122	3.31	3.54	8.7	92	station
087	94.5	168.7	64	52.6	2204	98	3.19	3.03	9	70	airport
088	91.3	170.7	67.9	49.7	3071	181	3.43	3.27	9	160	store
089	96.3	172.4	65.4	51.6	2403	110	3.17	3.46	7.5	116	movie
090	89.5	168.9	65	51.6	2756	194	3.74	2.9	9.5	207	restaurant
091	89.5	168.9	65	51.6	2800	194	3.74	2.9	9.5	207	station
092	97.2	172	65.4	52.5	2145	108	3.62	2.64	9.5	82	movie
093	107.9	186.7	68.4	56.7	3252	152	3.7	3.52	21	95	restaurant
094	99.2	178.5	67.9	49.7	3139	181	3.43	3.27	9	160	station
095	95.7	166.3	64.4	52.8	2140	98	3.19	3.03	9	70	airport
096	97.3	171.7	65.5	55.7	2261	97	3.01	3.4	23	52	store
097	94.5	159.3	64.2	55.6	2254	109	3.19	3.4	8.5	90	movie
098	98.8	177.8	66.5	53.7	2385	122	3.39	3.39	8.6	84	store
099	86.6	144.6	63.9	50.8	1819	92	2.91	3.41	9.2	76	movie
100	93.7	150	64	52.6	1837	79	2.91	3.07	10.1	60	restaurant
101	102.7	178.4	68	54.8	2910	140	3.78	3.12	8	175	station

Fig 4.4 Dataset (2)

The csv file of the dataset in opened in WEKA and by that all the attributes and there values are extracted for performing the analysis.



Relation: None
Instances: None

Look in: gps

c4.5r8
paper
0deec520b1a84f41f8000000
1310.2071
10749395_1565946850292922_1120242843_n
basepaper 3-1
c4.5algo
c4.5r8.tar
c45
C45
c45algo1
mobile envio22r.csv
mobile envior
mobile envior
mobile enviornment II
mobile enviornment
mobile enviornment123
mobile envormrnt
result

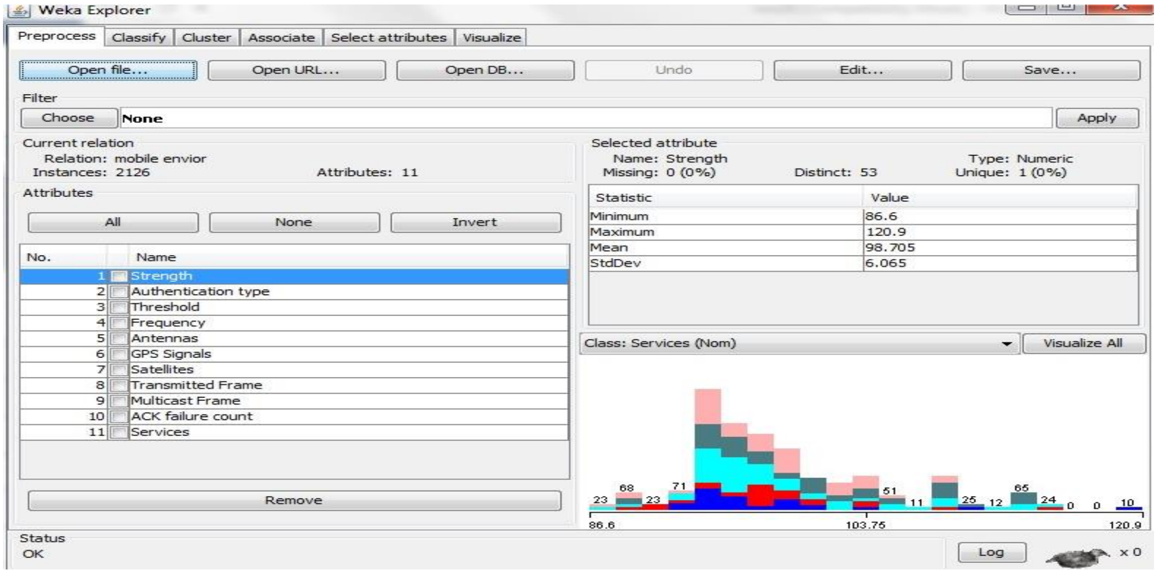
File name: mobile envio22r.csv.arff

Files of type: All Files

Status
Welcome to the Weka Explorer

Fig 4.5 File open in WEKA

The dataset mobile envio22r.csv is opened in WEKA and after opening it will automatically calculate the number of attributes and number of instances present in the dataset and this is known as pre-process of the dataset done by WEKA before analysis.



Filter
Choose None

Current relation
Relation: mobile envior
Instances: 2126
Attributes: 11

Attributes

No.	Name
1	Strength
2	Authentication type
3	Threshold
4	Frequency
5	Antennas
6	GPS Signals
7	Satellites
8	Transmitted Frame
9	Multicast Frame
10	ACK failure count
11	Services

Selected attribute
Name: Strength
Type: Numeric
Missing: 0 (0%)
Distinct: 53
Unique: 1 (0%)

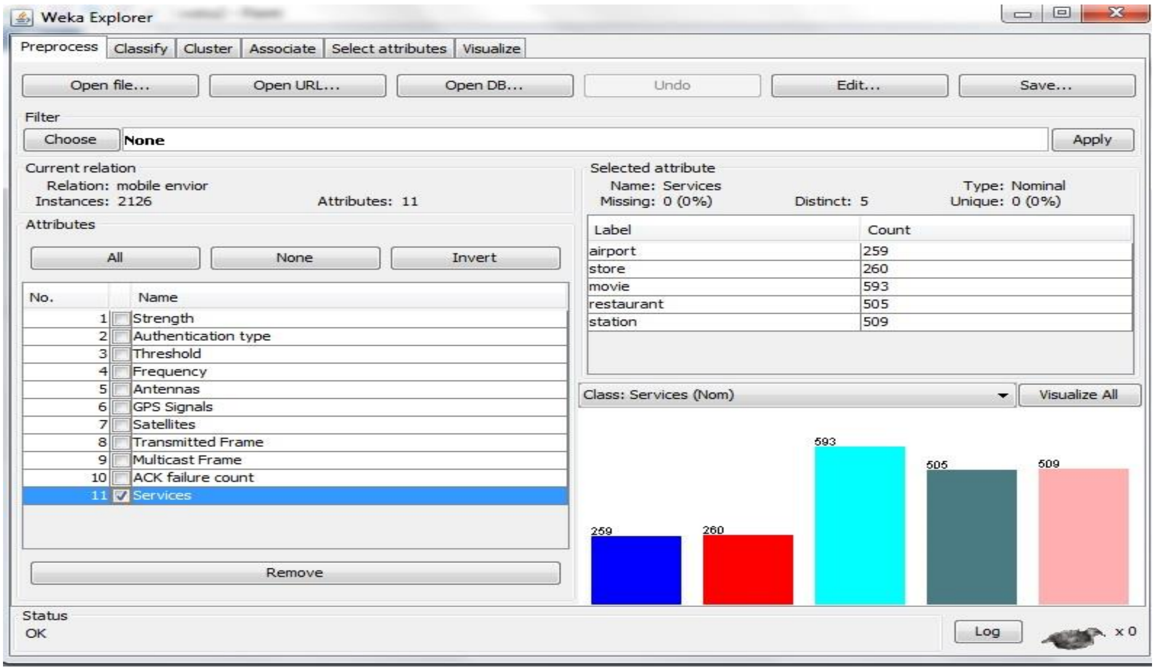
Statistic	Value
Minimum	86.6
Maximum	120.9
Mean	98.705
StdDev	6.065

Class: Services (Nom)

Status
OK

Fig 4.6 Pre-process in WEKA

The class attribute is the attribute for which the prediction is to be done. For this research the class attribute is selected as services.



Filter: None

Current relation
Relation: mobile envior
Instances: 2126 Attributes: 11

Attributes

No.		Name
1	□	Strength
2	□	Authentication type
3	□	Threshold
4	□	Frequency
5	□	Antennas
6	□	GPS Signals
7	□	Satellites
8	□	Transmitted Frame
9	□	Multicast Frame
10	□	ACK failure count
11	☑	Services

Selected attribute
Name: Services Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

Label	Count
airport	259
store	260
movie	593
restaurant	505
station	509

Class: Services (Nom)

Status
OK

Fig 4.7 Attribute service as class attribute

WEKA calculates the number of instances each value parameter of the class attribute contains. The class attribute services has five parameters namely airport, store, movie, restaurant and station. These are the places where network availability is searched and prediction that which place the services are used more.

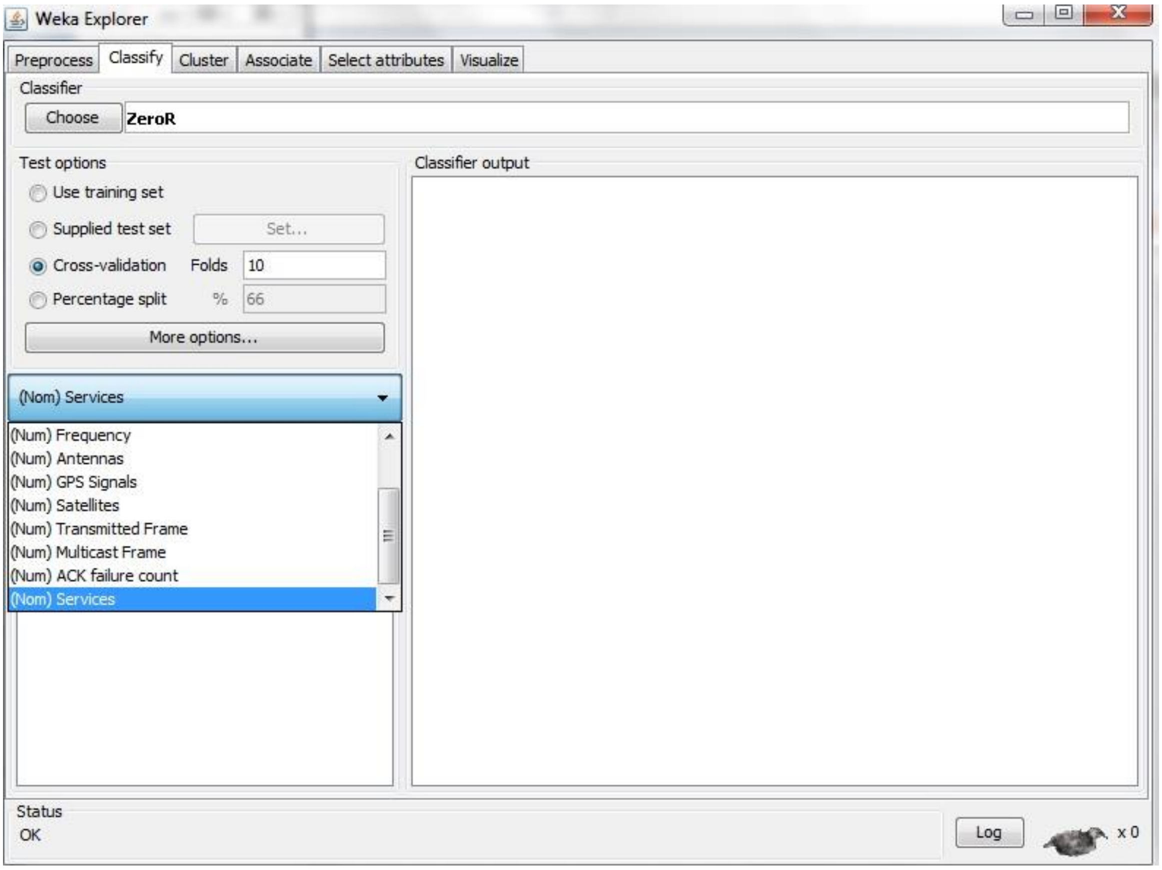
The number of instances corresponding to each service are

- Airport :- 259
- Store :- 260
- Movie :- 593
- Restaurant :- 505
- Station :- 509

Hence the total number of instances are 2126.

Along with classification WEKA is also used for clustering and association analysis. So for moving further selection is needed whether to classify, make cluster or association analysis is to be done. As this research is mainly focusing on classification so classify is to be selected.

After importing the dataset and selecting the classify option we need to check for the class attribute. All the eleven attributes are shown and as already defined the class attribute is same as services but if in case the class attribute needs to be changed it can be done by selecting from the list of attributes which are present in the dataset.



Classifier
ZeroR

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

(Num) Frequency
(Num) Antennas
(Num) GPS Signals
(Num) Satellites
(Num) Transmitted Frame
(Num) Multicast Frame
(Num) ACK failure count
(Nom) Services

Status
OK

Fig 4.8 Class attribute selected as services

There are number of algorithm under classification that is under decision trees. WEKA already have these algorithms installed in it. Along with this WEKA also gives the liberty to add more algorithms to its package. All the algorithms are written in java language. Every algorithm gives different prediction results on same dataset. In this research the comparison is done between new and already existing algorithm.

To select an algorithm an option "choose" is given. Under which all the algorithms are there. The required algorithm is selected from there.

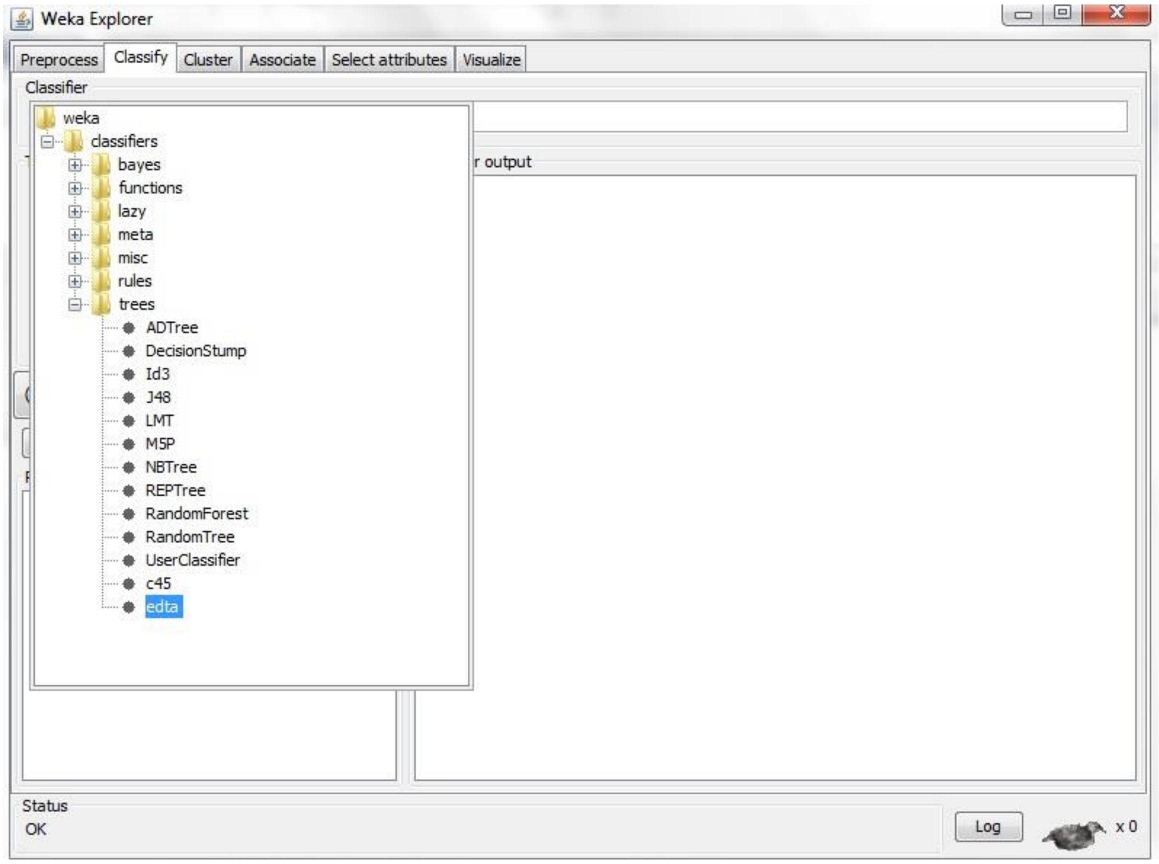
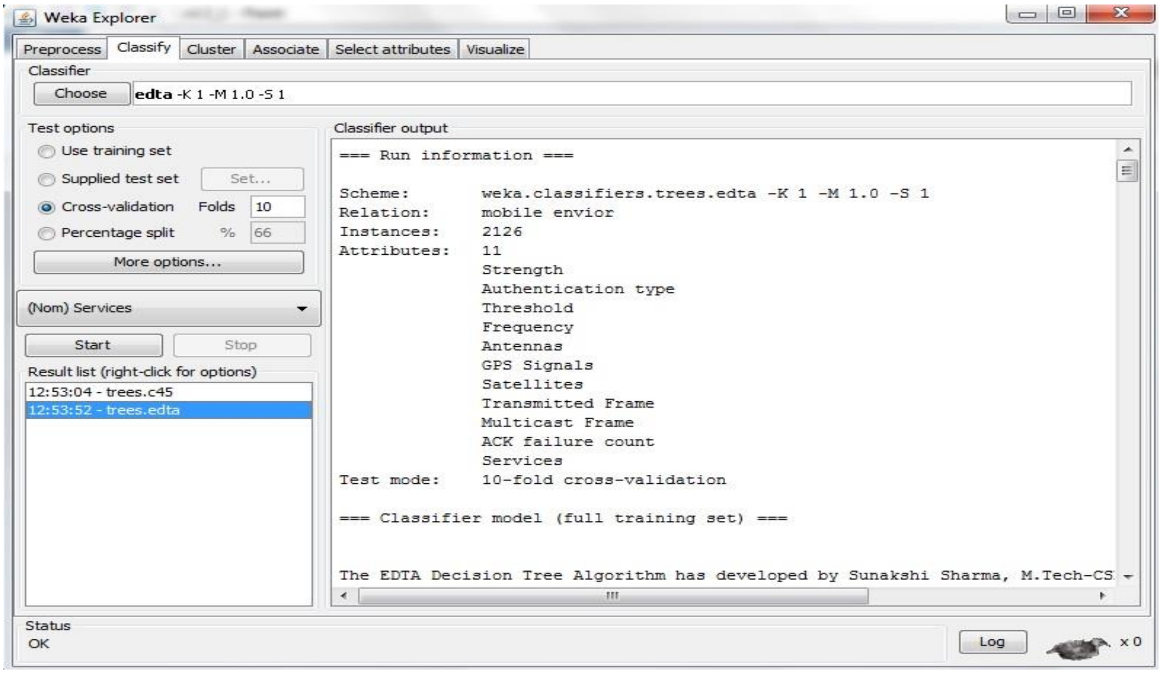


Fig 4.9 Selection of an algorithm

The proposed algorithm EDTA is selected to predict for the given dataset.



edta -K 1 -M 1.0 -S 1

Folds 10

% 66

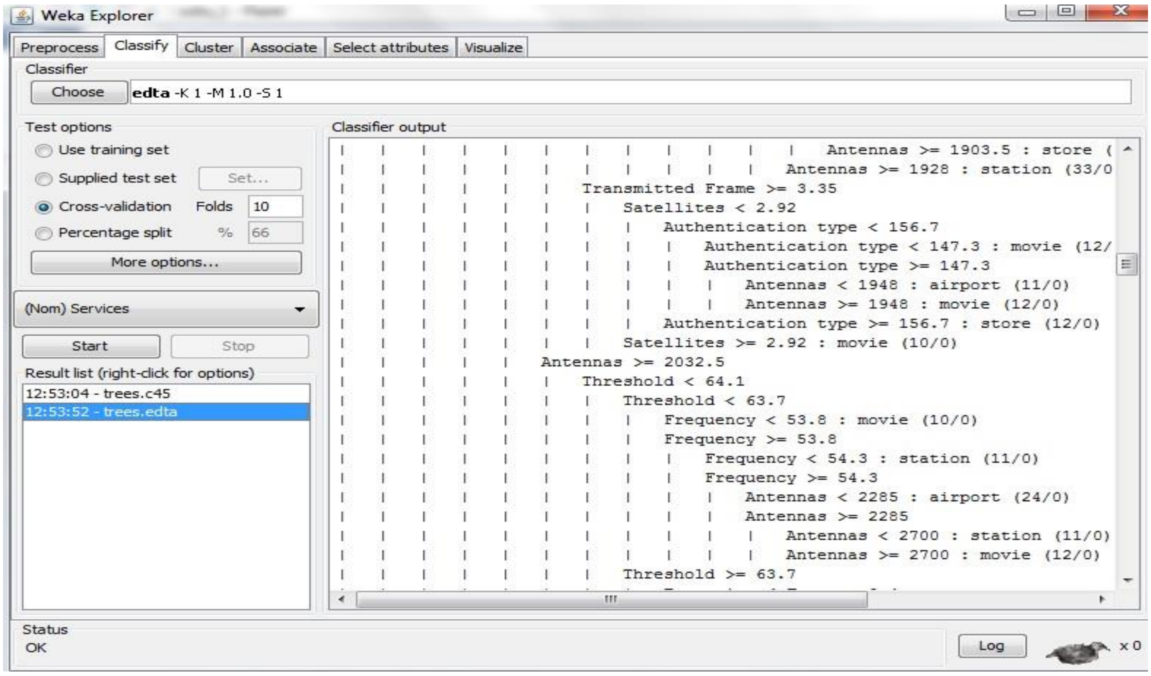
12:53:04 - trees.c45
12:53:52 - trees.edta

```
=== Run information ===

Scheme:       weka.classifiers.trees.edta -K 1 -M 1.0 -S 1
Relation:     mobile envior
Instances:    2126
Attributes:   11
              Strength
              Authentication type
              Threshold
              Frequency
              Antennas
              GPS Signals
              Satellites
              Transmitted Frame
              Multicast Frame
              ACK failure count
              Services
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


The EDTA Decision Tree Algorithm has developed by Sunakshi Sharma, M.Tech-CS
```

Status OK

Fig 4.10 Output of EDTA algorithm

EDTA shows the tree which is built for the analysis. There are only few algorithms which shows tree in the output window.



edta -K 1 -M 1.0 -S 1

Folds 10

% 66

12:53:04 - trees.c45
12:53:52 - trees.edta

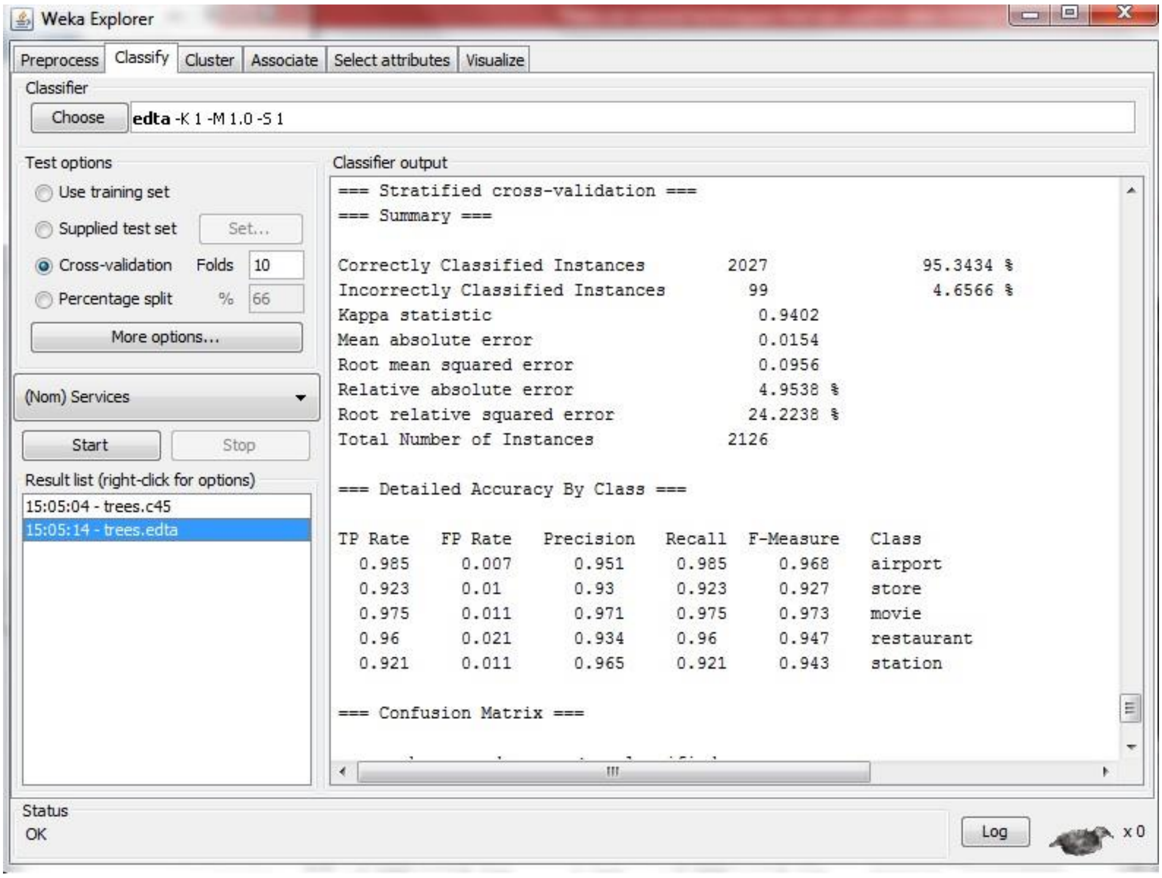
```
|   |   |   |   |   |   |   |   |   |   |   |   Antennas >= 1903.5 : store (
|   |   |   |   |   |   |   |   |   |   |   Antennas >= 1928 : station (33/0
|   |   |   |   |   |   Transmitted Frame >= 3.35
|   |   |   |   |   |   |   Satellites < 2.92
|   |   |   |   |   |   |   |   Authentication type < 156.7
|   |   |   |   |   |   |   |   |   Authentication type < 147.3 : movie (12/
|   |   |   |   |   |   |   |   |   Authentication type >= 147.3
|   |   |   |   |   |   |   |   |   |   Antennas < 1948 : airport (11/0)
|   |   |   |   |   |   |   |   |   |   Antennas >= 1948 : movie (12/0)
|   |   |   |   |   |   |   |   Authentication type >= 156.7 : store (12/0)
|   |   |   |   |   |   |   Satellites >= 2.92 : movie (10/0)
|   |   |   |   |   Antennas >= 2032.5
|   |   |   |   |   |   Threshold < 64.1
|   |   |   |   |   |   |   Threshold < 63.7
|   |   |   |   |   |   |   |   Frequency < 53.8 : movie (10/0)
|   |   |   |   |   |   |   |   Frequency >= 53.8
|   |   |   |   |   |   |   |   |   Frequency < 54.3 : station (11/0)
|   |   |   |   |   |   |   |   |   Frequency >= 54.3
|   |   |   |   |   |   |   |   |   |   Antennas < 2285 : airport (24/0)
|   |   |   |   |   |   |   |   |   |   Antennas >= 2285
|   |   |   |   |   |   |   |   |   |   |   Antennas < 2700 : station (11/0)
|   |   |   |   |   |   |   |   |   |   |   Antennas >= 2700 : movie (12/0)
|   |   |   |   |   |   |   Threshold >= 63.7
```

Status OK

Fig 4.11 Decision tree by EDTA in output window

The result summary for the EDTA algorithm shows the total number of correctly classified instances and total number of incorrectly classified instances and the more number of correctly classified instances will give more accurate result. For EDTA the total number of correctly classified instances are 2027 and the incorrectly classified instances are 99.

Along with this the error rate is also shown, less the error rate means prediction or classification is accurate. Basically accuracy is one minus error rate. The mean squared error is 0.0154 and the root mean squared error is 0.956. Along with this relative absolute error was calculated as 4.9538% and the root relative squared error is 24.2238%.



edta -K 1 -M 1.0 -S 1

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

15:05:04 - trees.c45
15:05:14 - trees.edta

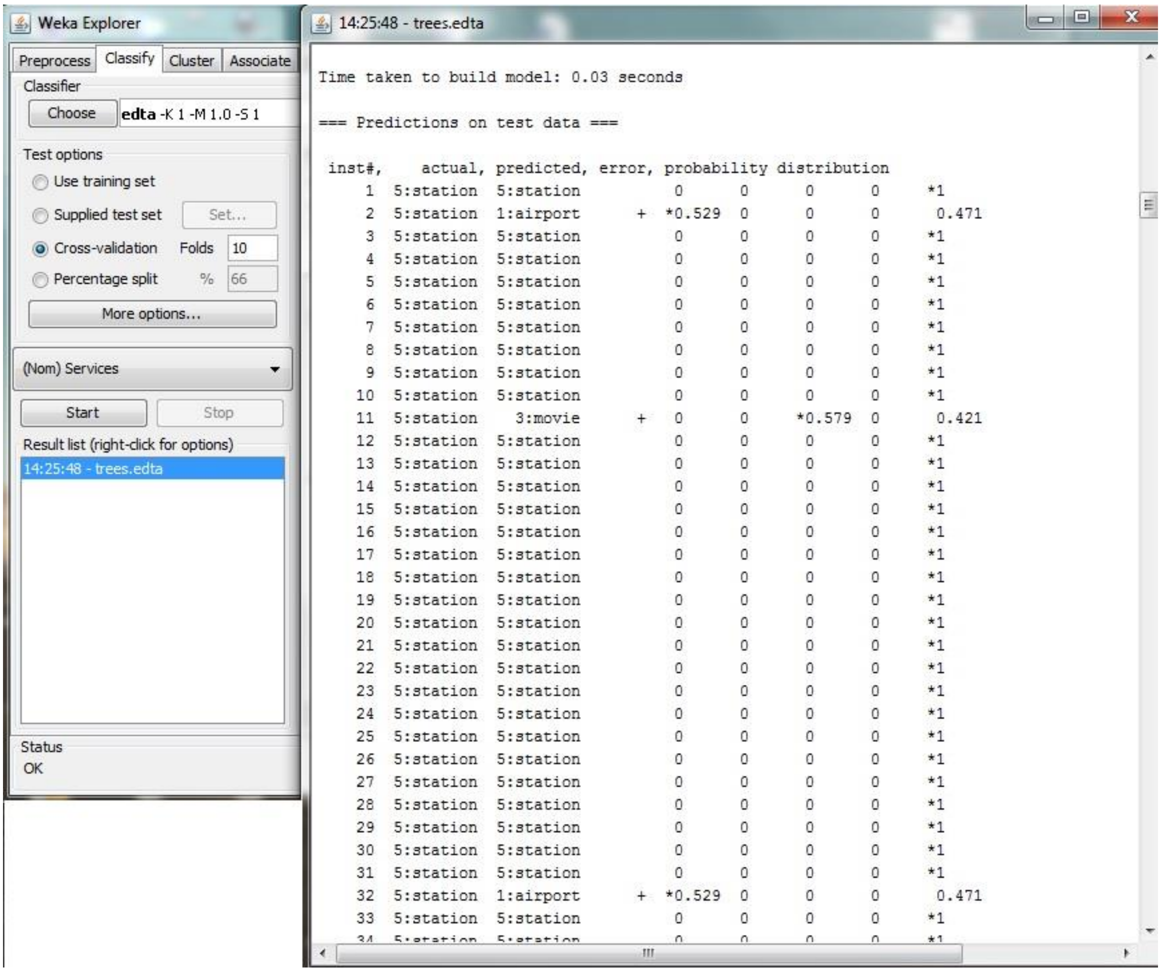
Status
OK

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2027               95.3434 %
Incorrectly Classified Instances        99                4.6566 %
Kappa statistic                          0.9402
Mean absolute error                      0.0154
Root mean squared error                  0.0956
Relative absolute error                  4.9538 %
Root relative squared error             24.2238 %
Total Number of Instances             2126

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.985     0.007      0.951     0.985     0.968    airport
  0.923     0.01       0.93      0.923     0.927    store
  0.975     0.011      0.971     0.975     0.973    movie
  0.96      0.021      0.934     0.96      0.947    restaurant
  0.921     0.011      0.965     0.921     0.943    station

=== Confusion Matrix ===
```

Fig 4.12 Result summary of EDTA

Prediction is done by the EDTA, and the prediction involves less errors. It includes instance number, actual value, predicted value, error, probability and distribution.

The instance which have error in its prediction has + marked under it. This also shows the time taken to build the model.



Classifier: edta -K 1 -M 1.0 -S 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

Result list (right-click for options)
14:25:48 - trees.edta

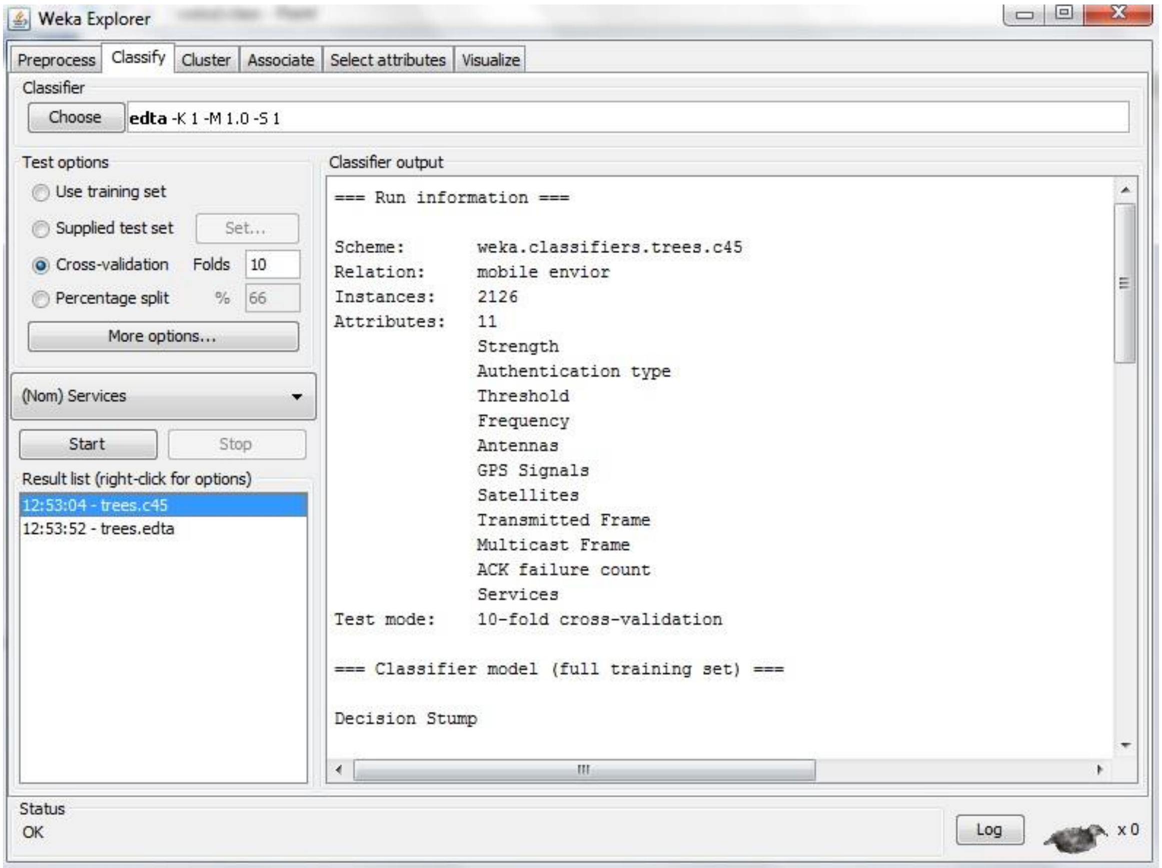
Status
OK

```
Time taken to build model: 0.03 seconds

=== Predictions on test data ===

 inst#,    actual, predicted, error, probability distribution
     1  5:station  5:station          0       0       0       0      *1
     2  5:station  1:airport      +  *0.529   0       0       0       0.471
     3  5:station  5:station          0       0       0       0      *1
     4  5:station  5:station          0       0       0       0      *1
     5  5:station  5:station          0       0       0       0      *1
     6  5:station  5:station          0       0       0       0      *1
     7  5:station  5:station          0       0       0       0      *1
     8  5:station  5:station          0       0       0       0      *1
     9  5:station  5:station          0       0       0       0      *1
    10  5:station  5:station          0       0       0       0      *1
    11  5:station    3:movie      +   0       0      *0.579   0       0.421
    12  5:station  5:station          0       0       0       0      *1
    13  5:station  5:station          0       0       0       0      *1
    14  5:station  5:station          0       0       0       0      *1
    15  5:station  5:station          0       0       0       0      *1
    16  5:station  5:station          0       0       0       0      *1
    17  5:station  5:station          0       0       0       0      *1
    18  5:station  5:station          0       0       0       0      *1
    19  5:station  5:station          0       0       0       0      *1
    20  5:station  5:station          0       0       0       0      *1
    21  5:station  5:station          0       0       0       0      *1
    22  5:station  5:station          0       0       0       0      *1
    23  5:station  5:station          0       0       0       0      *1
    24  5:station  5:station          0       0       0       0      *1
    25  5:station  5:station          0       0       0       0      *1
    26  5:station  5:station          0       0       0       0      *1
    27  5:station  5:station          0       0       0       0      *1
    28  5:station  5:station          0       0       0       0      *1
    29  5:station  5:station          0       0       0       0      *1
    30  5:station  5:station          0       0       0       0      *1
    31  5:station  5:station          0       0       0       0      *1
    32  5:station  1:airport      +  *0.529   0       0       0       0.471
    33  5:station  5:station          0       0       0       0      *1
    34  5:station  5:station          0       0       0       0      *1
```

Fig 4.13 EDTA Predictions

For comparison of EDTA with another algorithm one pre-existing algorithm is selected in the similar way as EDTA was selected. In this research for comparing the performance the algorithm C4.5 is used. Which is selected from the "choose" option.

Then its results are compared. This algorithm does not show the tree in its result window. This algorithm also shows the number of instances and number of attributes. The dataset used is also mentioned in the result window.



edta -K 1 -M 1.0 -S 1

Folds 10

% 66

(Nom) Services

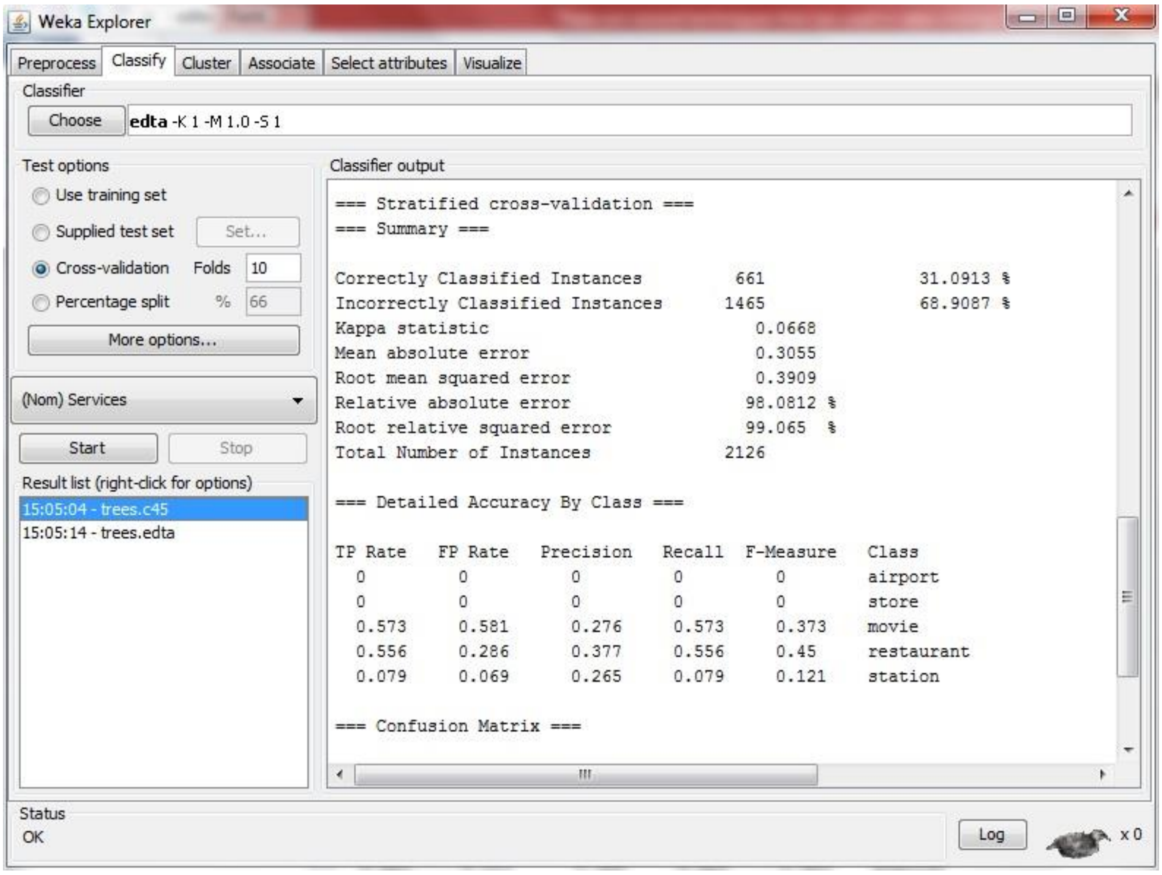
12:53:04 - trees.c45
12:53:52 - trees.edta

```
=== Run information ===

Scheme:       weka.classifiers.trees.c45
Relation:     mobile envior
Instances:    2126
Attributes:   11
              Strength
              Authentication type
              Threshold
              Frequency
              Antennas
              GPS Signals
              Satellites
              Transmitted Frame
              Multicast Frame
              ACK failure count
              Services
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Decision Stump
```

Status
OK

Fig 4.14 Output of C4.5 algorithm

The total number of correctly classified instances for C4.5 are 661 and the total number of incorrectly classified instances are 1465. Along with this the mean absolute error is 0.3055, the root mean squared error is 0.3909 and relative absolute error is 98.0812, root relative squared error is 99.065.



Classifier: edta -K 1 -M 1.0 -S 1

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

Result list (right-click for options)
15:05:04 - trees.c45
15:05:14 - trees.edta

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         661               31.0913 %
Incorrectly Classified Instances      1465               68.9087 %
Kappa statistic                          0.0668
Mean absolute error                      0.3055
Root mean squared error                  0.3909
Relative absolute error                 98.0812 %
Root relative squared error             99.065  %
Total Number of Instances             2126

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0         0         0          0         0        airport
  0         0         0          0         0        store
  0.573     0.581     0.276      0.573     0.373    movie
  0.556     0.286     0.377      0.556     0.45     restaurant
  0.079     0.069     0.265      0.079     0.121    station

=== Confusion Matrix ===
```

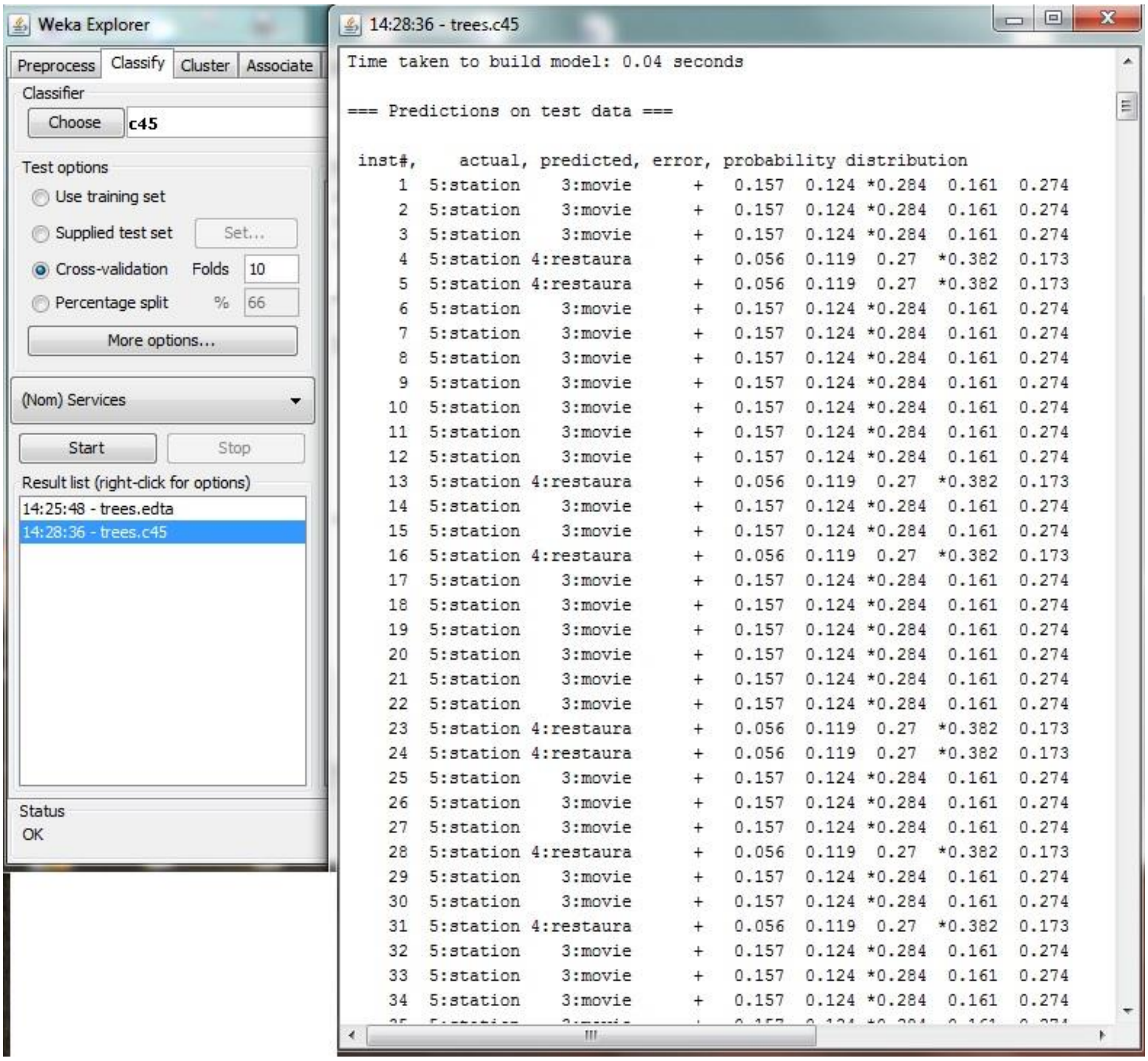
Status
OK

Fig 4.15 Result summary of C4.5 algorithm

From both the above result summaries it can be seen that the total number of correctly classified instances are more of EDTA than C4.5. The total number of incorrectly classified instances are less of EDTA from that of C4.5 algorithm. Along with this the error rate calculated in EDTA is far much less than that of C4.5 algorithm. Due to the more number of correctly classified instances and less error rate the accuracy of the classification result is more of EDTA than that of C4.5 algorithm.

Prediction is done by the C4.5 algorithms. It includes instance number, actual value, predicted value, error, probability and distribution.

The instances which have error in its prediction have + marked under it. This also shows the time taken to build the model.



Classifier: c45

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Services

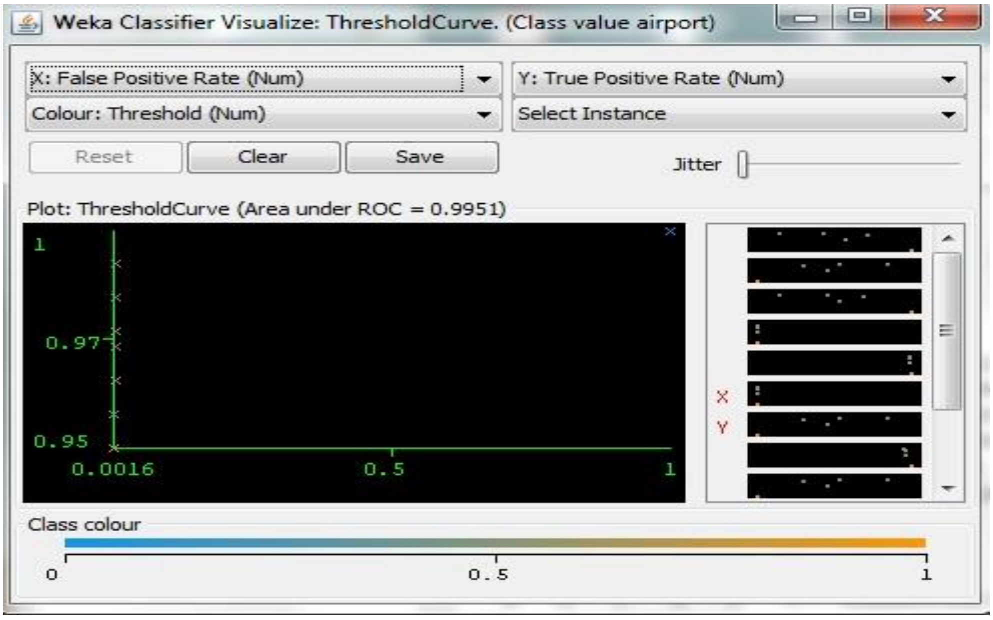
Result list (right-click for options)
14:25:48 - trees.edta
14:28:36 - trees.c45

Status
OK

```
Time taken to build model: 0.04 seconds

=== Predictions on test data ===

 inst#,     actual, predicted, error, probability distribution
     1  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     2  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     3  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     4  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
     5  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
     6  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     7  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     8  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
     9  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    10  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    11  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    12  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    13  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    14  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    15  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    16  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    17  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    18  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    19  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    20  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    21  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    22  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    23  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    24  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    25  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    26  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    27  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    28  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    29  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    30  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    31  5:station 4:restaura      +    0.056  0.119  0.27  *0.382  0.173
    32  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    33  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
    34  5:station    3:movie      +    0.157  0.124 *0.284  0.161  0.274
```

Fig 4.16 C4.5 Algorithm Predictions

WEKA shows the curves for threshold and cost of each parameter of the class attribute. The difference between the outcomes of both the algorithms can also be made with the help of threshold and cost curves



Weka Classifier Visualize: ThresholdCurve. (Class value airport)

X: False Positive Rate (Num)

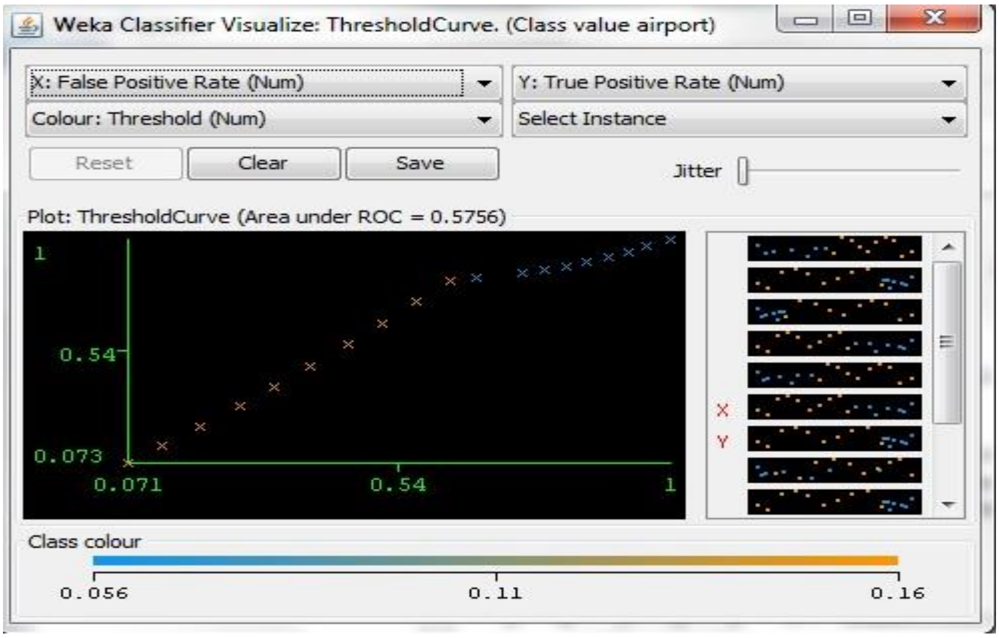
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.9951)

Fig 4.17 Threshold curve of class value airport for EDTA



Weka Classifier Visualize: ThresholdCurve. (Class value airport)

X: False Positive Rate (Num)

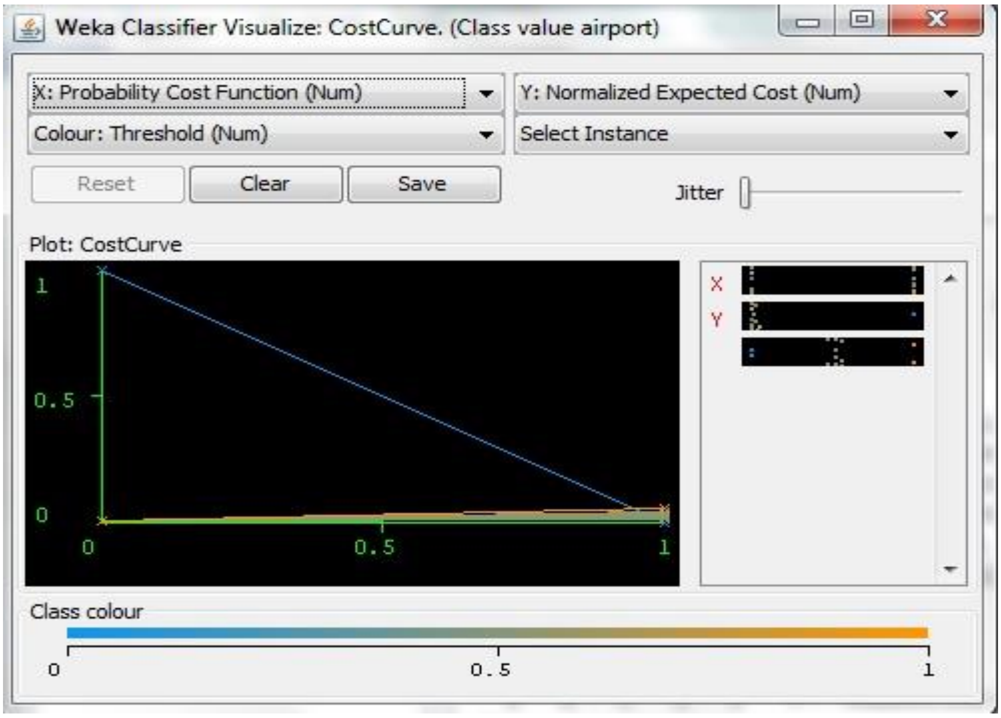
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.5756)

Fig 4.18 Threshold curve of class value airport for C4.5 algorithm



Weka Classifier Visualize: CostCurve. (Class value airport)

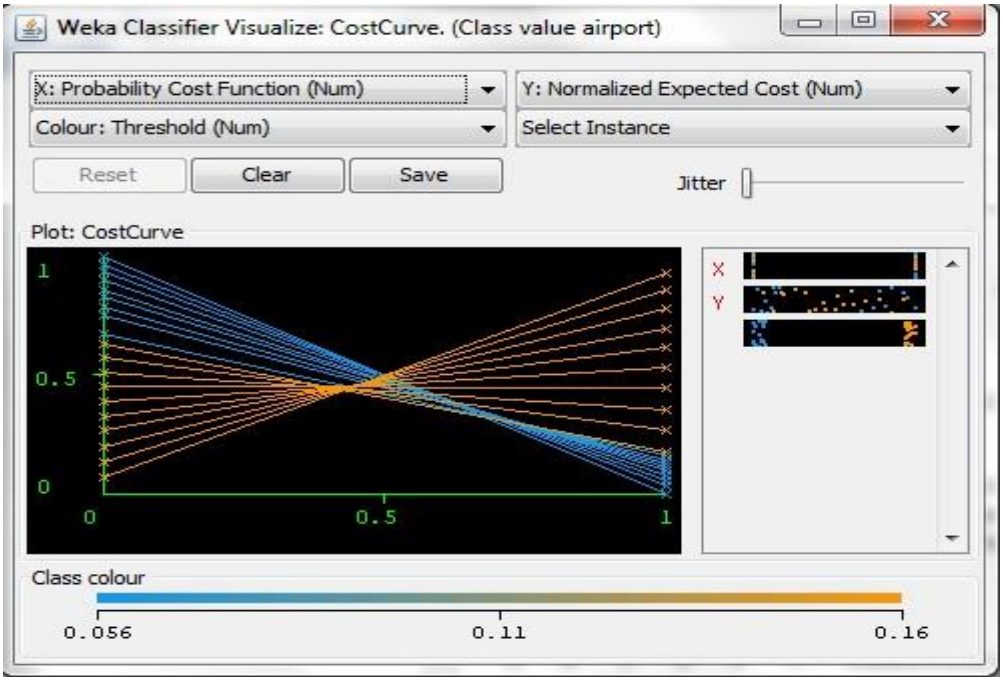
X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.19 Cost curve of class value airport for EDTA



Weka Classifier Visualize: CostCurve. (Class value airport)

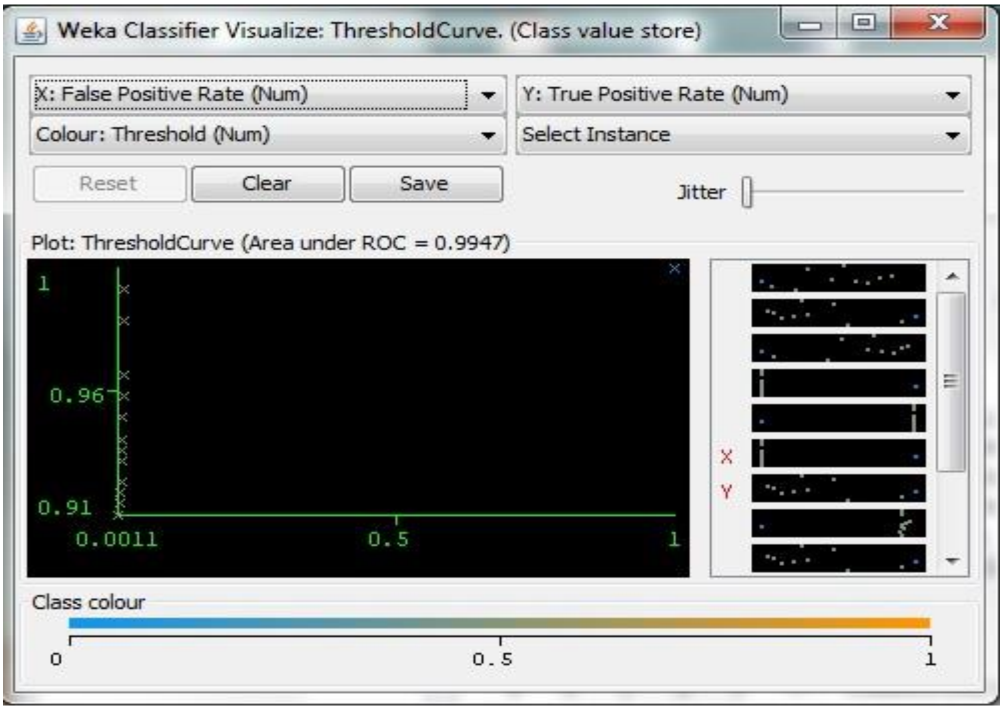
X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.20 Cost curve of class value airport for C4.5 algorithm



Weka Classifier Visualize: ThresholdCurve. (Class value store)

X: False Positive Rate (Num)

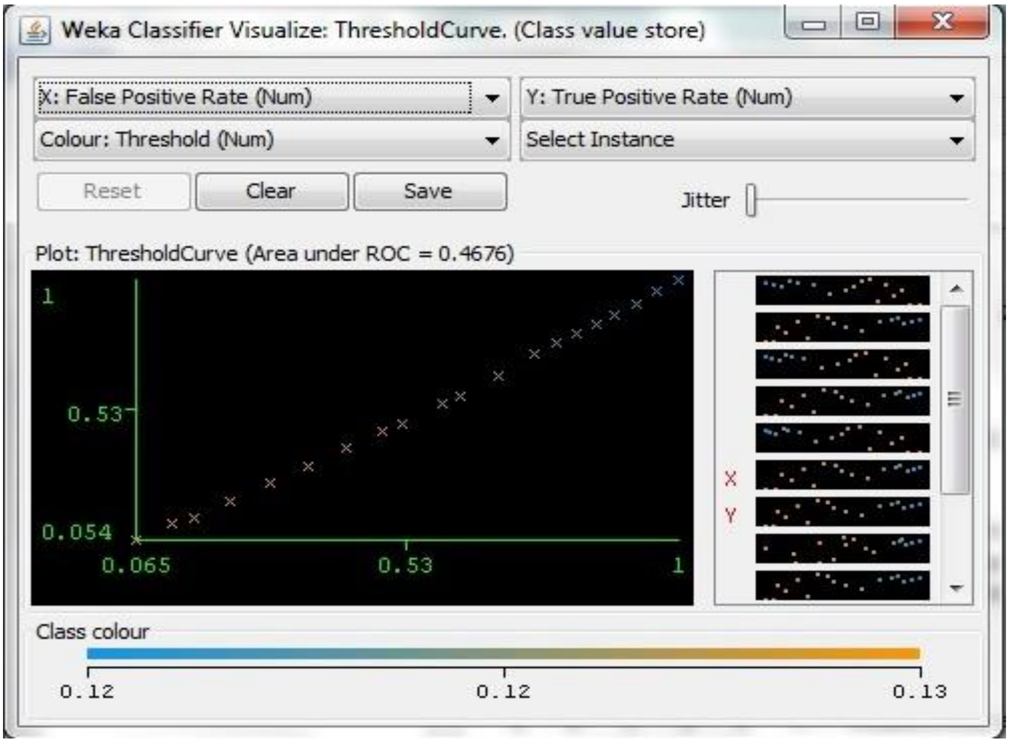
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.9947)

Fig 4.21 Threshold curve of class value store for EDTA



Weka Classifier Visualize: ThresholdCurve. (Class value store)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.4676)

Fig 4.22Threshold curve of class value store for C4.5 algorithm

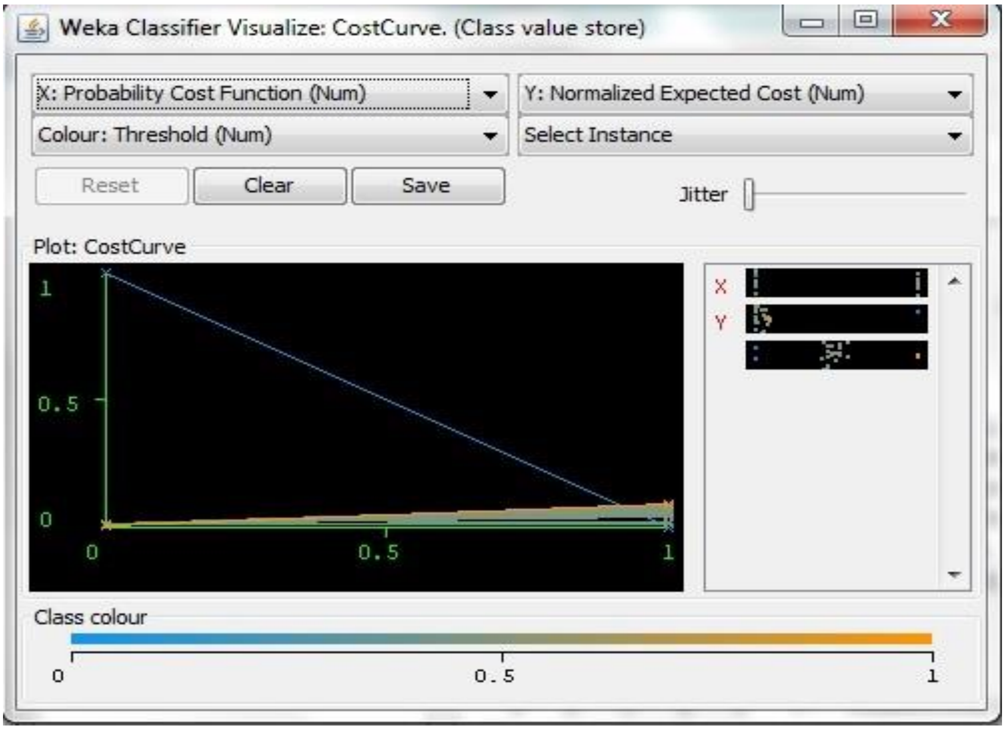

Weka Classifier Visualize: CostCurve. (Class value store)
X: Probability Cost Function (Num)
Y: Normalized Expected Cost (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Save
Jitter
Plot: CostCurve
1
0.5
0
0
0.5
1
X
Y
Class colour
0
0.5
1

Fig 4.23 Cost curve of class value store for EDTA

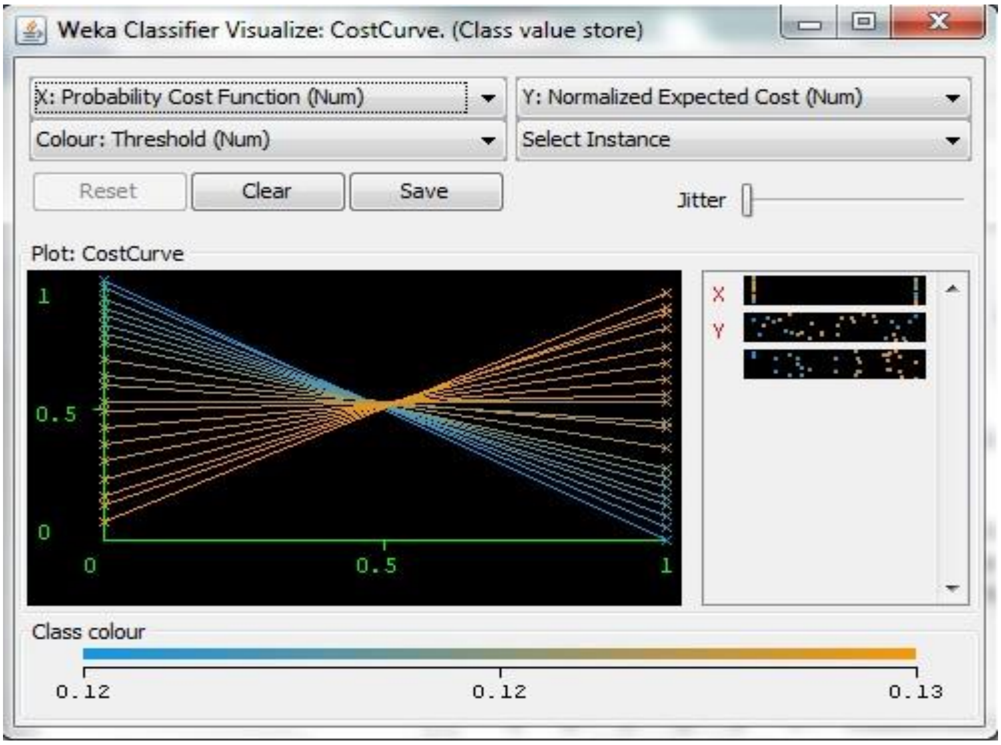
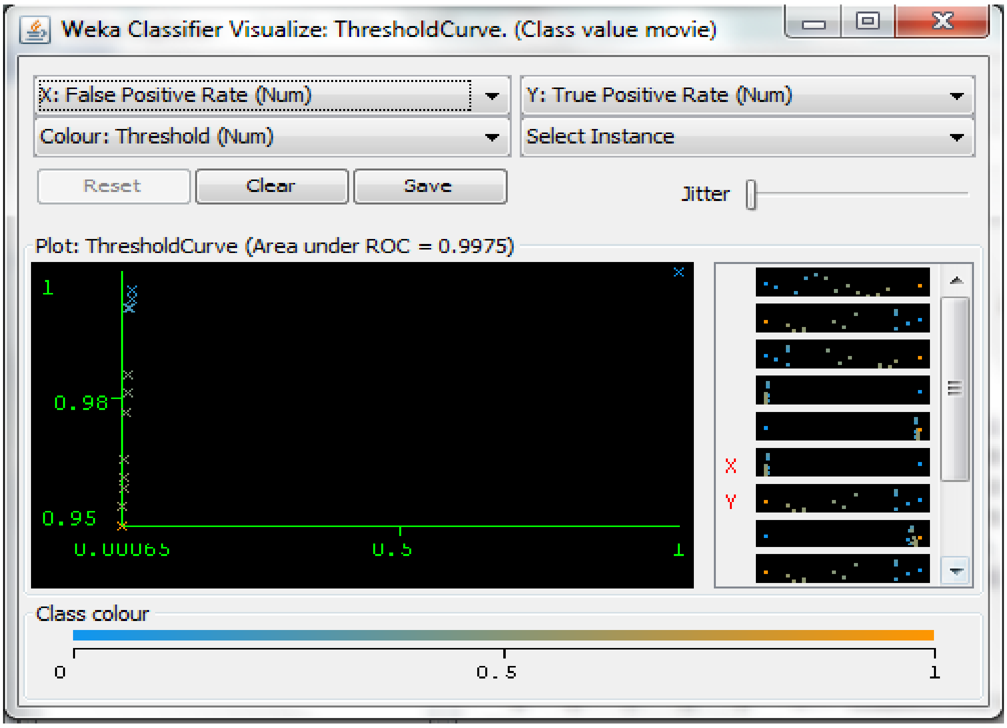

Weka Classifier Visualize: CostCurve. (Class value store)
X: Probability Cost Function (Num)
Y: Normalized Expected Cost (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Save
Jitter
Plot: CostCurve
1
0.5
0
0
0.5
1
X
Y
Class colour
0.12
0.12
0.13

Fig 4.24 Cost curve of class value store for C4.5 algorithm



Weka Classifier Visualize: ThresholdCurve. (Class value movie)

X: False Positive Rate (Num)

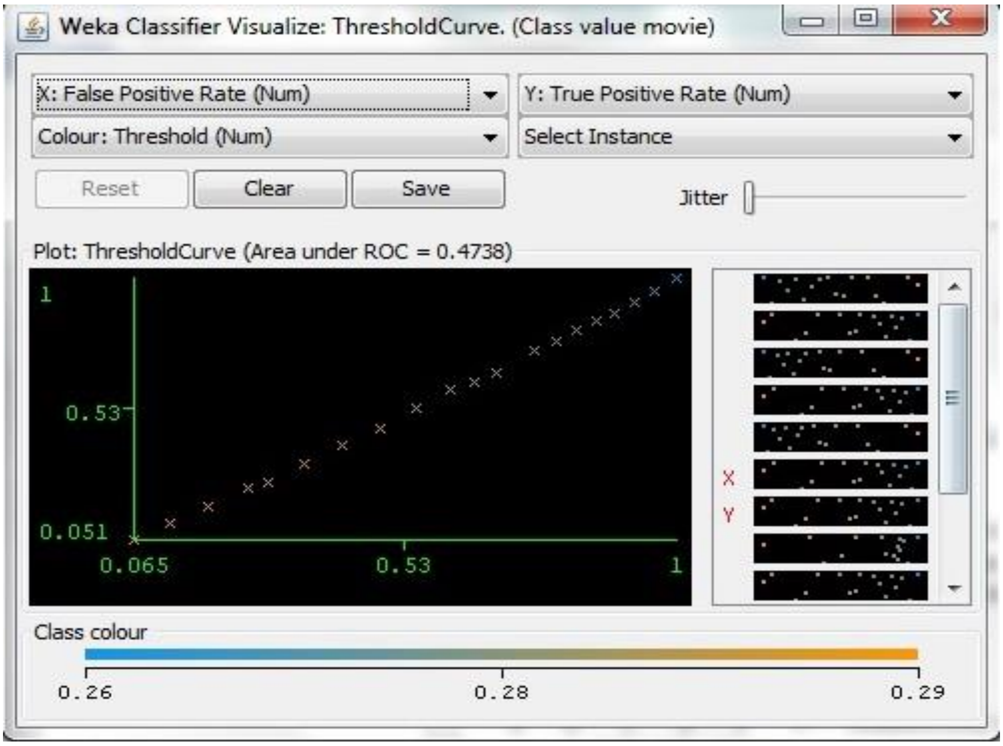
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.9975)

Fig 4.25 Threshold curve of class value movie for EDTA



Weka Classifier Visualize: ThresholdCurve. (Class value movie)

X: False Positive Rate (Num)

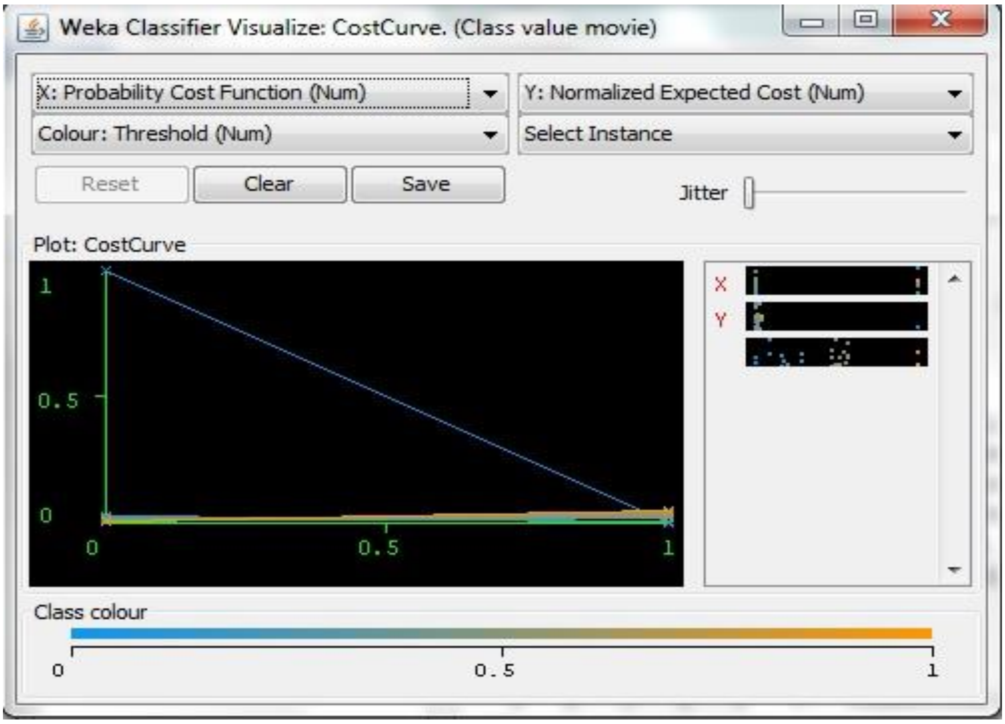
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.4738)

Fig 4.26 Threshold curve of class value movie for C4.5 algorithm



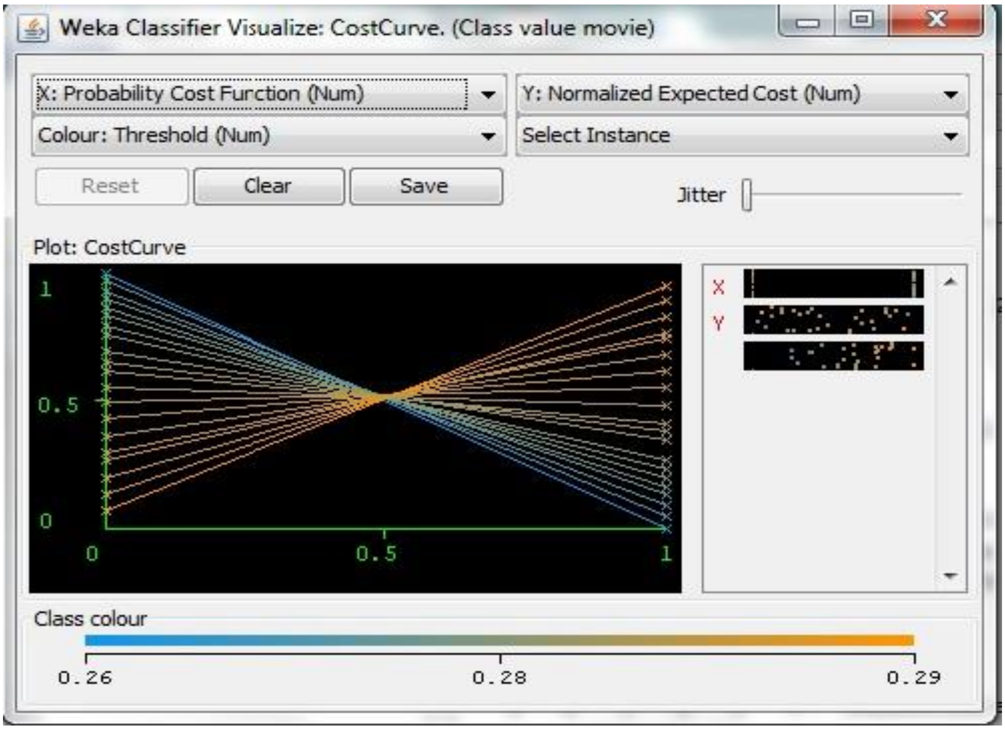
Weka Classifier Visualize: CostCurve. (Class value movie)

X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Fig 4.27 Cost curve of class value movie for EDTA



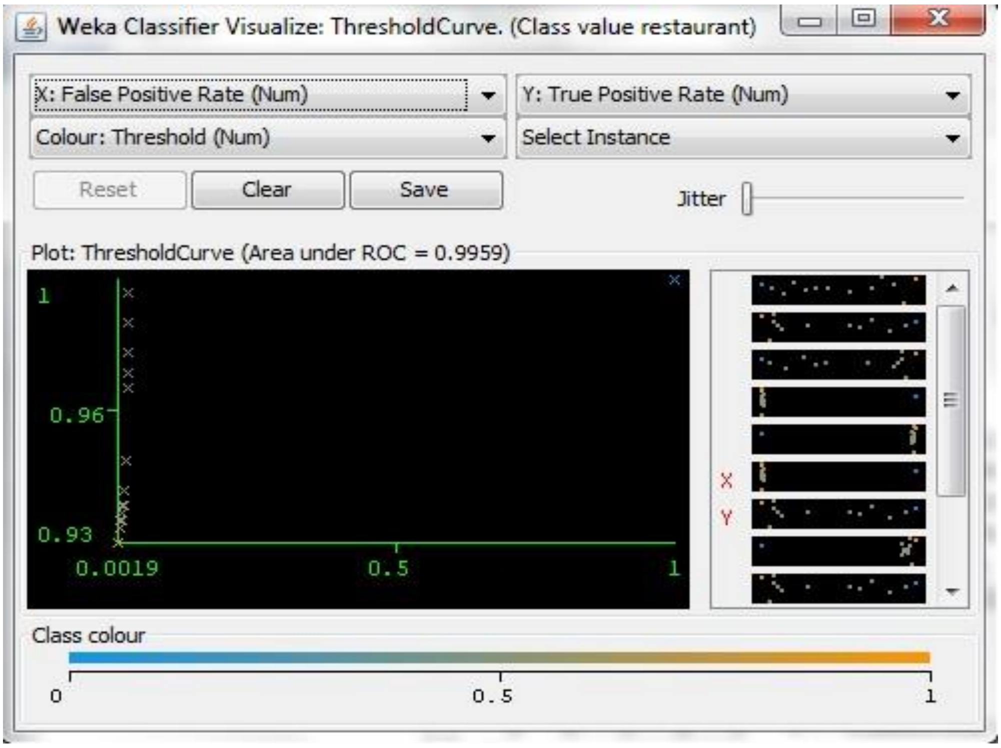
Weka Classifier Visualize: CostCurve. (Class value movie)

X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Fig 4.28 Cost curve of class value movie for C4.5 algorithm



Weka Classifier Visualize: ThresholdCurve. (Class value restaurant)

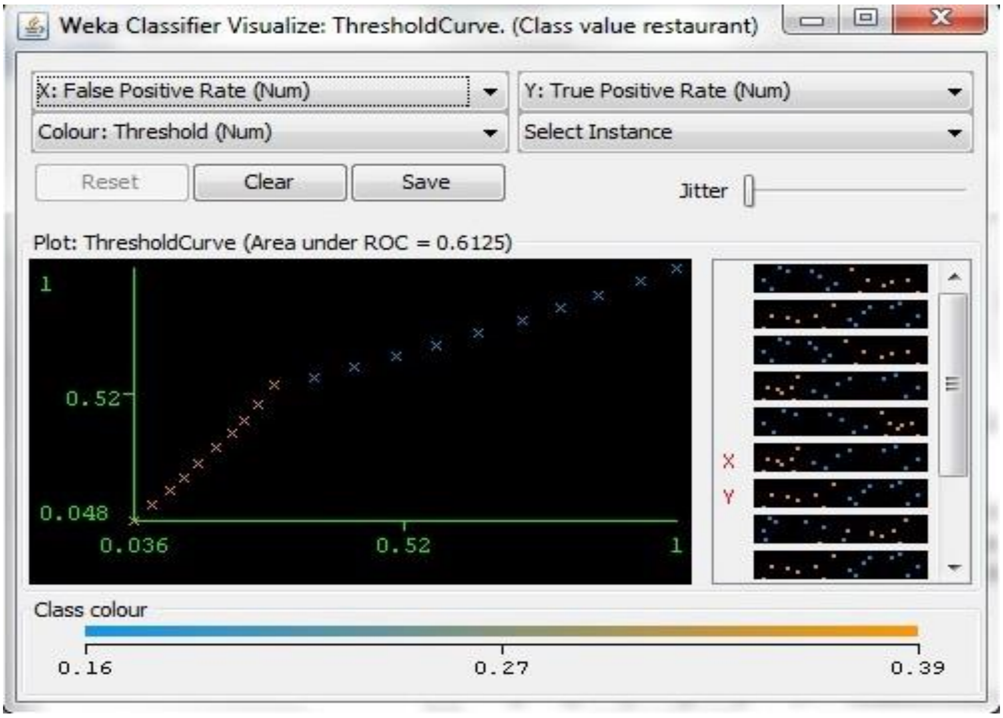
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot: ThresholdCurve (Area under ROC = 0.9959)

Fig 4.29 Threshold curve of class value restaurant for EDTA



Weka Classifier Visualize: ThresholdCurve. (Class value restaurant)

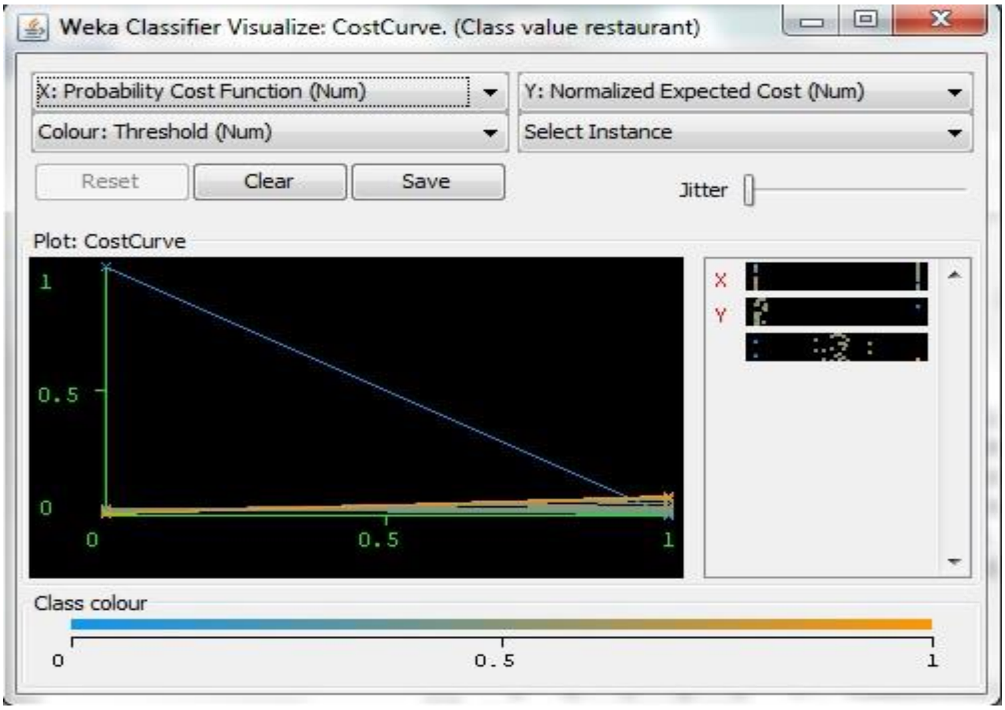
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot: ThresholdCurve (Area under ROC = 0.6125)

Fig 4.30 Threshold curve of class value restaurant for C4.5 algorithm


Weka Classifier Visualize: CostCurve. (Class value restaurant)

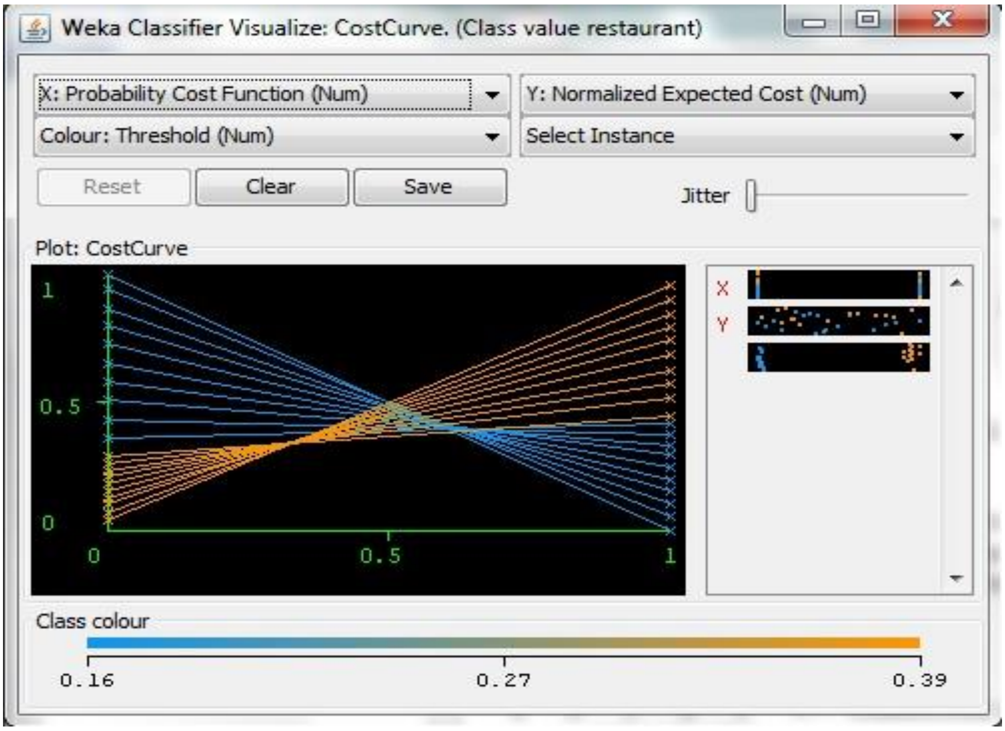
X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.31 Cost curve of class value restaurant for EDTA


Weka Classifier Visualize: CostCurve. (Class value restaurant)

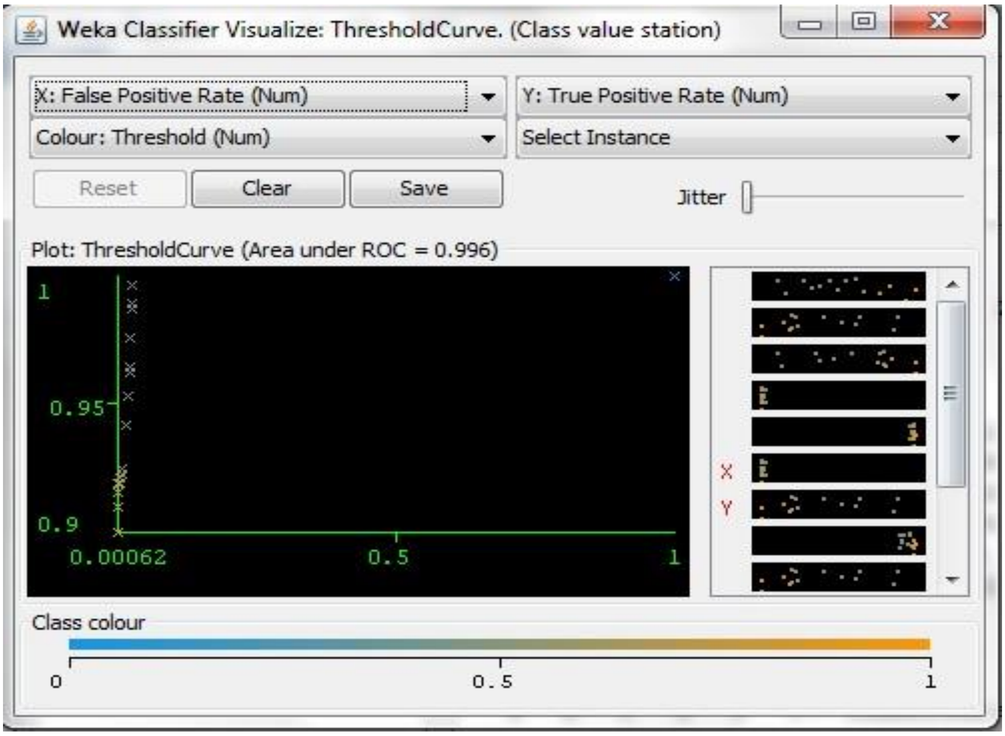
X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.32 Cost curve of class value restaurant for C4.5 algorithm



Weka Classifier Visualize: ThresholdCurve. (Class value station)

X: False Positive Rate (Num)

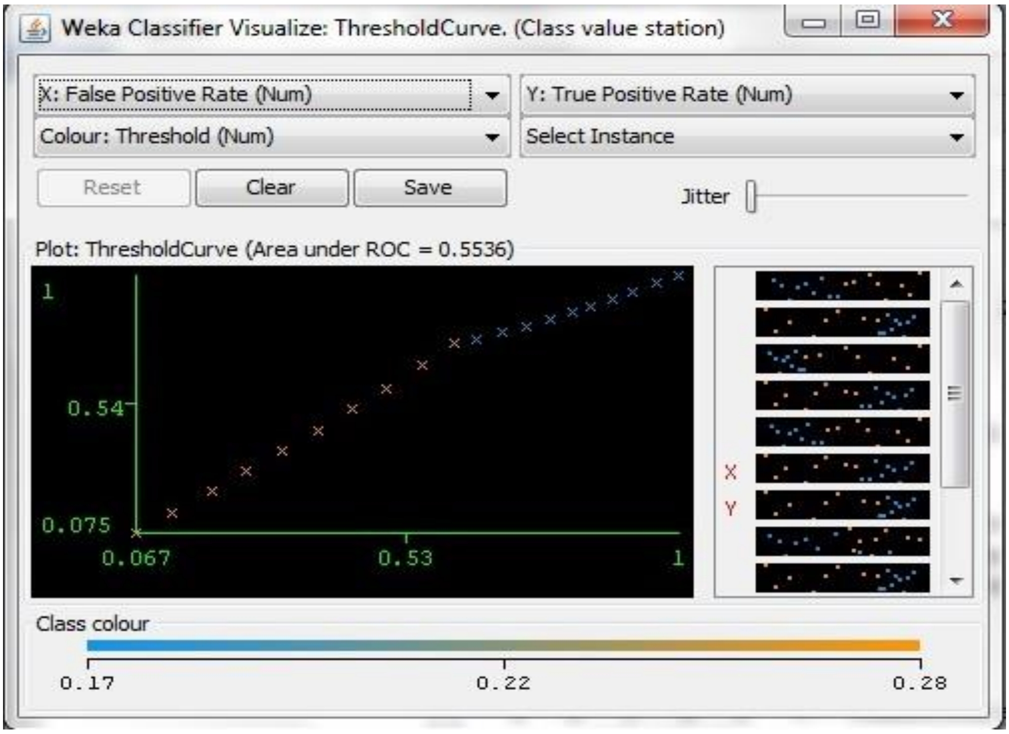
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.996)

Fig 4.33 Threshold curve of class value station for EDTA



Weka Classifier Visualize: ThresholdCurve. (Class value station)

X: False Positive Rate (Num)

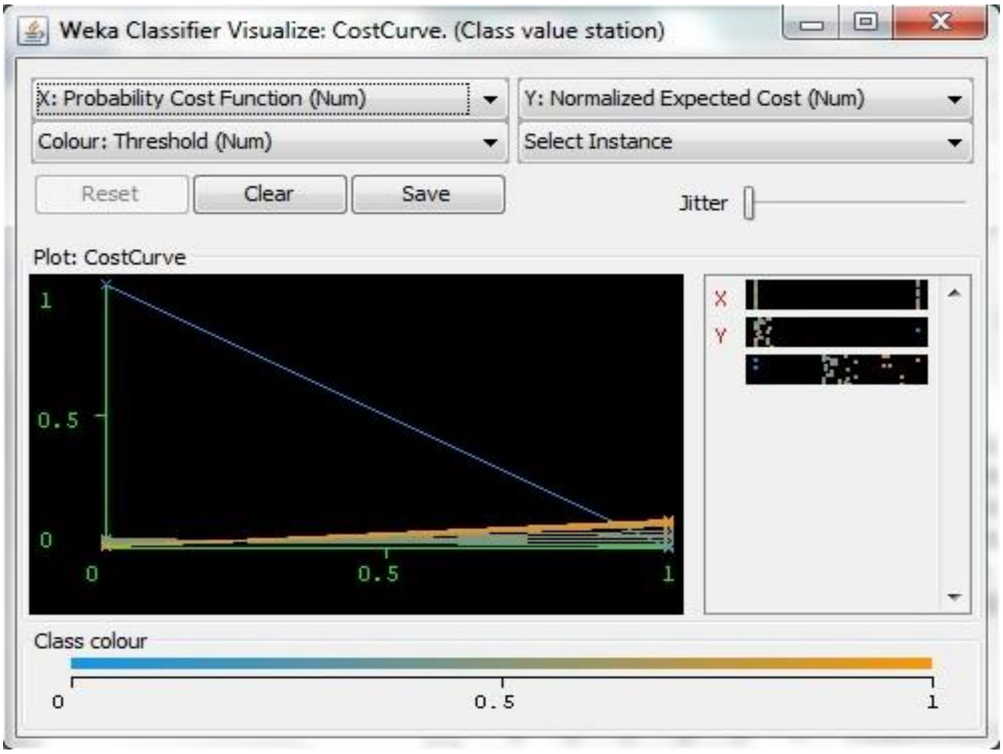
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: ThresholdCurve (Area under ROC = 0.5536)

Fig 4.34 Threshold curve of class value station for C4.5 algorithm



Weka Classifier Visualize: CostCurve. (Class value station)

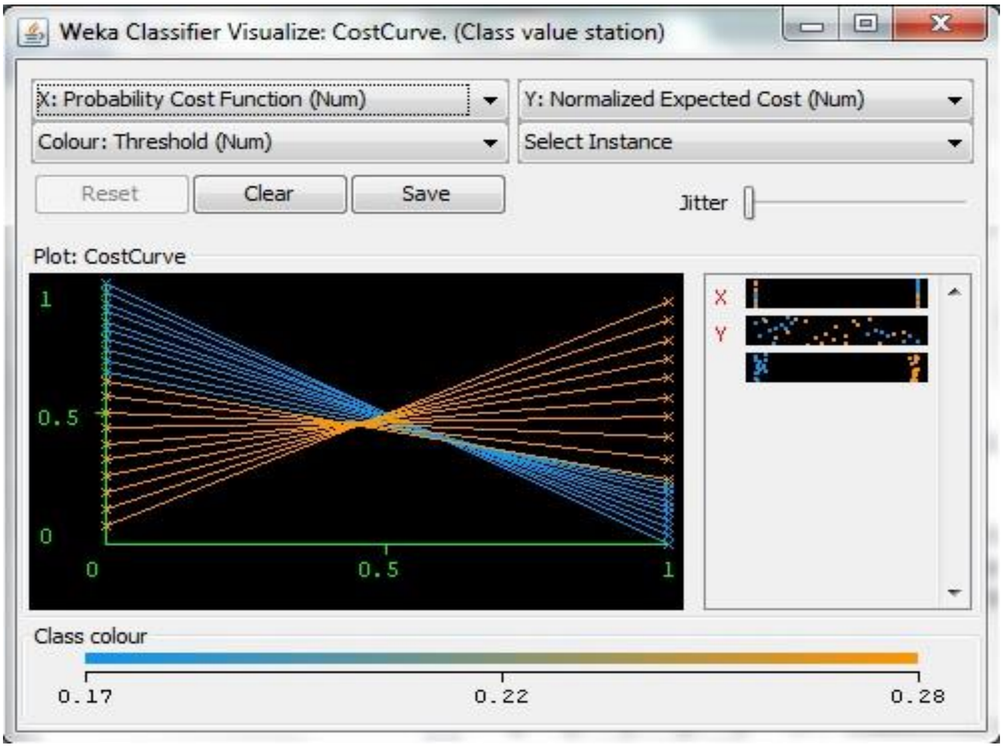
X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.35 Cost curve of class value station for EDTA



Weka Classifier Visualize: CostCurve. (Class value station)

X: Probability Cost Function (Num)

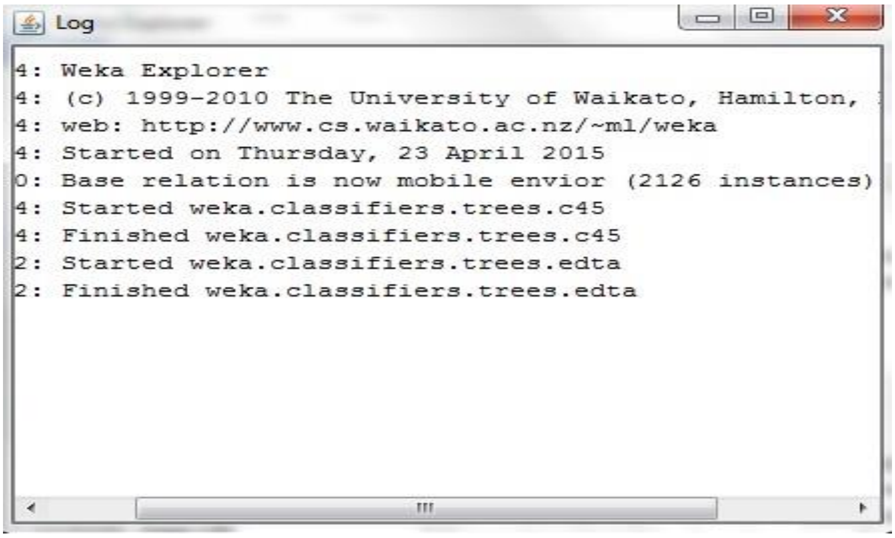
Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 4.36 Cost curve of class value station for C4.5 algorithm

Likewise there are threshold and cost curves for all the other parameters of the class attribute. WEKA stores the log also, this log is generated gives the details of what all operations were performed with the current used session. It tells the starting and execution of an algorithm and tells that which algorithm was executed first.



```
4: Weka Explorer
4: (c) 1999-2010 The University of Waikato, Hamilton,
4: web: http://www.cs.waikato.ac.nz/~ml/weka
4: Started on Thursday, 23 April 2015
0: Base relation is now mobile envior (2126 instances)
4: Started weka.classifiers.trees.c45
4: Finished weka.classifiers.trees.c45
2: Started weka.classifiers.trees.edta
2: Finished weka.classifiers.trees.edta
```

Fig 4.37Log generated by WEKA

Hence the results computed from this research are:

1. The EDTA algorithm is able to work on the large scale high dimensional dataset. Along with this EDTA is also able to work for both discrete and continuous data. It also works on both the nominal and alphabetical data.
2. The number of correctly classified instances of EDTA is more than that of c4.5 and the incorrectly classified instances are less than that of c4.5.
3. The EDTA algorithm have reduced error rate resulting to more accuracy as compared to the c4.5 algorithm. EDTA have relative absolute error rate as 24.2238%.

The results for the comparison of correctly classified instances can easily be computed in a tabular form and bar chart form.

Table 4.1 Number of correctly classified instances

Algorithm	C4.5 Algorithm	EDTA
Number of correctly classified instances	661	2027

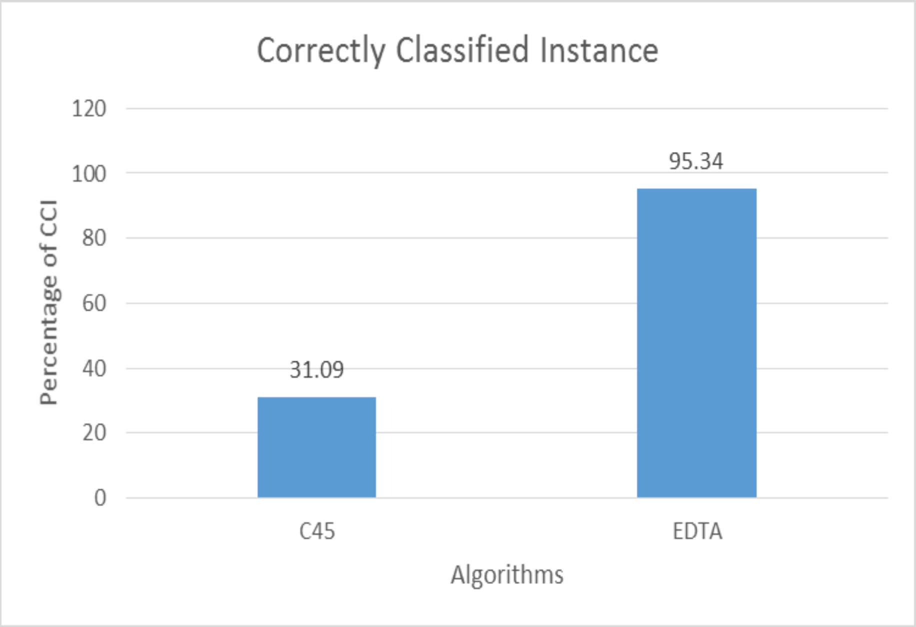


Fig 4.38 Bar graph showing correctly classified instances percentage

The results for the comparison of incorrectly classified instances can easily be computed in a tabular form.

Table 4.2 Number of incorrectly classified instances

Algorithm	C4.5 Algorithm	EDTA
Number of incorrectly classified instances	1465	99

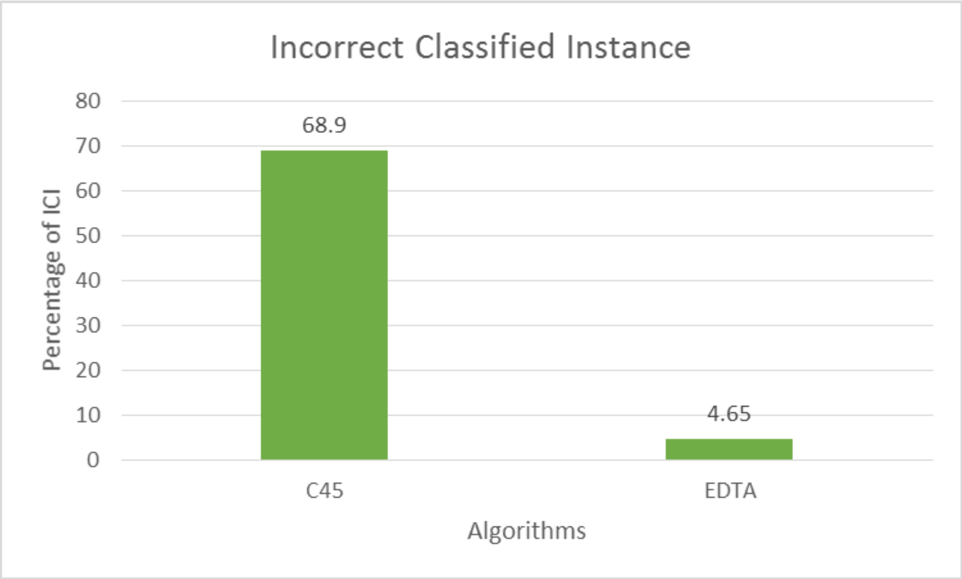


Fig 4.39 Bar graph showing incorrectly classified instances percentage

The results for the comparison of error rate can easily be computed in a tabular form.

Table 4.3 Error Rate

Algorithm	C4.5 Algorithm	EDTA
Error Rate	99.06%	24.22%

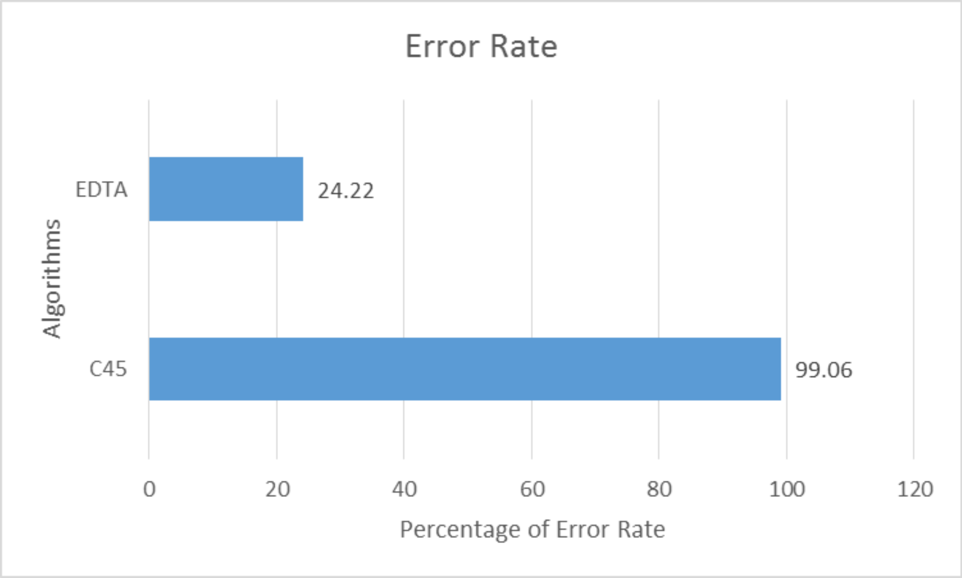


Fig 4.40 Error rate of EDTA and C4.5 algorithm

Chapter 5

SUMMARY AND CONCLUSION

The proposed approach is based on the analysis of mobile networks services and data. The analysis is done with the help of decision tree technique of data mining. The prediction is done on the basis of mobile services that where the particular services are present. The class attribute selected was services which have further five class values and the services for each instance was predicted that where the high usage of network is there out of the three services. This research is based on the analysis of mobile networks services and data. The analysis is done with the help of EDTA algorithm. This research proves that the result from the EDTA algorithm are more accurate than that of c4.5. This is done by the number of classified instances both correctly and incorrectly. This gave the result as well as gave the threshold and cost curves for each class value. This research proved to be a helpful because of the development of new algorithm as it helped to predict the decision in more accurate way with having accuracy of approximately 76%. The results of the algorithm are better when compared with the pre-existing C4.5 algorithm.

REFERENCES

[1] Yang, Q., Yin, J., Ling, C., & Pan, R. (2007). Extracting actionable knowledge from decision trees. Knowledge and Data Engineering, IEEE Transactions on,19(1), 43-56.

[2] Ye, M., Wu, X., Hu, X., & Hu, D. (2013). Multi-level rough set reduction for decision rule mining. Applied intelligence, 39(3), 642-658.

[3] Katz, G., Shabtai, A., Rokach, L., & Ofek, N. (2014). ConfDTree: A Statistical Method for Improving Decision Trees. Journal of Computer Science and Technology, 29(3), 392-407.

[4] Soldic-Aleksic, J. (2012). COMBINED APPROACH OF KOHONEN SOM AND CHAID DECISION TREE MODEL TO CLUSTERING PROBLEM: A MARKET SEGMENTATION EXAMPLE. Journal of Economics & Engineering, 3(1).

[5] Dan, J., Jianlin, Q., Xiang, G., Li, C., & Peng, H. (2010, June). A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree. In Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on(pp. 2722-2728). IEEE.

[6] Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. Information Technology in Biomedicine, IEEE Transactions on, 14(3), 559-566.

[7] Kargupta, Hillol, and Byung-Hoon Park. "Mining decision trees from data streams in a mobile environment." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

[8] Van Hieu, Duong, Nawaporn Wisitpongphan, and Phayung Meesad. "Analysis of factors which impact Facebook users' attitudes and behaviours using decision tree techniques." Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on. IEEE, 2014.

APPENDIX

List of Abbreviations

WEKA - Waikato Environment for Knowledge Analysis

ID3 - Iterative Dichotomiser 3

CART - Classification and Regression tree

CHAID - Chi-squared automatic interaction detector

MARS - Multivariate Adaptive Regression Splines

IDE – Integrated Development Environment