VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN EFFECTIVE PREDICTION ANALYSIS USING J48

Bhuvaneswari T[1], Prabaharan S.[2] and Subramaniyaswamy V.[3]
[1]Department of Computer Science and Engineering, India
[2]Vinayaka Missions Kirupananda Variyar Engineering College, Salem, India
[3]SASTRA University, Thanjavur, India
E-Mail: basweety4@gmail.com

ABSTRACT

Classification is the one of the well-known techniques in data mining. Based on the attributes of the object, classification assigns an object to one of numerous pre-defined categories. If information gain is not good then split attributes values into groups until we get better classification ratio. J48 is the one of the most frequently used classification techniques. In this paper, J48 is employed to effective prediction analysis of Iris data set. Three types of Iris flower with 250 instances and five attributes is used as test and training data. The results show that the accuracy of prediction is improved when compared with the existing ID3 method.

Keywords: predictive analysis, data mining, web mining, web documents, classification.

1. INTRODUCTION

Data Mining, popularly known as Knowledge Discovery in Databases (KDD), is a process of extracting hidden, previously unknown, possibly valuable information and knowledge from a huge number of incomplete, noisy, uncertain and arbitrary data. Many algorithms were developed and employed to excerpt information and discover knowledge patterns that may be suitable for decision support.

Classification is a method of discovering a set of models that depict and differentiate data classes and concepts. This model is then used to predict the class whose label is unknown [4]. The resultant model is based on the analysis of a set of data objects whose class label is known called training data. This resultant model can be represented in a variety of formats such as classification rules, mathematical formulae, decision trees, or neural networks. The aim of classification is to precisely predict the target class for each case in the data [5, 6]. Classification can be classified as binary or multiclass classification. In binary classification, data objects are assigned into one of the two groups. Multiclass classification is more complex than binary classification as three or more groups are involved [8]. Classification technique makes use of mathematical methods such as decision trees, linear programming, neural network and statistics [6].

Decision tree learning is a normally used method which uses a decision tree as a predictive model that maps observations about an item to conclusions about the item's target value. The goal is to build a model that foresees the value of a target variable based on numerous input variables [11]. Decision tree is a widely used method to model classification and prediction. Decision trees can handle high dimensional data and it can be simply converted to classification rules. The learning and classification process are simple and fast with superior accuracy. Decision tree induction algorithms have been used for classification in various applications [5].

2. RELATED WORK

NB Tree, a decision tree learner, is presented that consists of Naive Bayes classifiers as leaf nodes and used a split condition that is based on the performance of Naive Bayes classifiers in all initial-level child nodes [15]. Support Vector Machine (SVM) is indeed powerful classification methodology that has been applied in a wide range of applications. The essential idea in SVM is that the hyper plane classifier, or linear linear separability [21].

K-Nearest Neighbor (KNN) classification classifies instances supported their similarity. It is one in all the foremost well-liked algorithms for pattern recognition. It is a sort of Lazy learning where the function is merely approximated locally and every computation is delayed till classification. Associate object is classed by a majority of its neighbors. K is often a positive whole number. The neighbors are selected from a group of objects that the right classification is known [22].

Neural networks have begun as a vital tool for classification. The current research activities in neural classification have recognized that neural networks are an encouraging alternative to a number of conventional classification systems. The benefit of neural networks lies in the subsequent theoretical facets. Neural networks are data driven self-adaptive approaches that can correct themselves to the data without any explicit specification of functional or distributional form for the fundamental model [23].

A feed-forward back-propagation network called multilayer Perceptron (MLP) is the most often used neural network in pattern recognition. MLPs are supervised learning classifiers that contains input layer, output layer, and one or a lot of hidden layers that extract helpful information throughout learning and allot modifiable weighting coefficients to parts of the input layers [19, 20].

ID3 algorithm is a significant algorithm in the decision tree to this point. A new algorithm combining ID3 and Association Function (AF) is proposed due to the limitation of ID3 to select attributes with several values [1]. A random training subset is selected and a decision

tree is made from it. This tree classified all objects properly within the training subset. All alternate objects within the training set are then classified by means of the tree [6].

ID3 employs greedy approach to decide on the most effective attribute. The attribute with most information gain is selected and attributes values are split into groups if information gain is not good [11]. The improved classification algorithm had resolved the problems of ID3 after the principles and implementation steps are examined, though the classification accuracy and time are not sufficiently good enough [17].

J48 classifier is a simple and easy implementation of c4.5 algorithm for making decision trees for classification. A decision tree is constructed to model the classification method. After the tree is constructed, it is applied to every tuple in the database and results in classification for that tuple [13] [14].

3. MATERIALS AND METHODS

In this paper, J48 is employed for effective prediction analysis. Iris data set with three types of Iris flower with 250 instances and five attributes is used as test and training data. In decision tree structure, all internal node denotes a test on an attribute, every branch denotes an outcome of the test, and each leaf node having a class label. J48 classifier is a simple classification technique to create a binary tree. A classification process model is constructed to create a decision tree. The model is shown in the figure 1. A decision tree consists of two phases. In the initial phase, all the training sets are taken as root and based on the partition the attributes are selected. The second phase identifies the outliers and removes the branches. A set of training data builds the decision tree using the concept of entropy. The training data set $X_i=\{X_1, X_2 ..,X_n\}$ of the classified samples. For each sample X_i is a vector and $X_1, X_2 ..,X_n$ are the attributes or features. It is applied to each tuple on the data base and the tree is built for each tuple.

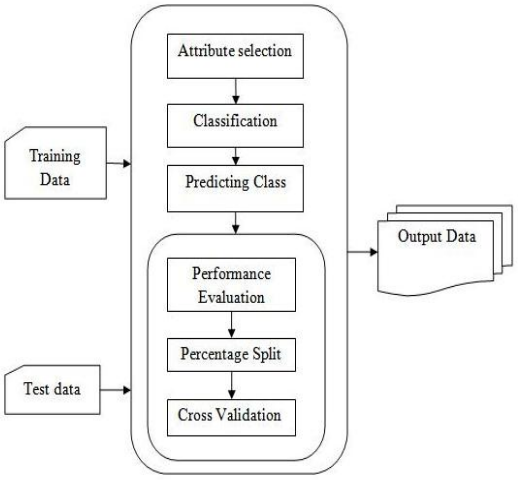

Figure-1. The proposed model.

3.1 Attribute selection

To select the test attribute at each node in the tree the measure called information gain is used. The attribute with the greatest entropy reduction or highest information gain is preferred as the test attribute for the present node. This attribute reduces the information required to classify the samples in the resultant partitions. Entropy measures the quantity of disorder or uncertainty in a system. In the classification setting, higher entropy matches to a sample that has a varied group of labels and lower entropy matches to a case where clean partitions are. The entropy of a sample *E* is given by

$$F(E) = -\sum_{j=1}^{n} Q(D_j / E)\log_2 Q(Dj / E)$$

where Q (D_j / E) is the probability of a data point in E being labeled with class D_j and *n* is the number of classes. Q (D_j / E) is estimated from the data as

$$Q(D_j / E) = \frac{|\{y_i \in E \mid y_i haslabelx_i = D_j}{|E|}$$

The weighted entropy of a decision or split is given by

$$F(E_{M,}E_S) = \frac{|E_M|}{|E|}F(E_m) + \frac{|E_s|}{|E|}F(E_s)$$

where *E* is partitioned into E_M and E_S due to some split decision.

The information gain for a given split can be defined as

$$Gain(E, E_{M,}E_R) = F(E) - F(E_{M,}E_R)$$

Gain is the anticipated decrease in entropy caused by knowing the value of an attribute.

3.2 Classification

J48 Decision tree, a predictive machine-learning model, decides the target value dependent variable based on several attribute values of the available data. The internal nodes of a decision tree represent diverse attributes. The branches amid the nodes state us the possible values that these attributes can have in the observed samples, while the terminal nodes state us the final value classification of the dependent variable.

In order to classify a new item, the J48 Decision tree classifier first desires to create a decision tree based on the attribute values of the available training data. Consequently, whenever it meets a set of items, training set finds the attribute that distinguishes the several instances utmost clearly. This feature that is capable to state us most about the data instances so that we can classify them the best is said to have the highest information gain. If there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target

variable, then that branch is terminated and the target value that we have obtained is assigned.

3.3 Predicting a class

Classification denotes to predicting categorical class label and prediction refers to modeling continuous-valued functions. First, build a model then use the model to predict unknown value. A significant technique for prediction is regression. Predict data values or build generalized linear models is based on the database data. One can only predict value ranges or category distributions. The main features which effect the prediction data relevance analysis are uncertainty measurement, entropy analysis, expert judgment, etc. Linear regression $X = \beta + \alpha X$ where two parameters β and α specify the line and are to be valued by means of the data at hand. Many nonlinear functions can be converted into multiple regression using the equation $X = c_0 + c_1Y_1 + c_2Y_2$.

4. EXPERIMENTAL SETUP AND RESULTS

The Iris flower data set collected from Wikipedia is used for experiment. The experiment is done using weka tool. The Iris flower data set consists of 3 classes of 250 instances each class. One class corresponds to one species of Iris flower named Setosa, Versicolor, and Virginica. Each class has 5 attributes. It represents Sepal Length, Sepal Width, Petal Length, Petal Width, and Species. J48 is a classifier technique which used to make a decision tree. The metrics used for evaluating the experiments results are Precision and Recall, F–Measure, and Accuracy using confusion matrix.

4.1 Percentage split

In percentage split the database is arbitrarily divided into two separate datasets. The first set called training set where the data mining system tries to extract knowledge. The mined knowledge is tested against the second set which is referred test set. The objective is to obtain nodes that contain cases of single class. A function of relative frequencies of the classes in that node is given by

$$j(s) = \sigma(q_{1,}q_{2,}...q_i)$$

with q_i (i = 1, ..., j) as the relative frequencies of the j different classes in that node.

4.2 Precision and recall

Precision refers to the probability that a (arbitrarily chosen) retrieved document is relevant. In information retrieval positive predictive value is called precision. It is calculated as number of correctly classified instances belongs to X divided by number of instances classified as belonging to class X. i.e. it is the proportion of true positives out of all positive results. Recall is the probability that a (randomly selected) relevant document is retrieved in a search. This is the percentage of fault-prone modules that are correctly classified.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

4.3 F-Measure

F-measure is a technique where recall and precision are combined into a single measure of performance. F-measure can be defined as

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

4.4 Accuracy

The accuracy of clustering is calculated using confusion matrix. Table I shows that the confusion matrix to calculates the actual and predicted classification i.e. the total number of true positives for class A is 50 and the total number of false positive for class B is 49. The total number of true negative for class C is 48.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FN + TN}$$

Table-1. Confusion matrix for Iris dataset.

Class A	Class B	Class C	Classified as
50	0	0	Setosa
0	49	1	Versicolor
0	2	48	Virginica

4.5 Sensitivity and specificity

Sensitivity and specificity are the two commonly used concepts to measure the performance. These notions are freely usable for the evaluation of any binary classifier.

$$\text{Sensitivity} = TPR = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Here TP is true positive, TN is true negative, FP is false positive, and FN is false negative. TPR is true positive rate that is equivalent to Recall.

4.6 Cost/Benefit analysis

The cost of J48 for the classes Setosa, Versicolor, and Virginica are 100%, 98%, and 98% respectively. Fig. 2, 3, and 4 shows the cost of J48 for the classes Setosa, Versicolor, and Virginica, respectively.

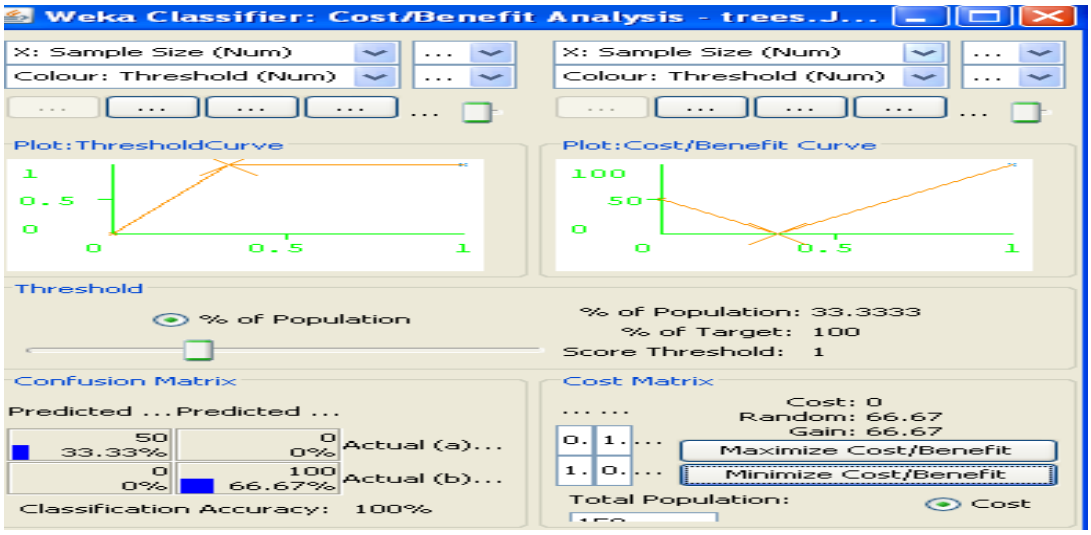

Figure-2. Cost analysis for class Setosa.

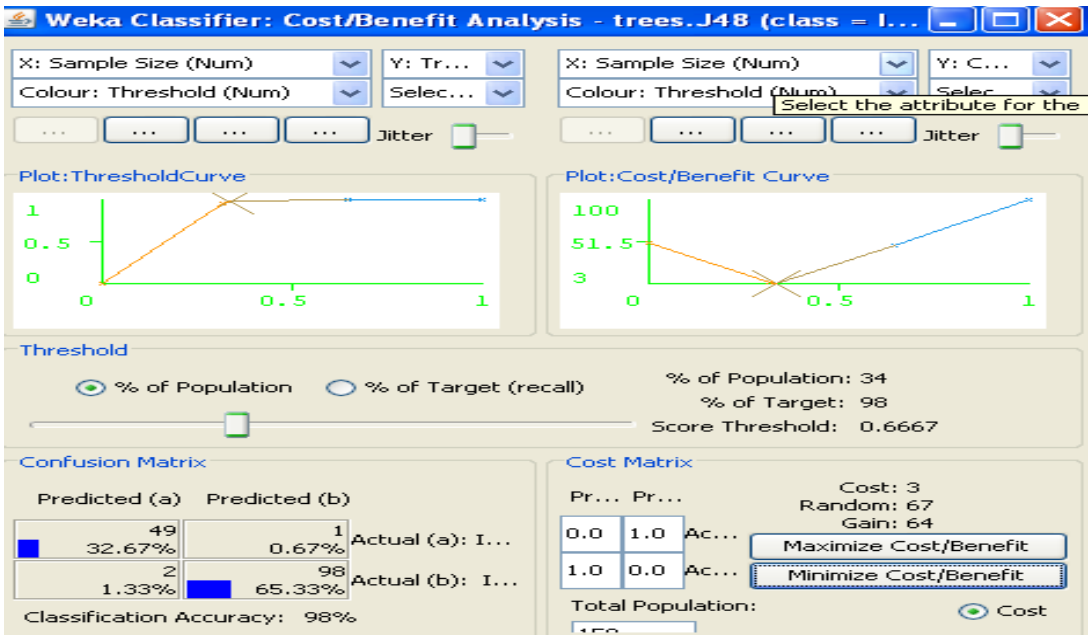

Figure-3. Cost analysis for class Versicolor.

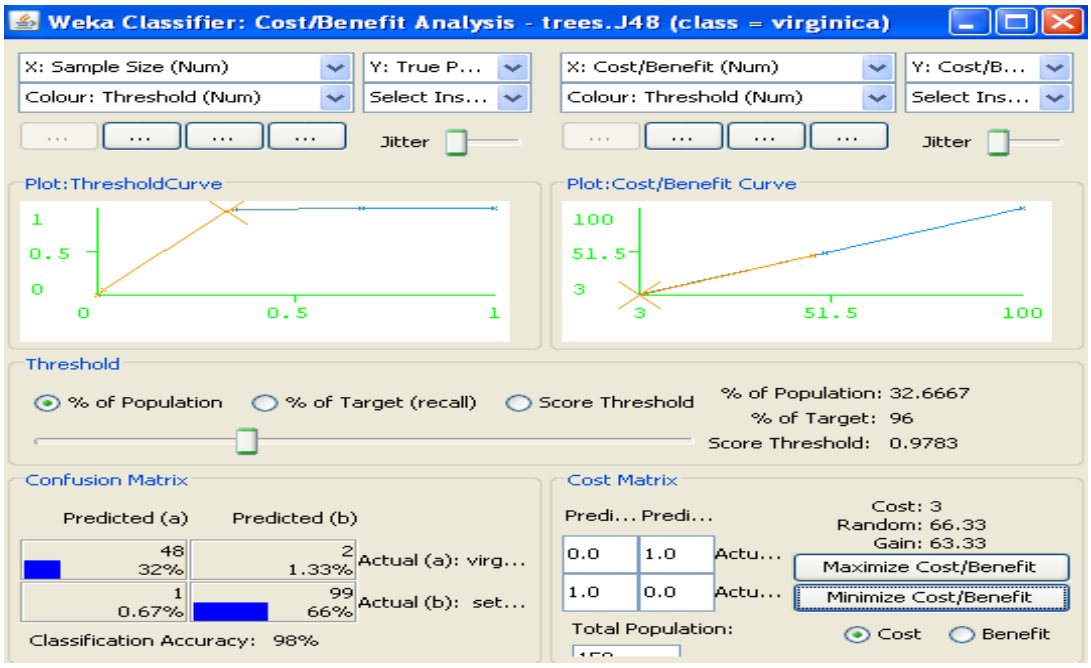

Figure-4. Cost analysis for class Virginica.

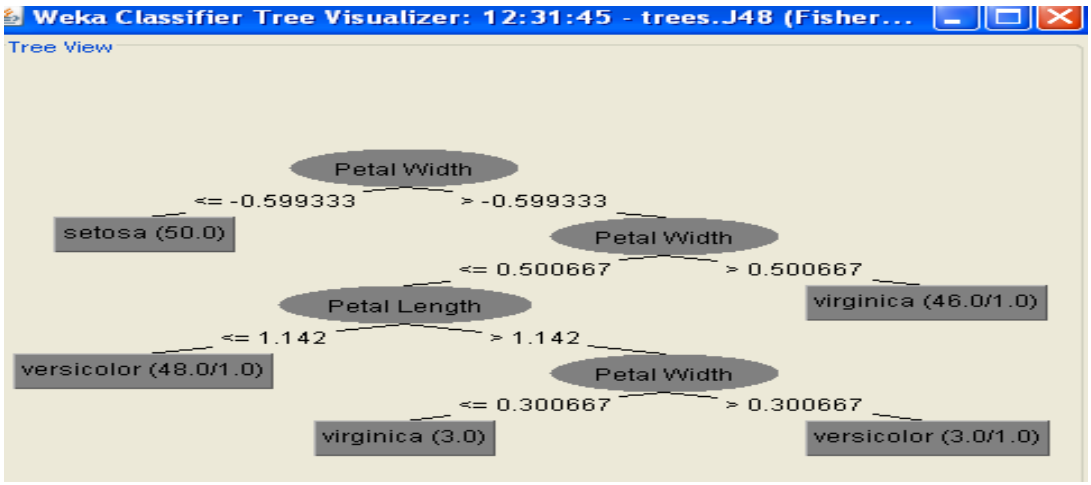

Figure-5. Decision tree using J48 during classification of Iris data.

Table-2. Accuracy of cost analysis for Iris data set.

Class	TP rate	FP rate	Precision	Recall	F-measure	ROC area
Setosa	98%	98%	96%	98%	97%	99%
Versicolor	96%	96%	98%	96%	97%	99%
Virginica	98%	98%	98%	98%	98%	99%

Table-3. Classification accuracy and cost analysis of J48.

Class	Classification accuracy	Cost analysis
Setosa	100%	100%
Versicolor	100%	98%
Virginica	100%	98%

Table-4. Comparison of ID3 and J48.

Parameters	ID3	J48
TP_Rate	0.987	0.96
FP_Rate	0.007	0.01
Precision	0.987	0.98
Recall	0.987	0.96
F-Measure	0.987	0.97
Roc_Area	0.987	0.99

Table 2, 3 and 4 shows the accuracy of cost analysis for Iris data set, classification accuracy and cost analysis of J48, and comparison of ID3 and J48. J48 gives more classification accuracy for class iris having three values Setosa, Versicolor and Virginica. The results on the datasets show that the efficiency and accuracy of J48 is better than ID3.

5. CONCLUSIONS

In this paper, J48 is employed for effective prediction analysis. Iris flower data set consists of 3 classes of 250 instances each class is collected for this experiment. In this experiment J48 classifier is used in Weka tool to make decision trees. The results of the experiments are compared with the results of the existing ID3 algorithm. The results show that J48 gives better classification accuracy than the existing ID3 algorithm.

REFERENCES

Chen Jin, Luo De-lin, Mu Fen-xiang. An Improved ID3 Decision Tree Algorithm. Proceedings of 2009 4th International Conference on Computer Science and Education, 978-1-4244-3521-0/09/$25.00 ©2009 IEEE 127.

Desouza K.C. 2001. Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands Institute for Healthcare Technology Management.

J. Han and M. Kamber. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.

Varun kumar,Nisha Rathee. 2011. Knowledge discovery from database using an integration of clustering and classification. (IJACSA) International Journal of Advanced Computer Science and Applications. 2(3).

Hem Jyotsana Parashar, Singh Vijendra and Nisha Vasudeva. 2012. An Efficient Classification Approach for Data Mining. International Journal of Machine Learning and Computing. 2(4).

J. R. Quinlan. 1996. Induction of decision trees. Machine Learning. 1(1).

Dipankar Dutta. Classification Rules Generation for Iris Data Using Lexicographic Pareto Based Multi Objective Genetic Algorithm.

Maysam Eftekhary, Peyman Gholami, Saeed Safari1 and Mohammad Shojaee2. 2012. Ranking Normalization Methods for Improving the Accuracy of SVM Algorithm by DEA Method. Modern Applied Science; 6(10); ISSN 1913-1844 E-ISSN 1913-1852 Published by Canadian Center of Science and Education.

Anshul Goyal and Rajni Mehta. 2012. Performance Comparison of Naïve Bayes and J48 Classification Algorithms. International Journal of Applied Engineering Research, ISSN 0973-4562, 7(11).

Tina R. Patil and Mrs. S. S. Sherekar. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Computer Science And Applications. 6(2), ISSN: 0974-1011.

Ravijeet Singh Chauhan. 2013. Predicting the Value of a Target Attribute Using Data Mining. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, 3(2).

Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M.Reena Pawar. 2013. Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727, 10(6): 01-06.

Margaret H. Danham, S. Sridhar. 2006. Data mining, Introductory and Advanced Topics. Pearson Education, 1st Ed.

Aman Kumar Sharma, Suruchi Sahni. 2011. A Comparative Study of Classification Algorithms for Spam Email Data Analysis. IJCSE. 3(5): 1890-1895.

Yang Y., Webb G. 2003. On Why Discretization Works for Naive-Bayes Classifiers. Lecture Notes in Computer Science. 2003: 440-452.

Raman Pathrey, Yogesh Kumar, Nitin and Nisha Rathee. 2013. Discovering Knowledge Patterns from Integration of Clustering and Classification Techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 3(4), ISSN: 2277 128X.

CHAI Rui-min, WANG Miao. A more efficient classification scheme for ID3. 978-1-4244-6349-7/10/$26.00 ©201 0 IEEE Vl-329.

Wenke Lee, Salvatore J. Stolfo, Kui W. Mok. A Data Mining Framework for Building Intrusion Detection Models.

Duda R.O., Hart P.E. 1973. Pattern Classification and Scene Analysis. In: Wiley-Interscience Publication, New York, USA.

Bishop C.M. 1999. Neural Networks for Pattern Recognition. Oxford University Press, New York, USA.

Vapnik, V.N, The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag, New York, USA.

Angeline Christobel. Y, Dr. Sivaprakasam. 2011. An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems. 3(2).

M. Kass, A. Witkin and D. Terzopoulos. 1987. Snakes: Activecontourmodels, Int,' IJ. comp. Vis. 1: 321-331.