Breast Cancer Classification with Weighted CART Decision Tree Approach

Khin Thuzar Win
Department of Information Technology Supporting and Maintenance
University of Computer Studies, Hinthada
khinthuzarwin87@gmail.com

Aung Nway Oo
Faculty of Computer Science
University of Information Technology
aungnwayoo78@gmail.com

Abstract: Data mining (DM) is a process of inferring knowledge from large data. Classification is a major technique in data mining and widely used in various fields. There are many classification algorithms. The Decision Tree (DT) approach is most useful in classification problem. The method of Decision tree generally used for the classification because it is the hierarchical structure for the user understanding and decision making. This paper we proposed the weighted CART decision tree algorithm 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 CART decision tree algorithm with traditional CART decision tree algorithm.

Keywords: Data mining (DM), Classification, Decision Tree (DT), CART

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. Concerning numeric attributes the condition deals with a range. Each leaf is allocated to one class which indicates most appropriate target value. On the other hand, the leaf may hold a probability vector that indicates the probability of the target 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 can be used as in the form of data analysis that can be used to extract models describing important data classes. In this study, weighted CART algorithm was used for efficient classification. Breast cancer data set was used for

testing of proposed method and compares the results of normal CART algorithm.

The rest of the paper is organized as follows. Section 2 reviews the related work and section 3 presents the overview CART algorithm. Naïve Bayes theorem and weighted CART algorithm were described in section 4 and 5. 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, conclusion of this study was provided in section 9.

2. RELATED WORK

There are many research works that proposed efficient decision tree for classification. B.Padmapriya & T.Velmurugan [11], searched for accuracy of classification techniques is evaluated based on the selected attributes of mammogram images with CART algorithm. Yamuna and Venkatesan [12] proposed the kidney transplant survival rate prediction with decision trees and comparative analysis is performed by using C4.5 and CART decision tree algorithms. E. Venkatesan* and T. Velmurugan [13] researched the performance analysis of decision trees algorithms for breast cancer classification. They used the j48, AD tree and CART algorithms and then compared and evaluate the results of different classifiers. In the paper of Hyontai Sug [14] Comparison of Decision Tree Algorithms for Medical Data Sets were performed using the C4.5 and CART. To evaluate the algorithm, used the 17 medical datasets and 10 fold cross validation was performed and compare the accuracy of algorithms. According to literature survey C4.5 has been used in some wide range of areas [15] like financial areas [16] and engineering areas [17], but CART has been favored mostly in medicine domain, because most researchers in medicine domain reported good performance of CART in their data mining tasks [18].

In this paper, comparative studies of weighted and normal CART algorithms are made to predict the breast cancer dataset.

3. CART ALGORITHM

CART is capable of handling both numerical and categorical variables. Gini index measures how well a given attribute separates training samples into targeted class. Here binary splitting of attributes takes place. It is most widely used statistical procedure. It provides a bi of binary decision [10].The Gini index is CART. If a data set D contains examples classes, gini index, $gini(D)$ is defined as

$$gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

where p_i is the relative frequency of class *where* $p_i = |C_i D| / |D|$, $|C_i D|$ = total tuple fo $|D|$ = total tuple. The sum is computed classes.

The Gini index considers a binary split fo attribute. Let's first consider the case attribute A is a discrete-valued attribute ha distinct values, $\{a_1, a_2, \ldots, a_v\}$, occurring Each subset, $SubA$, can be considered as a test for attribute A of the form "$A \in SubA$" a tuple, this test is satisfied if the value of A tuple is among the values listed in $SubA$. If possible values, then there are 2^v possible s Therefore, there are $2v$-2 possible ways t two partitions of the data, D, based on a split on A.

If a data set D is split on A into two sub and D_2, the *gini* index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

where $|D_1|$ = the set of tuples in D satisfyi split-point, $|D_2|$ = the set of tuples in D sat A > split-point and $|D|$ = total tuple.

Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

The attribute provides the smallest $gini_{split}$ the largest reduction in impurity) is chosen the node (*need to enumerate all the p splitting points for each attribute*) [5].

Decision trees are formed by a collection o based on variables in the modeling data set

1. Rules based on variables' valu selected to get the best split to differ observations based on the dep variable
2. Once a rule is selected and splits into two, the same process is app each "child" node (i.e. it is a re procedure)
3. Splitting stops when CART dete further gain can be made, or some stopping rules are met. (Alternative

data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

4. NAVIE BAYES THEROEM

Naïve Bayesian (NB) classifier is a simple probabilistic classifier based on probability model, which can be trained very efficiently 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, $X = (x_1, x_2, \ldots, x_n)$, depicting 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

$$P(C_i|X) > P(C_j|X) \quad for \ 1 \le j \le m, j \ne i. \tag{4}$$

Thus we maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the *maximum posteriori hypothesis*.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{5}$$

3. As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed 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):

$$\text{arg max } P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{6}$$

We can easily estimate the probabilities $P(x_1|C_i)$, $P(x_2|Ci)$, ... , $P(x_n|Ci)$ from the training tuples. Weight value for each attribute is calculated by equation (6) which is the maximum weight value.

5. WEIGHTED CART 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 CART algorithm calculates the gini index by using these weights and builds the decision tree model for classification. Given a training dataset, the weighted CART algorithm initializes the weights of each training instance, Wi by highest posterior probability for that training instance. The algorithm uses the weight value calculated from Naïve Bayes probabilistic model to initialize the weights of each training instance.

The Gini index of dataset D is calculated by applying equation (1). In this case, p_i is the relative frequency of class i in D, *where* , $p_i = \sum Wi / \sum_{j=1}^{n} |Wj|$, W_i = weight for Class C_i and W_j = weight for tuple j. The sum is computed over m classes.

To determine the best binary split on attribute A, we examine all the possible subsets that can be formed using known values of attribute A. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. Dataset D is partitions into D1 and D2 depends on the value attribute A. And then $gini_A(D)$ is calculated by applying equation (2). In this time, the value of equation (2) is defined as follows:

- $|D_1|$ = the set of tuples with weight value in D satisfying A ≤ split-point
- $|D_2|$ = the set of tuples with weight value in D satisfying A > split-point
- $|D|$ = total weight value tuple

For each attribute, each of the possible binary splits is considered. For a discrete-valued attribute, the subset that gives the minimum gini index for that attribute is selected as its splitting subset.

The reduction in impurity that would be incurred by a binary split on a discrete-value attribute A is by applying equation (3). The attribute that provides the minimum gini index is selected to split the node. The decision tree is constructed based on the weights of training data which results from naïve Bayes probabilities.

6. SYSTEM FLOW OF PROPOSED SYSTEM

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

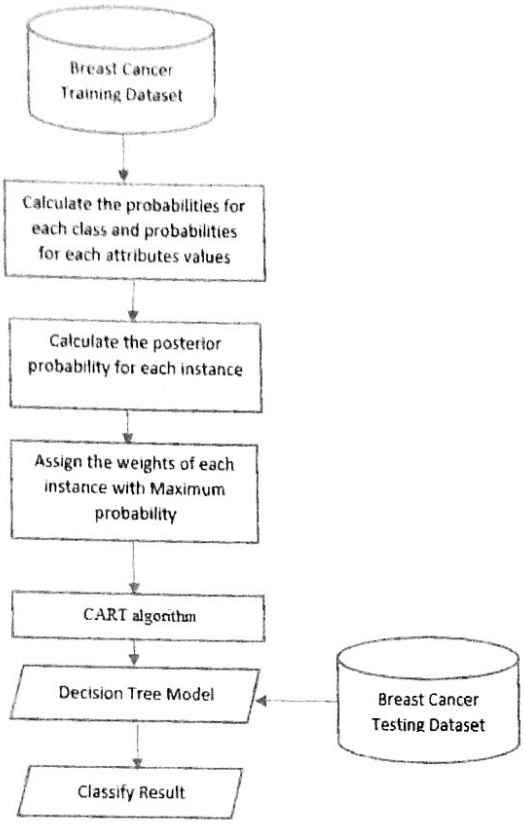

Figure 1. Overview of the proposed system

7. 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.

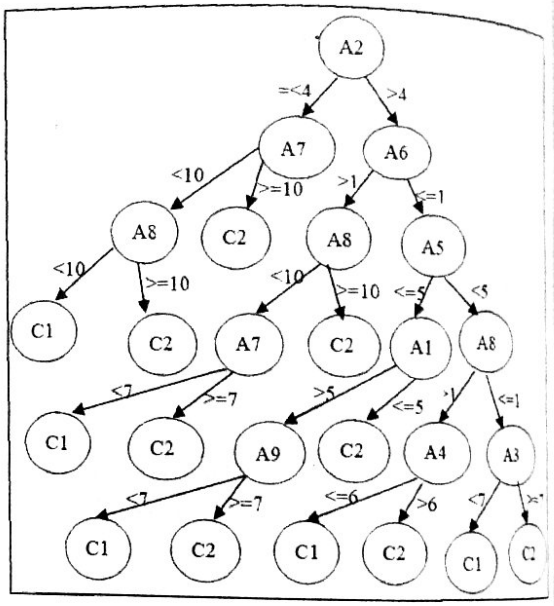

Figure 2. Sample decision tree for breast cancer classification

8. EXPERIMENTAL RESULTS

The experimental results of classifiers are discussed in this section. The main aim of this research is to analyze weighted CART decision tree and traditional CART 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 dataset is used for testing.

The following formula will calculate the accuracy which is the proportion of the total number of predictions that were correct.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

The following table compares the accuracy results of two classifiers.

Table 2. Comparison of classification accuracy

Data Record	CART algorithm	Weighted CART algorithm
100	85%	90%
200	90%	92.5%
400	92.5%	93.75%
683	94.16%	96%

The following figure visualizes the accuracy results of classifiers.

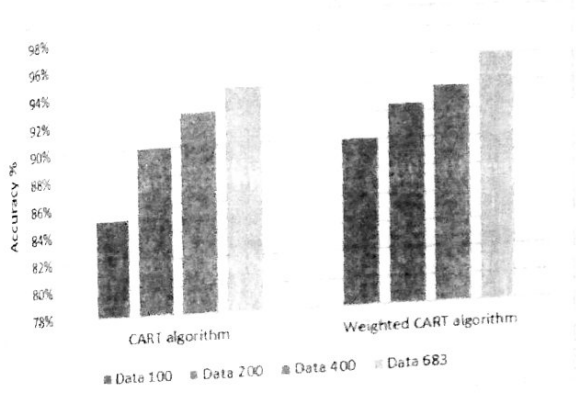

Figure 3. Comparison of classification accuracy

9. CONCLUSIONS

In this paper, the comparative analysis of CART and weighted CART algorithms classification on Breast Cancer classification was presented. From this study it is found that accuracy of weighted CART algorithm is better than traditional CART algorithm. The experimental results proved that the CART algorithm with weight value is more suitable for prediction of breast cancer dataset.

10. ACKNOWLEDGMENTS

I would to express my gratitude to University of Computer Studies Yangon for allow me to do this research work. Thanks to Dr. Aung Kway Oo for discussions that helped clarify our ideas and his support and encouragement. The author carried out this work while at Technological University (Kyaukse) and the Institute for the Study of Learning and Expertise.

11. 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 '(a)' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)

[8]. https://en.wikipedia.org/wiki/CART_algorithm

[9]. L. Breiman, J. H. Friedman, R. A. Olshen and C.J. Stone, "Classification and Regression Trees,"Statistics probability series, Wadsworth, Belmont, 1984.

[10]. E.Venkatesan and T.Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classificaiton", Indian Journal of Science and Technology, Val 8(29), November 2015.

[11]. B.Padmapriya and T.Velmurugan "Classification Algorithm Bayes Analysis of Breast Cancer Data", Internation Journal of Data Mining Techniques and

Applications Volume 5, Issus 1, June 2016, Page No.43-49, ISSN: 2278-2419.

[12]. A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival, International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 5, No. 3, March-April 2014, Page No. 225-229.

[13]. E. Venkatesan* and T. Velmurugan , Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, Indian Journal of Science and Technology, ISSN (Online) : 0974-5645, Vol 8(29), Page No. 1-8.

[14]. Hyontai Sug, Performance Comparison of Decision Tree Algorithms for Medical Data Sets, International Journal of Mathematics and Computers in Simulation, ISSN No: 1998-0159, Volume 8, 2014, Page No. 107-115.

[15]. S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, 2007, Page No. 249-268.

[16]. Z. Chang, "The application of C4.5 algorithm based on SMOTE in financial distress prediction model," in Proceedings of 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, Page No. 5852-5855.

[17]. S. Gao, "The Analysis and Application of the C4.5 Algorithm in Decision Tree Technology," Advanced Materials Research, vol. 457-458, 2012, Page No.754-757

[18]. R.J. Lewis, An Introduction to Classification and Regression Tree (CART) Analysis, Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.