Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 156 – 165

Procedia
Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

A Predictive Model Construction for Mulberry Crop Productivity

Ramya M.C.[a],Lokesh V.[b], Manjunath T.N.[c], Ravindra S. Hegadi[d]

[a]*Research Scholar, Jain University, Bangalore, Karnataka, India*
[b]VSK University, Bellary, Karnataka, India
[c]Acharya Institute of technology, Bangalore, Karnataka, India
[d]Solapur University, Solapur, Maharastra, India

Abstract

India accounts for more than fifty percent of sericulture production in the world. The modern Sericulture methods that have evolved demand, accurate classification of soil suitable for Mulberry crop productivity. But the most prevalent method adopted currently in soil testing is manual, which often fails to give the correct prescription to make soil suitable for Mulberry crop. A scientific approach of soil testing could aid farmers in dynamic decision-making, which would significantly increase Mulberry crop productivity. Such analysis is possible with the help of data analysis, thanks to the advent of modern computer technology. Due to significant advances in the area of Information Technology and agriculture, there is scope of interdisciplinary work, application thereof to solve agricultural problems. Hence effort was made to explore and develop an automated system for the analysis of range of soil characteristic suitable for Mulberry crop production, which in turn contribute to increase in Cocoon productivity. The experiment was carried out by collecting soil samples from different irrigated regions of Karnataka, India, to deduce the range of soil parameters supporting the healthy growth of Mulberry crop. Further, different classification technique was applied on parameters of soil suitable for Mulberry crop using Hunt's algorithm, and J48 Decision tree was more applicable in decision making. The statistical information obtained from data mining technique were validated through mathematical model for developing a forewarning predictive system for crop productivity.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: *Data mining; classification; J48 decision tree; Hunt's algorithm; Soil testing; Sericulture.*

*corresponding author :+91-9845950391
Email address :ramyamc@acharya.ac.in

1877-0509 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.108

1. Introduction

Sericulture Industry, buoyed by revolutionary developments has the potential to play a vital role in rural development, in most parts of India. However there are few intertwined factors restricting its immense growth potential, by limiting cocoon production and productivity. The most critical and significant factors are the lack of knowledge in the farming community, not adopting scientific and advanced tools related to prediction of quantity and quality of the cocoon. Mulberry plant is an important nutrient supplier for the growth of silkworms, and cocoon production in turn. soil nutrient content plays a major role in the growth of Mulberry leaves. Efficient techniques were developed and modified to improve the efficiency and accuracy of soil data sets using data mining classification algorithms [8]. An automated system for checking fertility range of the soil has been developed in our research process. The fertility class family obtained by an automated system is generated using classification techniques employed in data mining tool known as WEKA (Waikato Environment for Knowledge Analysis). Soil examining research laboratory in Sericulture Research and Development Institute (Thalaghattapura, Bangalore) provided dataset required .This research has implemented a sound classification Hunt's algorithm and J48 decision tree for predicting the soil parameter ranges for the growth of quality mulberry plant. The outcome of this research will help in checking the soil for growth of Mulberry plant.

2. Literature review

There are copious factors affecting the healthy growth of mulberry plants. General consensus is that the soil condition, climate condition are the key primary factors which significantly affects the growth. Other secondary factors, which are not as significant as the farmer, but nevertheless of crucial importance are use of chemical fertilizers (and its application thereof at different stage), effects of folic acids, organic manures, influence of phosphate, nitrogen, zinc et al. In nutshell, soil plays a vital role for healthy growth of mulberry plant. There are many significant works and studies conducted related to ideal type of Soil and Climate conditions. Summary of various works, almost always lead to the conclusion that the Ideal conditions are the temperature range varying between 22 – 30^{0}C, rainfall of around 1000-2000 mm, and humidity level of 65-80 percent humidity are optimum for luxuriant growth of mulberry. Further, various studies have been conducted on secondary factors too and unearthed information like the effect of use of chemical fertilizers applied at various stage, the correlation of folic acid fraction of soil organic matter on growth of plant, the effect of different kinds of organic manures and microbial inoculants on leaf yield. In effect it's concluded with a great degree of certainty that the application of organic manures helps in the revival of soil health, use of mulberry wood ash as fertilizer in combination with other mineral and organic fertilizers improves the soil fertility etc.

In this era of computerization and mechanization, no significant study was conducted to automate the whole system. Information on the range of soil fertility parameters, its effect on biochemical composition of mulberry leaf and its relevance to the silkworm breed affecting cocoon production is extremely scanty. With that backdrop, we attempt to do exactly that, with the help of data mining techniques. Further to also evaluate the range of soil fertility by automating the system through collecting soil data from different garden. We have evaluated a range of soil parameters by using data mining technique which is very much suitable for the growth of mulberry plant which in turn affect the characteristic of cocoon.

3. Methodology

The populated data of surveys were carried out in the districts of Mysore and Bangalore, Karnataka, India. Primary data for the soil survey were acquired by field sampling. These sampling were then sent for physical and chemical analysis at the soil testing laboratories in Thalaghattapura, Sericulture Research and Development Institute, Bangalore. Dataset had 10 attributes of soil samples collected from Mysore and Bangalore district. Table-1 describes Nutrients found in soil sample.

Table 1. Parameter contributing for the growth of mulberry leaves

N	Nitrogen
P	Phosphorous
K	Potassium
Fe	Iron
Mn	Manganese
Zn	Zinc
Cu	Copper
PH	PH
EC	Electrical conductivity
OC	Organic Carbon

4. Automated System

The Current manual classification of soil systems, is time consuming, and at times inaccurate. Hence reliable, quick and automated system is needed to make classification better and effective use of technician's time [19]. We propose such an automated system that classifies soil based on fertility for the growth of mulberry leaves. Such a system works as a very powerful tool in classifying soil quickly and accurately based on properties. Input soil data were collected from different region of Karnataka and tested in soil laboratory. The soil was classified into two class labels: Ideal and Not Ideal which was obtained with the help of our automated system and has been further used for decision making.

5. Data Mining & Classification Model

Data mining techniques and algorithms were used and developed to study the fertility of the soil, ideal for the growth of the Mulberry plant. Fig.1 shows flow diagram for classifying soil fertility. Owing to the limited sample size there may be slight differences in depicting the range of soil for predicting the idealness of soil for the growth of mulberry plant. The Hunt's algorithm was used to classify the soil factor for decision making. Relative data from the nearest lab was used to draw decision using Hunt's algorithm. Quality data was accomplished by performing data pre-processing methodology such as data reduction, elimination of noise and null values. Soil samples were collected from Mysore and Thalaghattapura district for the current experiment.

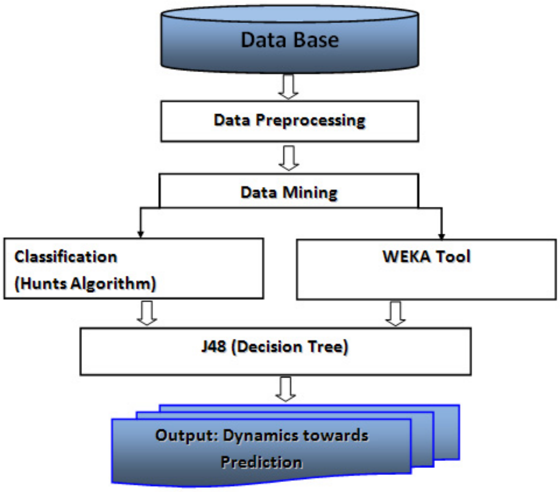
Data Base
Data Preprocessing
Data Mining
Classification (Hunts Algorithm)
WEKA Tool
J48 (Decision Tree)
Output: Dynamics towards Prediction

Fig.1. Soil Classification model.

The values of soil parameter were collected for each samples, and finally Computing maximum and minimum values contributing for ideal growth of Mulberry leaves and these values are supported by soil experiment laboratory in Thalaghattapura silk board. The growth of the Mulberry plant depends on the parameters associated with soil. Data Mining (DM) techniques were used to understand fertility of the soil and to decide the ideal status for the growth of healthy Mulberry plants. The classification technique is an efficient methodology to construct classification models using input data set [12]. For example, decision tree classifiers, rule-based classifiers, support vector machines, and Naïve Bayes classifiers are different techniques to solve a classification problem. Each technique adopts an algorithm that builds a relationship between the attributes and class label of the input data. Therefore, one of the key objective of these algorithms is to build a predictive model that predicts the class labels of previously unidentified records. Decision tree classifier is a straightforward and generally used classification technique. It applies a clear idea to tackle the classification problem. Decision tree classifier generates a series of questions concerning the attributes of the data set. Each time it gets an answer, a catch-up question is asked until a decision about the class label of the record is reached. The Hunt's algorithm was used in association with above decision tree to decide the soil classification. WEKA tools were used in comparison of different algorithm to support the decision and to develop a decision tree which helps in easy understanding. The class label as: Ideal and not Ideal for the fertility of soil obtained were further used to have a comparative study of classification algorithms. The following section describes Random tree, J48, Random forest, Decision stump and REP briefly.

5.1. Random tree

Random decision tree algorithm builds multiple decision trees randomly. In this context "random" implies that each tree in the set of trees has an equal probability of being tested. A tree stopping criteria are: A node becomes empty or the depth of the tree exceeds limit. Random tree models have been extensively developed in the field of Machine Learning in the recent years [14].

5.2. J48 (C4.5)

The J48 Decision tree classifies a new item by creating a decision tree based on the attribute values of training data. When a training set is encountered it identifies the attribute that discriminates the various instances most clearly. The features consisting of data instance that can classify them the best is said to have the highest information gain. The values of these feature which has no ambiguity i.e. having the same value has target variable, will terminate branch by assigning target value obtained. We then continue with other features searching for highest information gain. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [6].

5.3. Random Forest

Random forest tree generally exhibits a significant performance progress over the single tree classifier such as CART and C4.5.

5.4. Decision Stump

A decision stump is basically a one-level decision tree where the split up at the root level depends on a particular attribute/value pair.

5.5. REP

REP is one of the simplest forms of pruning which has the advantage of simplicity and speed. Each node is replaced with class. Change is kept until prediction accuracy is not affected.

In this paper, different Decision tree classifier techniques of data mining were compared and evaluated on the basis of error rate, true positive rate, false positive rate and accuracy. The study showed that J48 model is one of the easiest and best classifier techniques for test data for generating accurate soil range.

Table-2: Accuracy Rate of different Decision tree classifiers

Decision Tree	Accuracy rate
Decision stump	75
J48	100
Random Forest	100
Random Tree	100
REP tree	99

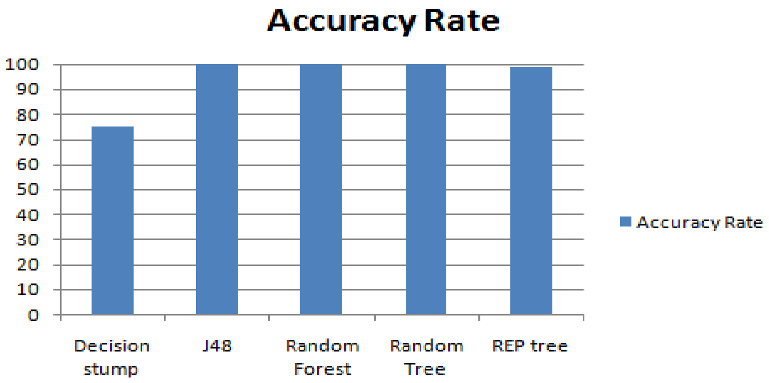

Fig.2. Accuracy rate for Classification Algorithm

Statistical model of classification rules is easily obtained. Moreover these rules are clear and easy to understand for decision making. Hunt's algorithm which acts as base is employed. All decision tree approaches investigated are available in the WEKA package. In this paper we will evaluate the statistical model and hence prove mathematically based on J48 decision tree. Following is the Hunt's algorithm for checking the soil fertility for Mulberry growth.

```
Attributes: PH, EC, OC, PHO, K, SUL, ZN
CLASS:
IDEAL, NOT IDEAL
INPUT: D//Training data
OUTPUT T//Decision tree

$D_n$: the set of training records for node $n$
$Yi$: {Ideal, Not ideal} class labels

1. Formtree (training database $D$)
2. If ($D$ Є $Yi$)
    Create a decision node $N$
    Return a leaf
    Else
3. For each attribute $Ai$
    $D$`= split ($D$, $Ai$)
4. If (N≠ leaf node)
    Create Child of node
    $N$

5. Child of $N$ = Formtree ($D$`)
```

6. Measuring Performance

The performance of Decision Tree classifier is generally examined by calculating the accuracy of each tree. Determining which is best depends on how the user interprets the problem. In this paper J48 Decision tree is employed.

6.1. J48 Statistics

J48 algorithm of classification was used for developing multi variant equation for the training test data. This technique provides basis to understand how the typical values in each parameter splits to form a growing tree. While constructing a tree, J48 ignores the missing values [3].When developing the decision tree using J48 with multivariate data we can evaluate the accuracy rate. It is a data mining technique under classification that generates decision tree by evaluating various soil parameters for checking the idealness of mulberry growth. J48 allows classification through decision trees or generated rules from them.

6.2. Confusion Matrix

A confusion matrix is a table with two rows and columns that contains information about predictive and actual classification performed by a classifier system. Each cell in the table represents the number of true positives, false negatives, false positives and true negatives. Zero values outside the diagonal represent the best solution [16]. The Table-3 shows the confusion matrix for J48 classifier. The entries in the confusion matrix in Table-3 have the following meaning [5] :

1. a is the number true positives

2. b is the number of false negatives

3. c is the number false positives

4. d is the number true negatives [17].

6.3. Interpretation of Confusion Matrix.

1. True positive (TP) : Positive cases that were identified correctly.

2. False Negative (FN) : Positive cases that were falsely identified as negative.

3. False positive (FP) : Negative cases that were incorrectly identified as positive

4. True Negative (TN) : Negative cases that were identified correctly as negative.

In this paper we have used WEKA tool (Waikato Environment for Knowledge Analysis) for calculating accuracy based on correct and incorrect classes generated by confusion matrix .The attributes are soil parameter consisting of ph, electrical conductivity, organic carbon, phosphorus, Potassium, Sulphur and Zinc.

7. Experimental Work and Results

Classification is performed using J48 classifier. Visualize tree shows the J48 classification to classify soil parameters into Ideal or Not for the growth of Mulberry plant in Fig: 2 and Confusion matrix Table-3 depicts the decision for soil dataset. Confusion matrix classifies into two possible values i.e. IDEAL or NOT IDEAL.

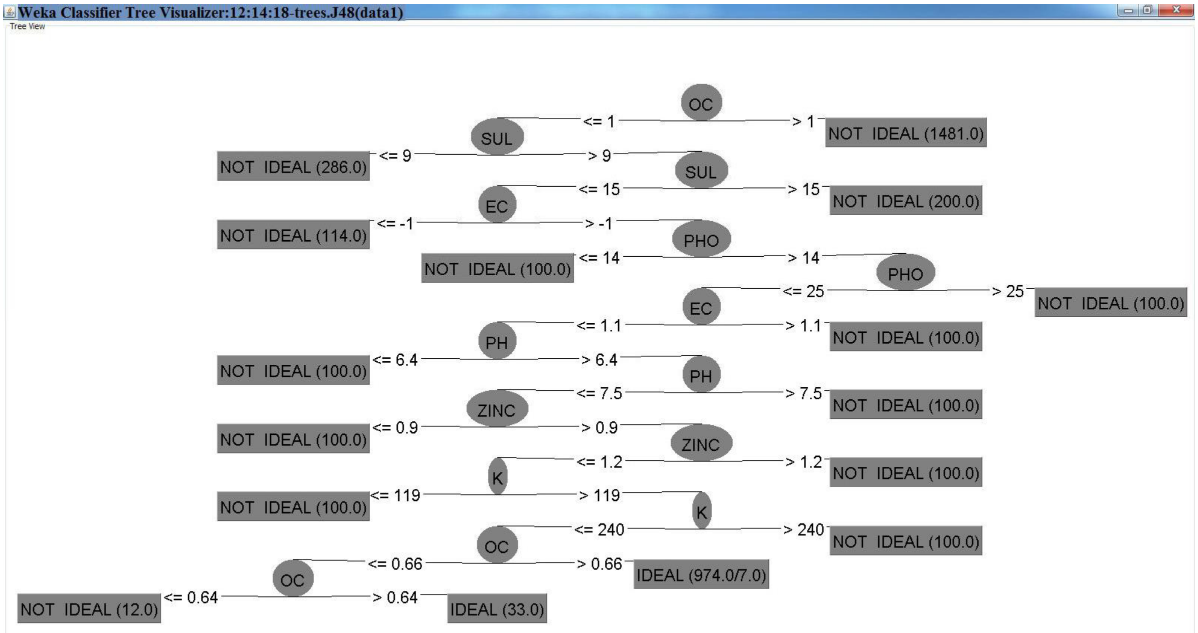

Fig.3. Visualize tree

Decision tree algorithms such as ID3, C4.5, J48 NBTree can be applied on large amount of data and valuable predictions can be produced. Decision tree learning starts from root node. Each node of the decision tree in turn represents some discreet information. Discrete values, which act as target function, are produced at each node by testing the values of attribute. Then by using target function, value of attribute for next node is evaluated. This cycle is repeated for each new node. The learned tree is represented by if-then rules.

In our research work J48 decision tree algorithm using WEKA is made use of, along with Decision tree that depicts soil parameters. The result shows the range of each parameter ideal for the growth of mulberry plant.

Table-3: Confusion Matrix

Decision	A(Ideal)	B(Not Ideal)	Total
YES	1000 (a)	0 (b)	1000
NO	7 (c)	2993 (d)	3000
Total	1007	2993	4000

For above confusion matrix:

1. True positive (TP) : 1000 positive cases were interpreted correctly.

2. False Negative (FN) : 0 Positive cases were interpreted incorrectly.

3. False positive (FP) : 7 Negative cases were interpreted correctly.

4. True Negative (TN) : 2993 Negative cases were interpreted correctly.

Since True positives are 1000 and True Negative are 2993 i.e. diagonal elements of matrix 1000+2993 =3993 represents correctly classified instances and other elements False Negative and false positive i.e. 0+7 = 7 represents the incorrect instances.

Mathematical Results:

Methodical proof for True Positive Rate, False Positive Rate and Precision is given below.

$$\begin{aligned}\text{TP Rate} &= TP / (TP+FN) \\ &= 1000 / (100+0) \\ &= 1\end{aligned} \tag{1}$$

$$\begin{aligned}\text{FP Rate} &= FP / (FP+TN) \\ &= 7 / (7+2993) \\ &= 0.002\end{aligned} \tag{2}$$

$$\begin{aligned}\text{Precision} &= TP / (TP+FP) \\ &= 1000 / (1000+7) \\ &= 0.99\end{aligned} \tag{3}$$

$$\begin{aligned}\text{FP Rate} &= 2TP / (2TP+FP+FN) \\ &= (2*1000) / (2*1000+7+0) \\ &= 0.99\end{aligned} \tag{4}$$

7.1. Kappa Statistic:

Kappa Statistic compares the accurateness of the system with the random system. The Measurement of Observer Agreement for Categorical Data, is an observational probability of agreement and is a hypothetical expected probability of agreement under a baseline constraint for appropriate set [5].

$$\text{Kappa} = (\text{total accuracy} - \text{random accuracy})/(1 - \text{random accuracy}) \tag{5}$$

In Table-3 confusion matrix

$$\begin{aligned}\text{Total Accuracy} &= (1000+2993)/4000 \\ &= 0.998\end{aligned}$$

$$\begin{aligned}\text{Random Accuracy} &= (5.0017+2244.75)/4000 \\ &= 0.5624\end{aligned}$$

$$\begin{aligned}\text{Kappa} &= (0.998-0.5624)/(1-0.5624) \\ &= 0.9954\end{aligned}$$

Hence to support our experiment few Equations

7.2. Mean Absolute Error(MAE):

It is the average of the difference between predicted (output value) and actual (target value) value in all tested cases.Mean Absolute Error is one of the way to compare predicted value with actual outcomes.

$$\frac{1}{n}\sum_{i=1}^{n}| t_i - o_i | \qquad (6)$$

7.3. Root Mean Squared Error(RMSE):

It is similar to mean absolute error but taking squares and then square root and it used to measure the average magnitude of the error. Lower values are better. Since MAE=RMSE in our experiment errors are of same magnitude.

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{|t_{i\_}o_i|}^{2} \qquad (7)$$

7.4. Relative Absolute Error:

It is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the target values.

$$\frac{\sum_{i=1}^{n}|t_{i\_}b_i|^{2}}{\sum_{i=1}^{n}|t_{i\_}\bar{t}|^{2}} \qquad (8)$$

7.5. Root Relative Squared Error:

This is similar to relative absolute error but taking squares and then square root

$$\sqrt{\frac{\sum_{i=1}^{n}|t_{i\_}b_i|^{2}}{\sum_{i=1}^{n}|t_{i\_}\bar{t}|^{2}}} \qquad (9)$$

Table-4: Summaries the results for the formula defined above.

Statistics	Result
Kappa Statistic	0.99
Mean Absolute Error	0.00
Root mean squared Error	0.04
Relative absolute Error	0.92
Root relative squared Error	9.62

8. Conclusion

In this paper, we have analyzed the soil test data collected from different region of Karnataka and predicted the range of soil parameter best suited for mulberry growth through an automated supported by classifier algorithm.

Data has been tested using different classification algorithm. We have drawn decision tree using J48 (C4.5) from data mining tool WEKA to generate a predictive model which predicts whether soil is "Ideal" or "Not Ideal" based on soil parameters, for the growth of Mulberry plant. Further, the experiment is validated through mathematical formulas. The Hunt's algorithm is generalized for soil parameter. In future, we intend to build an automated system for Predicting type of Mulberry plant suitable for different soil type.

References

1. A Sravani, DND Harini. A Comparative Study of the Classification Algorithms Based on Feature Selectio .*Springer* 2014; **249**: 97-104.
2. SS Baskar, L Arockiam and S Charles. Applying Data Mining Techniques on Soil Fertility Prediction. *International Journal of Computer Applications Technology and Research* **2013**; 2:660 - 662.
3. Tina R Patil, Mrs.S S Sherekar Sant Gadgebaba. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications* 2013; **6**:2.
4. AK Tripathy, J Adinarayana,SN Merchant, UB Desai, K Vijayalakshmi, D RajiReddy, S Ninomiya, M Hirafuji, T Kiura.Data Mining and Wireless Sensor Network for Groundnut Pest Thrips Dynamics and Predictions. *Journal of Emerging Trends in Computing and Information Sciences* 2012; **3**:6.
5. Anshul Goyal,RajniMehta. Performance Comparison of Naïve Bayes and J48 Classification Algorithms. *IJAER* 2012; **7**: 11.
6. Naïve Bayes. Wikipedia, February 2012.
7. C4.5 (J48).Wikipedia. February 2012.
8. A Kumar and N Kannathasan. A Surveyor Data Mining and Pattern Recognition Techniques for Soil Data Mining. *International Journal of Computer Science* 2011; **8**: 1.
9. A Sharma. A Comparative Study of Classification Algorithms for Spam Email Data Analysis. *International Journal on Computer Science and Engineering* 2011; **3**.
10. Milan Kumari and Sunila Godara. Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction. *IJCST* 2011; **2**:304-308.
11. M Kumari and S Godara. Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. *International Journal of Computer Science and Technology* 2011; **2**.
12. http://mines.humanoriented.com/classes /2010/fall/csci568/portfolio_exports/lguo/ decisionTree.html 2010.
13. Mitchell TM. Generative and Discriminative Classifier: Naive Bayes and Logistic Regression. *In Machine Learning McGraw Hill* 2010;1-17.
14. Yongheng Zhao and Yanxia Zhang. Comparison of decision tree methods for finding active object. *Advances in Space Research* 2007.
15. Hong Hu, Jiuyong Li, Ashley Plank. A Comparative Study of Classification Methods for Microarray Data Analysis. *CRPIT* 2006; 61.
16. Margaret H Danham, S Sridhar. *Data mining, Introductory and Advanced Topics*. Person education 2006.
17. Xiang yang Li,Nong Ye. A Supervised Clustering and Classification Algorithm for Mining Data with Mixed Variables. *IEEE* 2006; 36:396-406.
18. Sally Jo Cunningham and Geoffrey Holmes. Developing innovative applications in agriculture using data mining. *SEARCC* 1999.
19. A Al-Rawas, S Al-Alwai, A Bisma and Y Al-Alwai. *Soil Classification Decision Support System Using An Expert System Approach. Engineering Journal of University of Qatar* 1998; **11**:103-115.
20. P W Eklund, S D Kirkby and A Salim. Data Mining and Soil salinity Analysis. *International Journal of Geographical Information Science* 1998; **12**:247-268.
21. George, H J and Pat L. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* 1995; 338-345.