Classification with Weighted C4.5 Decision Tree Approach

Khin Thuzar Win, Aung Nway Oo
University of Computer Studies, Yangon
khinthuzarwin87@gmail.com, aungnwayoo78@gmail.com

Abstract

Data mining techniques is increasing becoming on medical data for discovering useful trends and patterns that are used in diagnosis and decision making. Classification is a data mining technique which addresses the problem of constructing a predictive model for a class attribute given the values of other attributes and some examples of records with known class. This paper we implemented the weighted C4.5 decision tree algorithms for Breast Cancer classification. Naïve Bayesian theorem was used to calculate the weight value to set the appropriate weights of training instances before trying to construct a decision tree model. The research work focuses the predictive comparative analysis of weighted C4.5 decision tree algorithm with traditional C4.5 decision tree algorithm by using Breast Cancer Datasets.
Key words: Data mining (DM), Classification, Decision Tree (DT), C4.5

1. Introduction

The term 'data mining' is devised to refer to the action of moving through large databases investigating appealing and new patterns. Data mining has become considerably important and a necessity today when data are bountiful and easily accessible. The automatic analysis of large numbers of data is possible through the methods and instruments that the field of data mining provides. Data mining is one aspect of the process of Knowledge Discovery in Databases (KDD). Some searchers think if data mining as an ambiguous expression and uses the term "Knowledge Mining" as it bears a better resemblance to gold mining. Data mining approach are mostly grounded on inductive learning i.e., constructing a mode explicitly or implicitly by forming a generalization from enough training examples. The inductive approach forms a basic assumption that the trained model is related to future unseen examples. Specifically, any form of conjecture is considered an induction on conditions that conclusions are not logically drawn from premises.

Data collection was conventionally accepted as one pivotal period in data analysis. An analyst would be able to select the variables to be collected by the application of the available domain knowledge. The number of specified variables was usually restricted and their values could be recorded by hand or using oral interviews. If computer-aided analysis was to be used, the collected data had to be entered into statistical computer package or an electronic spreadsheet. Because the process of data collection was expensive, analysts had to learn to make decisions on available information. Decision trees are regarded as well-known methods for representing classifiers. A decision tree is a classifier viewed as the repetitive subdivision of the instance space.

The decision tree is composed of nodes forming a 'rooted tree' i.e., a 'directed tree' with a node known as 'root' with no incoming edges. There is exactly one incoming edge in all other nodes. An internal node is a node with outgoing edges. All other nodes are known as leaf node. In a decision tree, it is each internal node subdivides the instance space into two or more sub-spaces by an assured discrete function of the input attributes values. Simply and most frequently, each test takes a single attribute such that the attribute's values subdivide the instance space. On the other hand, the leaf may grip a probability vector that indicates the probability of the goal attribute having a definite value. Instance, from the root of a tree to a leaf, are navigated and organized, following the outcome of the tests along the path. There have been many decision tree algorithms like ID3 [1], C4.5 [2], CART [9] etc.

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification has been successfully applied to wide range application areas, medical diagnosis, weather prediction, credit approval, customer segmentation, fraud detection among the different proposals. Classification is clearly useful in many decision problems, where for a data item a decision is to be made (which depend on the class to which data item belongs).

The rest of the paper is organized as follows. Section 2 reviews the related work and section 3 presents the background theory and section 3.1 presents the overview C4.5 algorithm. Naïve Bayes theorem and weighted C4.5 algorithm were described in section 3.2 and 3.3. Overview of the system flow was illustrated in section 6. Description of dataset is presented in section 7.The experimental results are presented in section 8. Finally, conclude of this study was provided in section 9.

2. Related Work

There are many research works that proposed efficient decision tree for classification. Kohavi & John [8], who researcher of parameter settings of C4.5 decision trees made a result in optimal performance on a particular data set. Liu Yuxun and XieNiuniu [10], decision tree algorithm of based on attribute importance. It is suggested by solving the problem. Liu X.H 1998 [9], presented a new optimized algorithm of decision

trees. On the basis of ID3, this algorithm considered attribute selection in two levels of the decision tree and make the classification accuracy improve. Gaurav & Hitesh [11], proposed C4.5 algorithm which is improved by the use of L'Hospital Rule, this simplifies the calculation process and improves the efficiency of decision making algorithms. S.VijayaRani et.al [12] the authors analyzed the performance of C4.5, RIPPER and PART algorithm. Time and Number of rules produced were provided as the measures to analyze the data for Breast cancer data and heart disease data. Dewan Md. Faraid and Chowdhury [6] proposed the method for assigning weight value to training instances to increase the classification accuracy. In this paper, comparative studies of weighted and normal C4.5 algorithms are made to approximation the breast cancer dataset.

3. Background Theory

Classification can be used as in the form of data analysis that can be used to extract models describing important data classes. Classification can be used for making intelligent decision. In this study, weighted C4.5 algorithm was used for efficient classification. Breast cancer data set was used for testing of proposed method and compares the results of normal C4.5 algorithm.

3.1. C4.5 Algorithm

The C4.5 algorithm is the modified version of ID3 algorithm and which choose splitting attributes from a dataset with the maximum information gain.

The attribute with the maximum gain ratio is selected as root node or the splitting attribute. The expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \qquad (1)$$

Let p_i is the probability that an arbitrary tuple in D belongs to class C_i and m is the quantity in class label. The log function to the base 2 is used, because the information is encoded in bits. The information is based on the proportions of tuples of each class.

Information needed (after using attribute A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2)$$

where $Info_A(D)$ is the expected information of each attribute in data D and v is types of data in that attribute. Information gained by branching on attribute A.

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

In other words, *Gain(A)* tells how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|}) (4)$$

where SplitInfo(A) is the expected split information of each attribute in data D and v is types of data in that attribute. The attribute that yields the largest Gain Ratio is chosen for the decision node. The attribute with the maximum gain ratio is selected as the splitting attribute

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \qquad (5)$$

Gain Ratio for each attribute may be computed by equation 5. The attribute that yields the largest Gain Ratio is chosen for decision node. For building decision trees of C4.5 algorithm [14]:

Algorithm: Generate Decision Tree by C4.5
Input: Dataset D, attribute_list
Output: Tree
Begin
    Check for the base cases.
    For each attribute *a* in attribute_list
        Find the normalized information gain from splitting on *a*
    a_best be the attribute with the maximum normalized information gain.
Create a decision *node* that splits on *a_best*.
Recur on the sublists gained by splitting on *a_best*, and add those nodes as children of *node*.
End

3.2. Naïve Bayes Theorem

Naïve Bayesian (NB) classifier is a simple probabilistic classifier based on probability model, which can be trained very competently in a supervised learning [3-4]. The naïve Bayesian classifier, or simple Bayesian classifier [5], works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, illustrating n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \quad for \ 1 \le j \le m, j \ne i. \qquad (6)$$

Thus we maximize $P(C_i|X)$.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (7)$$

where $P(C_i|X)$ is the posterior probability or the probability that the value, $P(C_i)$ is the probability class based on the hypothesis, $P(X|C_i)$ is the predictor probability based on the given class. $P(X)$ is a predictor probability.

3. As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ necessity be highest. If the class prior probabilities are not known, then it is commonly presumed that the classes are equally likely, that is, $P(C_1) = P(C2) = \ldots = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$W_i = argmax\ P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (8)$$

The naïve Bayesian classifier is sample to use and efficient to learn. It requires only one scan of the training data. Despite the fact that the independence assumption is often violated in practice, naïve Bayes often competes well with more sophisticated classifiers. In other words, the predicted class label is the class Ci for which P(X|Ci) P(Ci) is the maximum [4].Weight value for each attribute is calculated by equation 8 which is the maximum weight value.

3.3. Weighted C4.5 Algorithm

Weighted decision tree learning algorithm was developed by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training instances are calculated using maximum posteriori hypothesis of Naïve Bayesian theorem. Weight of each training instance is calculated with the maximum value of the class conditional probabilities.

Weighted C4.5algorithm chooses splitting attributes from a dataset with the maximum information gain by using these weights value and constructs the decision tree model for Breast Cancer classification. Given a training dataset, the weighted C4.5 algorithm initializes the weights of each training instance, *Wi* by maximum likelihood of posterior probability by assigning weights of training dataset in D. This algorithm uses the weight value calculated from Naïve Bayes probabilistic model to initialize the weights of each training instance.

The expected information needed to classify a tuple in dataset D is calculated by applying equation (1). In this case, p_i is the relative frequency of class *i* in *D*, *where* , p_i is the probability that an arbitrary tuple in *D* belongs to class C_i and m is the quantity in class label. The log function to the base 2 is used, because the information is encoded in bits. The information is based on the proportions of tuples of each class. The sum is computed over *m* classes.

To determine the information required to classify D, we examine all the possible subsets that can be formed using known values of attribute A. When considering a split, we calculate a weighted sum of the impurity of each resulting partition. And then $Info_A(D)$ is calculated by applying equation (2). In this time, the value of equation (2) is defined as follows:

$|D_j|$ = the set of tuple with weight value in training dataset that have outcome aj of attribute,

$|D|$ = total weight value tuple

Information gain is defined as the dissimilarity between the original information requirement (i.e., based on just the proportion of classed) and the new requirement (i.e., gained after partitioning on A) by using equation (3) and gain ratio to overcome the problem by using equation (4) and equation (5). We are calculated InfoA(D), Gain(A) ,SplitInfoA(D) and GainRatio to assign weight value.

The decision tree is built established on the weights of training data which results from naïve Bayes probabilities.

4. System Flow of Proposed System

The system flow for classification of breast cancer dataset with weighted C4.5 algorithm was described in the following figure.

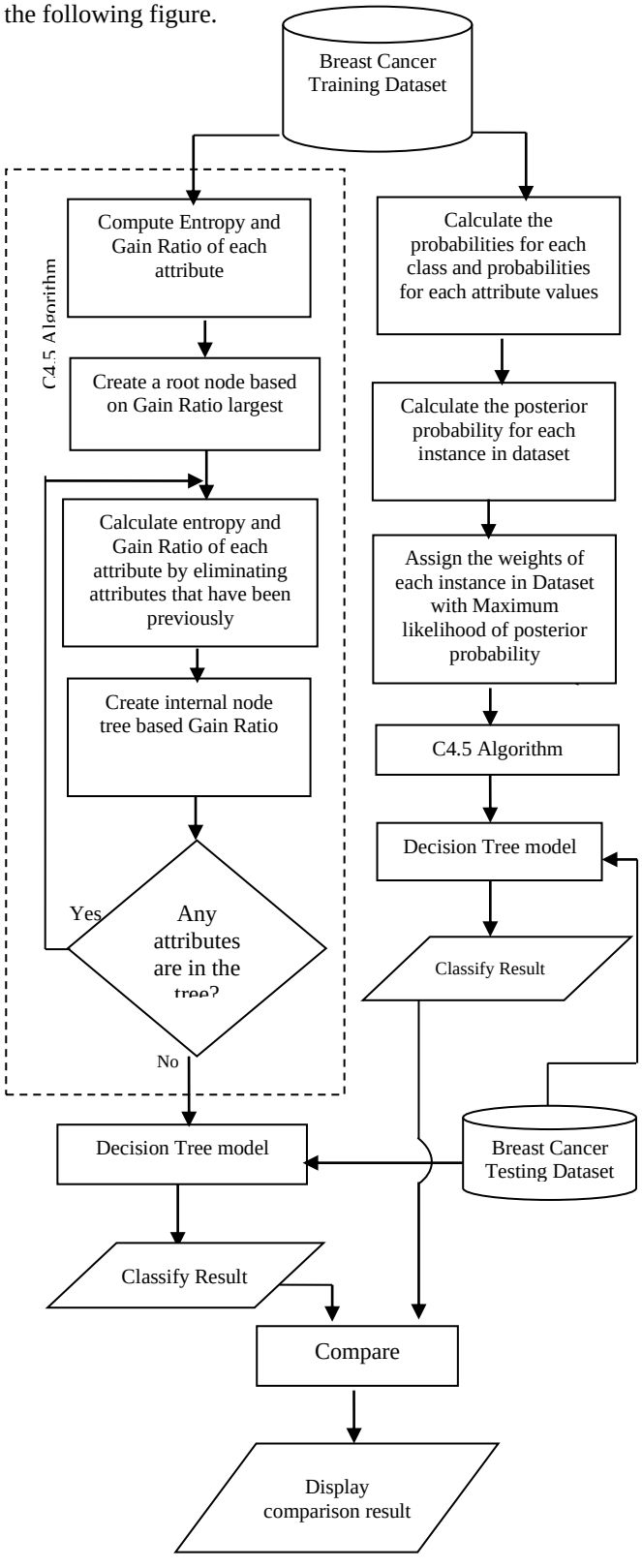

Figure 1. Overview of the proposed system

5. Dataset Description

The breast cancer dataset contains 683 instances and 10 attributes. Each of the characteristics is assigned a value from 1 to 10 by the pathologist. The larger the value of attribute the greater the likelihood of malignancy.

The following table lists the attribute information of breast cancer dataset.

Table 1. Dataset Description

ID	Attribute Name	Value
A1	Clump Thickness	1 – 10
A2	Uniformity of Cell Size	1 – 10
A3	Uniformity of Cell Shape	1 – 10
A4	Marginal Adhesion	1 – 10
A5	Single Epithelial Cell Size	1 – 10
A6	Bare Nuclei	1 – 10
A7	Bland Chromatin	1 – 10
A8	Normal Nucleoli	1 – 10
A9	Mitoses	1 – 10
A10	Class	Benign(C1), or malignant(C2)

There are two types of classes in dataset, benign (It does not invade nearby tissue or spread to other parts of the body), or malignant (It is serious and likely to spread other parts of the body).

The following figure described the sample decision tree of Breast Cancer detection. The figure is illustrated by using attribute id.

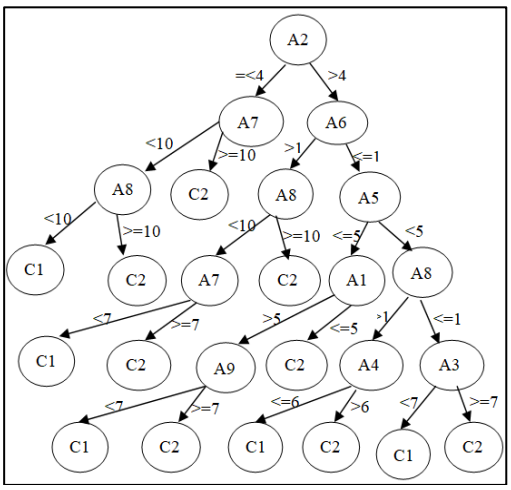

Figure 2. Sample decision tree for breast cancer classification

6. Experimental Results

The main aim of this research is to analyze weighted C4.5 decision tree and traditional C4.5 decision tree algorithm. The breast cancer dataset from UCI [7] is used for comparative analysis. For each classifier, 2/3 of the dataset is used for training and 1/3 of datasets is used for testing. The following table compares the accuracy and performance results of two classifiers by using confusion matrix and Biometric evaluation system.

The performance of the classifiers in detecting the breast cancer can be evaluated from the analysis of confusion matrix and below parameters are calculated.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$F-measure = \frac{2\times Recall\times Precision}{Precision+Recall} \quad (12)$$

Biometric evaluation system that assigns all authentication attempts a 'score' between closed interval [0, 1]. 0 means no match at all and 1 means a full match.

False Acceptance Rate (FAR) is calculated as a fraction of negative scores exceeding your threshold.

$$FAR = \frac{FP}{(FP+TN)} \quad (13)$$

False Rejection Rate (FRR) is calculated as a fraction of positive score falling below your threshold.

$$FRR = \frac{FN}{(TP+FN)} \quad (14)$$

Table 2. Comparison Accuracy for Breast Cancer Dataset

Data Record	100	200	400	683
C4.5	84.84 %	92.42 %	93.98 %	96.03 %
Weighted C4.5	84.84 %	92.42 %	93.98 %	96.03 %

Table 3. Confusion matrix of Precision, Recall and F-Measure of C4.5 Algorithm classification result

Data Record	100	200	400	683
Precision	0.78	0.79	0.91	0.96
Recall	0.73	0.85	0.92	0.96

F-Measure	0.82	0.87	0.92	0.94

Table 4. Confusion matrix of Precision, Recall and F-Measure of Weighted C4.5 Algorithm classification result

Data Record	100	200	400	683
Precision	1	1	1	1
Recall	0.93	1	1	1
F-Measure	0.96	0.98	1	1

Table 5. Biometric evaluation for C4.5 Algorithm classification result

Data Record	100	200	400	683
FAR	0.095	0.136	0.045	0.019
FRR	0.85	0.041	0.27	0.15

Table 6. Biometric evaluation for Weighted C4.5 Algorithm classification result

Data Record	100	200	400	683
FAR	0	0	0	0
FRR	0	0	0	0

Table 7. Comparison of evaluation time complexity (seconds)

Data Record	100	200	400	683
C4.5 Algorithm	0.467	0.827	1.453	2.905
Weighted C4.5 Algorithm	0.797	1.248	2.921	3.935

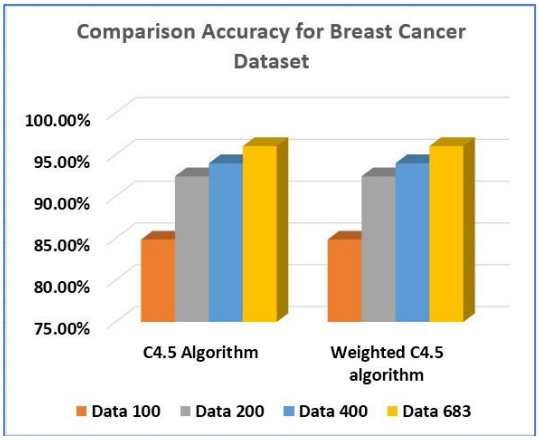

Figure 3. Comparison of classification accuracy

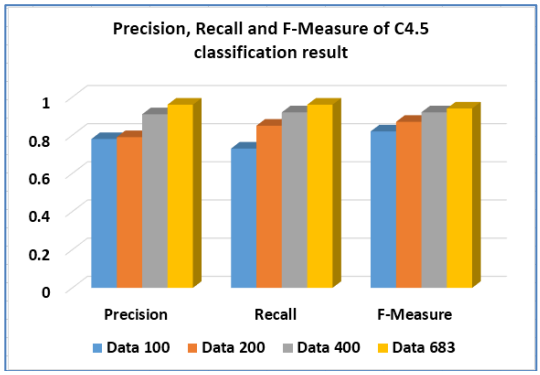

Figure 4. Confusion matrix of C4.5 Algorithm Classification result

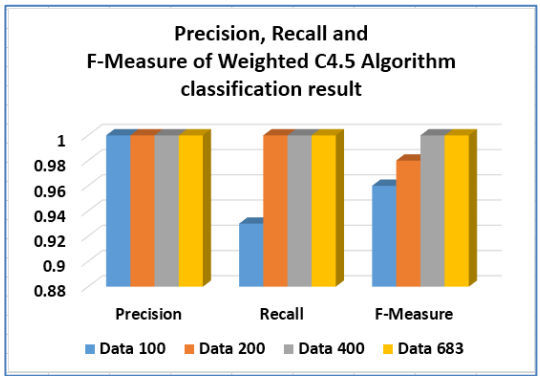

Figure 5. Confusion matrix of Weighted C4.5 Algorithm Classification result

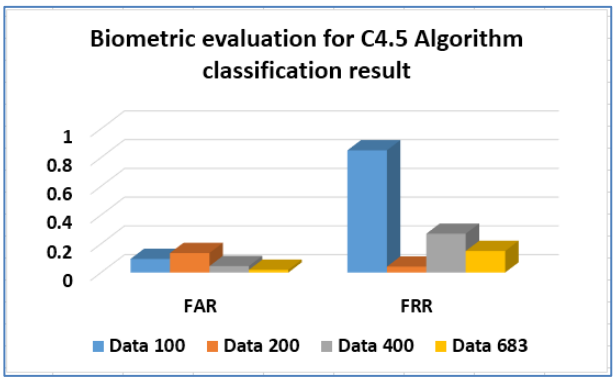

Figure 6. Biometric evaluation for C4.5 Algorithm classification result

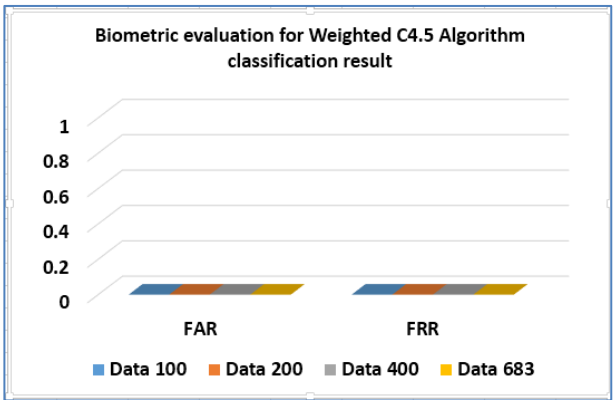

Figure 7. Biometric evaluation for Weighted C4.5 Algorithm classification result

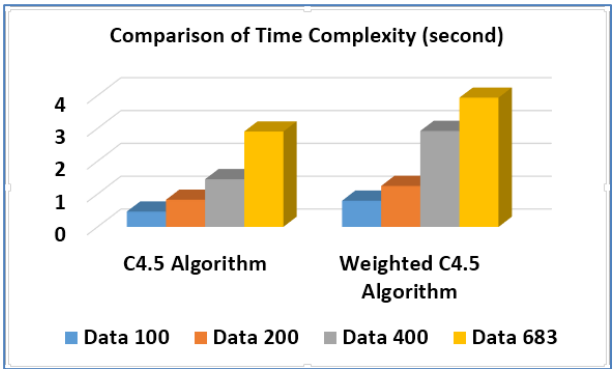

Figure 8. Comparison of evaluation time complexity

7. Conclusion

In this paper, the comparative analysis of C4.5 and weighted C4.5 algorithms classification on Breast Cancer classification was presented. From this study it is found that accuracy of weighted C4.5 algorithm is better than traditional C4.5 algorithm. The time complexity of weighted C4.5 algorithm is also slower than C4.5 algorithm. The experimental results proved that the weighted C4.5 algorithm can achieve high classification rate with better performance.

References

[1]. J. R. Quinlan, "Induction of Decision Tree," Machine Learning Vol. 1, 1986, pp. 81-106.

[2]. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Mateo,CA, 1993.

[3]. Kononenko I, "Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition," in Wieling, B. (Ed), Current trend in knowledge acquisition, Amsterdam, IOS press. 1990.

[4]. Langely, P., Iba, W., Thomas, and K., "An analysis of Bayesian classifier," in Proceedings of the 10th national Conference on Artificial Intelligence (San Matro, CA: AAAI press), 1992, pp. 223-228.

[5]. Han, Jiawei and Kamber, Micheline "Data Mining Concepts and Techniques" 2nd ed., Morgan Kaufmann Publishers, San Franciso, CA, 2007 ISBN 1-55860-901-3.

[6]. Dr. Dewan Md. Farid1 and Prof. Dr. Chowdhury Mofizur Rahman2 "ASSIGNING WEIGHTS TO TRAINING INSTANCES INCREASES CLASSIFICATION ACCURACY" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013.

[7]. UCI Machine Learning Repository: "Breast Cancer Wisconsin (Original) Data Set", Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, is consin, USA , Donor: Olvi Mangasarian (mangasarian '@' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)

[8]. Ron Kohavi & George H. John, "Automatic Parameter Selection by Minimizing Estimated Error". In Proceedings of the Twelfth International Conference, Morgan Kaufmann Publishers, San Franciso, CA.

[9]. Weiguo Yi, Jing Duan, &Mingyu Lu, "Optimization of Decision Tree Based on Variable Precision Rough Set", International Conference on Artificial Intelligence and Computational Intelligence, 2011.

[10]. Liu Yuxun, &XieNiuniu, "Improved ID3 Algorithm", IEEE, 2010.

[11]. Gaurav L. Agrawal, & Prof. Hitesh Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013.

[12]. Vijayarani, S. and Divya, M. (2011) "An Efficient Algorithm for Generating Classification Rules", International Journal of Computer Science and Technology,Vol.2,Issue 4,.

[13]. Rokach & Maimon "Data Mining with Decision Tree Theory and Applications" 2nd Edition ,2014.

[14]. https://en.wikipedia.org/wiki/C4.5_algorithm.