Analyzing the Comparison of C4.5, CART and C5.0 Algorithms on Heart Disease Dataset using Decision Tree Method

Khin Lay Myint[1], Hlaing Htake Khaung Tin[2]
[1]Faculty of Information Science,University of Computer Studies, Hinthada, Myanmar,
khinlaymyint.cu@gmail.com
[2]Faculty of Information Science,University of Computer Studies, Hinthada, Myanmar,
hlainghtakekhaungtin@gmail.com

Abstract. Data acquisition methods can be expected for patients suffering from heart disease. The resolution of this learning was to compare a similar data mining algorithm to the calculation of heart diseaseThis research paper proposed the traditional decision tree procedure and weighted decision tree procedure. Traditional decision tree process consists of C4.5, C5.0, CART processes. The weighted decision process is established suitable weights of training cases based on naïve Bayesian theorem before trying to construct a decision tree model. The main objectives of this research paper are (1) to know the operation of C4.5 process, Cart and C5.0 process, (2) to learn how to analysis the 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, C4.5, Cart, C5.0.

1 Introduction

Searching for information is about collecting unknown knowledge that cannot be used by large data groups. It is a data cleaner; Data integration; Data selection; Data transfer; Data mining; It is charity for some types of records, containing catalogue data, data storage, or transaction data. [1]

Decision tree study is some of the greatest powerful and standard decision-making implements that can help diagnose and treat medical problems in clinical trials. Radar signal classification; Weather forecasts, etc. The decision is easy to understand and can handle large amounts of dataset. Because the size of the plant is not related to the size of the datasets. The new decision model can be integrated with other machine learning models. The attribute values of each attribute can be determined from the Database. Once the decision tree is completed, it can be used to separate visual or invisible exercises.

ICIDSSD 2020, February 27-28, New Delhi, India

DOI 10.4108/eai.27-2-2020.2303221

Classification is a procedure of facts investigation which can be charity to excerpt prototypes representing significant data programs or to forecast forthcoming information movements that are unknown to a person. It can be used to make decisions. Classification is useful for solving many problems. The place to determine data for items (depending on the level of the item). [1] The C5.0 is a classification tree; it develops a feature based on the investigation of input data. It indicates the use for each node.

2 Literature Review

The study of the effects of organ selection on the neural network algorithm used in the identification of patients with ischemic heart disease, using 12 neural networks, was used. The results from their study were 89.4% of the training model and 82.2% of the trials when all the features were used. An important feature of the decision is that the reduction of features reduces the accuracy of training and testing methods. [3] The fuzzy instrument on cardiac patients intended precision was 94.11% .[4]

Categorization method is popular field of data mining because it actions in agreement with “if and then” rule. Its drive goes to the forecast a constant established on the more characters which are identified by way of forecasters.

The algorithms of NN, SVM and Decision Tree are separated method of categorization procedures. An expansion of the C4.5 algorithm is the C 5.0 algorithm. C4.5 saves all errors in the same way. There is no mistaken interpretation of their position or meaning. A complete development in C5.5 and above C5.0 provides an accurate classification of all errors, depending on their impact on the systemThis is the kind of classification information that helps reduce the cost of misdiagnosis rather than heighten the mistakes. This is known as the standard variable of C5.0 as the misclassification cost. So far we have discussed errors for each feature of a data set. This problem is best solved by using the so-called case weight attribute in C5.0.

This feature is used and the C5.0 reduces the cost of computational predictions. C5.0 holds many types of data compared to C4.5 or previous algorithms. Another weakness is the missing value. This means that if some values are not in the given datasets, or it may be because of an increase in error rates due to other reasons for logging into that account.

C5.0 defines this category for the new name 'irrelevant'. It easily integrates new features as a function of another feature. Compared to C4.5 in computing, C5 is faster. The set of rules requires memory, and there is less room for regulation in standard C5 construction. Figure 1 below shows the accuracy of the C5 and C4.5 comparisons. Display speed and memory. [5]

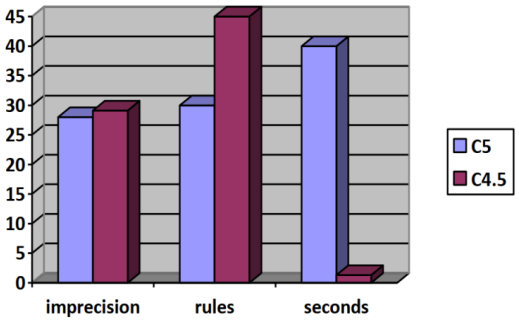

Fig. 1. Comparison of C 5 and C4.5 [5]

C5.0's algorithm is faster contrasted to C4.5 in computing. Decision Tree evaluation is a diagram that is not exact, Displayed in leaves and seconds. The following figure 2 is shown in the comparison of the C 4.5 and C 5.0 algorithm.

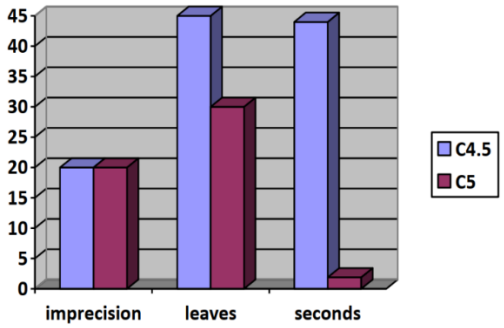

Fig. 2. Decision Tree Comparison of C 5 and C4.5 [5]

3. Investigation of C4.5, CART, C5.0 Algorithms

The C4.5 algorithm stands the upgraded version of ID3 decision tree learning algorithm. CART is a procedure of producing a binary tree, which can handle missing data and contain pruning strategy. C4.5 algorithm finds the best splitting attribute with greatest information gain value using the weights of training instances in training dataset to type a decision process. CART algorithm finds the *gini*(*D*) using the weights of training instances in training records to appear a decision tree. We intended into the novel decision tree learning procedure by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training occurrences 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 inheritor of ID3)
- Categorization and Reversion Trees (CART)
- C5.0 procedure
- Naïve Bayes theorem

3.1. C4.5 Algorithm

Choosing the feature through the greatest materials improvement. Assume p_i stands the possibility assume an subjective tuple into D goes to group C_i, assessed through $|C_{i,D}|/|D|$. Predictable fact required into the grouping 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 required (in searach of spending A to split D into v separations) to arrange D:

For Traditional C4.5

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

where $|D_j|$ = total tuples in D that have outcome a_j of A,
$|D|$ = total tuple

For Weighted C4.5

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

where $|D_j|$ = total weight tuples in D that have result a_j of A,

|D| = total weight tuple

Information gained by branching on attribute A. Information gain extent is influenced regarding the features with a huge amount of values. The algorithm of C4.5 (replacement of ID3) customs gain ratio into affect the difficulty (normalization to info-gain).

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

GainRatio(A) = Gain(A)/SplitInfo(A)

The attribute with the greated info-gain ratio is designated as the splitting feature.

3.2 CART Algorithm

If a data set *D* take into the samples from *m* classes, gini index, *gini*(*D*) is defined as, where p_i stays the comparative 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

When a data set *D* is separted 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

$|D_1|$ = set of tuple D satisfy by A ≤ split-point,
$|D_2|$ = set of tuple in D satisfy by A > split-point
|D| = total tuple

For Weighted C4.5

where

$|D_1|$ = the total weight of tuple D satisfy by A ≤ split-point,
$|D_2|$ = the total weight of tuple D satisfy by A > split-point,
|D| = total weight tuple

Decrease in Contamination:

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

This feature support minimum $gini_{split}(D)$ (or the greatest drcrease into contamination) remains selected towards divide the point (essential to compute all the feasible splitting facts for individually aspect).

3.3. C 5.0 Algorithm

The C5.0 set of rules is an addition of the C4.5 process that too extends ID3. This is a good categorization algorithm for large databases C4.5 faster than memory and performance. C5.0 Model by maximum weight training data records. C5.0 handles missing attributes from very valuable attribute and pest training records. In this research, preparation pest information is charity intended for predicting pest information.

Generating the Algrithm of C5.0 Process

Input

a. Data partition, *D*, a set of training tuples and their associated class labels
b. *attribute_list, the set of candidate attributes*
c. *attribute_selection_method*, a procedure to determine the splitting criterion partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and ,either a *split-point* or *splitting subset*

ouput: C5.0 decision tree

Method:

1. create a node N
2. if tuples in *D* are all of the same class, *C*, then
3. return *N* as a leaf node labelled with the class *C*
4. if *attribute_list* is empty, then
5. return *N* as a leaf node labelled with the majority class in *D*
6. apply *attribute_selection_method(D, attribute_list)* to find the best *splitting_criterion*
7. label node *N* with *splitting_criterion*
8. if *splitting_attribute is discrete-valued and multiway splits allowed then*
9. *attribute_list* ←*attribute_list- splitting_attribute*
10. For each outcome *j* of *splitting_criterion*
    Let D_j be the set of data tuples in *D* satisfying outcome *j*
    if D_j is empty then attach a leaf labelled with majority class in *D* to node *N*
    else, attach the node returned by Generate C5.0 decision tree(D_j, *attribute_list*) to node *N*
11. Return *N*

In this research, the arrangement of C4.5 , CART and C5.0 is used in collecting pesticide training data. C5.0 remains comparable to C4.5, but C5.0 is continuous data. Complete type of information, such as time and duration, is implemented. It can allocate with missing standards of crop information. It mostly maintained improving to advance the classifier accuracy.[2]

3.4 Naive Bayesian Theorem

Assume D stands is a collection of tuples. They are connected labels, and for separate tuple stands signified through an n-D element vector **X** = (x_1, x_2, …, x_n). Assume there are *m* classes C_1, C_2, …, C_m. The most common posteriori means the highest P (Ci | X). It is available since Bayes' theory.

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

Since P(X) is constant for all classes, only wants to be exploited. Attributes are conditional.

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

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

4 Implementation of Dataset

Data management is often ignored, but is an siginificant stage in the procedure of finding important and relevant data. Low Qualitative data can leading to low property mining consequences. Therefore, records can be preprocessed to progress quality and data quality. It may be viewed into positions of comprehensiveness and harmony. In this section, the implementation of C4.5, Cart and weighted C4.5 and weighted Cart algorithms implemented on heart patients' data with the objective of extracting information primary in the data under investigation. In this concern, participating field of diagnosis comprising category model label table 1 was considered as a class-labeled training tuples after the heart diseases record. The sample data was divided into two groups. The training is 70% and the testing is 30% from database.

4.1. Class-Labelled Preparation Tuples from the Heart Disease Catalogue

Table 1. Heart Disease

Age	ST by exercise	exerc ind angig	blood sugar	Max HR	vessles cc	Cholestero	Rest SBP	neter narrov	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

4.2. Accuracy Measurement

Such accuracy, accuracy, and reliability are required to assess the legitimacy of the models. These indices are estimated by the equilibrium matrix below the figure 4. This remains a valuable implement for the analysis of the arrangement process for facts analysis or for different types of data. Ideal situation; Most of the data related to the findings. It would be situated on the central intersection of the matrix. The value of that matrix should be zero. [2]

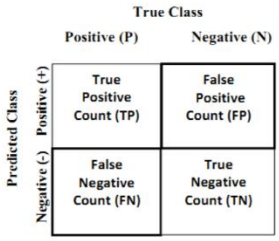

Fig. 4. Confusion Matrix [2]

FN=Number of positive label items that are incorrectly categorized as "negative".
TN= Number of negative label items that are categorized as "Correct".
TP= Number of positive label items that are categorized as "Correct".
FP= Number of negative label items that are incorrectly categorized as "Positive".

4.3. Experimental Results

The experimental results of heart disease classification accuracy results for C4.5 , Cart , Weighted C4.5 and Weight Cart are shown in the following figure 5.

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%

(a) 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

(b) Classification of Speed (seconds)

Fig. 5. Heart Disease Classification Results

The experimental results of heart disease classification time per seconds for C4.5 , Cart , Weighted C4.5 and Weight Cart are shown in the following figure 5(a) and (b).Analyzing the C4.5 and C5.0, they can provide classifiers stated as either decision tree or rulesets. In most presentations, the ruleset is modest. It's easy to evaluate. The main modifications are tree sizes and computation time. C5.0 is used to provide the smaller trees and very fast than C4.5. The C5.0 algorithm is additional accurate, short duration and reduction predictions compare the error rate with the C4.5 algorithm.

5 Conclusion

This paper presented the comparison of traditional decision tree algorithms and weighted decision tree algorithms classification on Heart Disease classification problems. The experimental outcomes proved that the weighted decision tree algorithm can achieve high classification rate on Heart Disease dataset. C5.0 algorithm has been intended to switch the huge record in mining that contains hundreds of thousands of archives. And another good point of C5.0 algorithm can take some kind of values, such as numeric, timestamps or other form. The time complexity of weighted decision tree was slower than conventional decision tree. The weighted CART decision tree algorithm was more accurate than C4.5, and C4.5 was faster than CART. And C5.0 Decision Tree were perfected with the greatest accuracy. The C5.0 algorithm has develop the industry standard for creating decision trees, since it does well for most types of complications directly out of the case. Associated to additional sophisticated machine learning simulations such as neural network, support vector machine and the decision trees under the C5.0 algorithm normally implement almost as well but are much easier to realize and arrange.

References

1. R. Revathy and R. Lawrance, "Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data, No. 5, Iss 1, March 2017.

2. M.Abdar and S. R. N.Kalhori, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases", IJECE, December 2015.

3. K.Rajeswari, V.Vaithiyanathan and T.R. Neelakantan, " Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. (IRIS 2012).

4. A.V. S. Kumar, "Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis Heart Disease", IJCA, 2013; 77(11): 6-12.

5. M.Arif, "Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review", IJDTA, Vol. 11, No.1 (2018), pp. 1-8.

6. Durg C.G, "Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection ", IJCA, Volume 133 – No.9, January 2016.

7. Dr. W. H.Wolberg (physician), "Breast Cancer Wisconsin (Original) Data Set", University of Wisconsin Hospitals Madison, USA.

8. M.T. Mon, Classification the Stages of Dental Caries using C4.5 Decision Tree University of Computer Studies, Yangon.

9. S.Hlaing, ICU Patients Risk Level Classification System using CART Algorithm, University of Computer Studies, Yangon.

10. J.Han, M.Kamber, J.Pei, "data mining concepts and techniques", 2006.