CLASSIFICATION OF C4.5 AND CART ALGORITHMS USING DECISION TREE METHOD

Khin Lay Myint[1], Aye Aye Cho[2], Aye Mon Win[3]

[1] *Lecturer, Faculty of Information Science, University of Computer Studies, Hinthada, Myanmar*
[2] *Associate Professor, Faculty of Computer Science, University of Computer Studies, Hinthada, Myanmar*
[3] *Lecturer, Faculty of Information Science, University of Computer Studies, Hinthada, Myanmar*

ABSTRACT

In this paper, we proposed the traditional decision tree algorithm and weighted decision tree algorithm. Traditional decision tree algorithm consists of C4.5 and CART algorithms. The weighted decision tree algorithm is to set appropriate weights of training instances based on naïve Bayesian theorem before trying to construct a decision tree model. We compare the proposed weighted decision tree algorithm with traditional C4.5 and CART algorithms. In this paper, traditional decision tree and weighted decision tree algorithms are compared results from both training and testing dataset for heart disease.

Keyword: - *Data mining, classification algorithms, decision tree, patient database*

1. INTRODUCTION

The decision tree learning is most powerful and popular decision support tools of machine learning in classification problems which is used in many real world applications like: medical diagnosis, radar signal classification, weather prediction, etc. Decision tree is simple to understand and can deal with huge volume of dataset, because the tree size is independent of the dataset size. Decision tree model can be combined with other machine learning models. Decision tree can be constructed from dataset with many attributes and each attribute having many attribute values. Once the decision tree construction is completed, it can used to classify seen or unseen training instances. Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends whose class label is unknown. Classification can be used for making intelligent decision. Classification is clearly useful in many decision problems, where for a given data item a decision is to be made (which depend on the class to which the data item belongs).

2. BACKGROUND THEORY

The C4.5 algorithm is the upgraded version of ID3 decision tree learning algorithm. CART (Classification and Regression Trees) is a process of generating a binary tree, which can handle missing data and contain pruning strategy. C4.5 algorithm finds the best splitting attribute with highest information gain value using the weights of training instances in training dataset to form a decision tree. CART algorithm finds the *gini*(*D*) using the weights of training instances in training dataset to form a decision tree. We proposed a new decision tree learning algorithm by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training instances are calculated using naïve Bayesian theorem. The C4.5 and CART algorithms are calculated to assign weight values. There have been many decision tree algorithms. We are used the following algorithms

- C4.5(a successor of ID3)
- Classification and Regression Trees (CART)
- Naïve Bayes theorem

2.1. C4.5 Algorithm

Select the attribute with the highest information gain. Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$. Expected information (entropy) needed to classify a tuple in D:

For Traditional C4.5

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

where $p_i = |C_iD| / |D|$
$|C_iD|$ = total tuple for C_i,
$|D|$ = total tuple

For Weighted C4.5

where $p_i = \sum Wi / \sum_{j=1}^{n} |Wj|$
W_i = weight for Class C_i,
W_j = weight for tuple j

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

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

For Traditional C4.5
where $|D_j|$ = total tuples in D that have outcome a_j of A,
$|D|$ = total tuple

For Weighted C4.5
where $|D_j|$ = total weight tuples in D that have outcome a_j of A,
$|D|$ = total weight tuple

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Information gain measure is biased towards attributes with a large number of values. C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain).

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

GainRatio(A) = Gain(A)/SplitInfo(A)
The attribute with the maximum gain ratio is selected as the splitting attribute

2.2. CART Algorithm

If a data set *D* contains examples from *m* classes, gini index, *gini*(*D*) is defined as, where p_i is the relative frequency of class *i* in *D*

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

For Traditional CART
where $p_i = |C_iD| / |D|$
$|C_iD|$ = total tuple for C_i,
$|D|$ = total tuple

For Weighted CART
where $p_i = \sum Wi / \sum_{j=1}^{n} |Wj|$

W_i = weight for Class C_i,
W_j = weight for tuple j

If a data set *D* is split on A into two subsets D_1 and D_2, the *gini* index *gini*(*D*) is defined as:

For Traditional C4.5

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

where

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

For Weighted C4.5

where

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

Reduction in Impurity:

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

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

2.3. Naive Bayesian Theorem

Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x_1, x_2, …, x_n). Suppose there are *m* classes C_1, C_2, …, C_m. Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$. This can be derived from Bayes' theorem:

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

Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized. A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$\equiv \arg\max P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

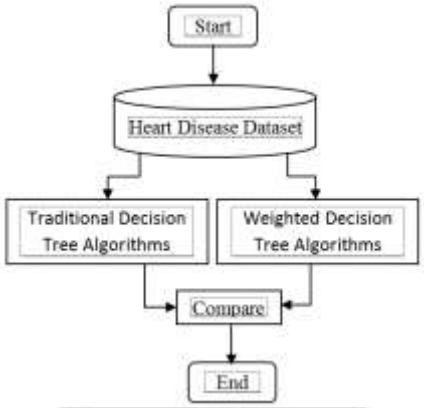

Fig - 1: Over View of System Flow Diagram

3. ABOUT DATASET

3.1 Class-Labeled Training Tuples from the Heart Disease Database

Table - 1: Heart Disease

Age	ST by exercise	exerc ind angig	blood sugar	Max HR	vessles cc	Cholestero	Rest SBP	neter narro	Class
7	6	3	2	5	10	7	4	6	defect
3	1	1	1	2	?	3	1	1	normal
3	1	1	1	2	1	3	1	1	normal
5	4	6	10	2	10	4	1	1	defect
1	1	1	1	2	1	3	1	1	normal
3	2	2	1	2	1	2	3	1	normal
10	1	1	1	2	10	5	4	1	normal
1	1	1	1	2	1	2	1	1	normal
8	10	3	2	2	4	3	10	1	defect
10	4	6	4	2	10	7	1	1	defect
10	4	7	2	2	8	6	1	1	defect
5	1	1	1	2	1	3	1	2	normal
5	2	2	2	2	1	2	2	1	normal
5	4	6	6	4	10	4	3	1	defect
8	6	7	3	3	10	3	4	2	defect
1	1	1	1	2	1	1	1	1	normal
6	5	5	8	4	10	3	4	1	defect
1	1	1	1	2	1	3	1	1	normal
1	1	1	1	1	1	2	1	1	normal
8	5	5	5	1	10	4	3	1	normal
10	3	3	1	2	10	7	6	1	defect
1	1	1	1	2	1	3	1	1	normal
2	1	1	1	2	1	1	1	1	normal

3.1 Experimental Results

Data Record	C4.5	CART	Weighted C4.5	Weight CART
100	80%	85%	85%	90%
200	87.5%	90%	90%	92.5%
400	91.25%	92.5%	92.5%	93.75%
600	92.5%	92.5%	94.16%	95%

Fig - 2 : Heart Disease Classification of Accuracy Result

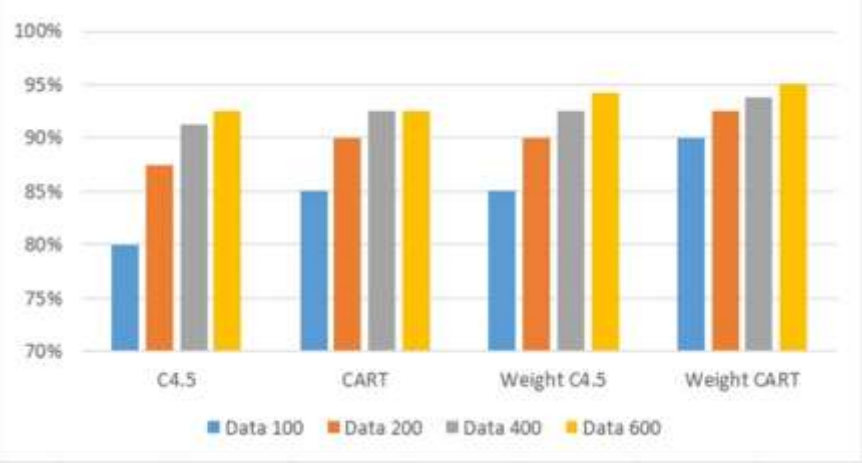

Chart -1 : Heart Disease Classification of Accuracy Result

Data Record	C4.5	CART	Weighted C4.5	Weight CART
100	.375s	.188s	.078s	1.39s
200	.672s	.351s	.23s	2.182s
400	2.183s	.811s	.3s	4.555s
600	3.055s	1.062s	0.521s	7.521s

Fig - 3: Heart Disease of Classification Time (seconds)

4. CONCLUSION

This paper presents comparison of traditional decision tree algorithms and weighted decision tree algorithms classification on Heart Disease classification problems. The experimental results proved that the weighted decision tree algorithm can achieve high classification rate on Heart Disease dataset. The time complexity of weighted decision tree is slower than conventional decision tree. The weighted CART decision tree algorithm is more accurate than C4.5, C4.5 is faster than CART.

5. REFERENCES

[1] Shweta Kharya Bhilai Institute of Technology, Durg C.G. India, Sunita Soni Bhilai Institute of technology, Durg C.G. India :" Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection ", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016

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

[3] Classification the Stages of Dental Caries using C4.5 Decision Tree Algorithm, May Thet Mon, University of Computer Studies, Yangon.

[4] ICU Patients Risk Level Classification System using CART Algorithm, Swe Hlaing, University of Computer Studies, Yangon.

[5] DATA MINING Concepts and Techniques, Jiawei Han | Micheline Kamber | Jian Pei, 2006.

[6] Comparative Study of Decision Tree Algorithms: ID3 and CART, Su Myat Thu, University of Computer Studies, Yangon.