International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-3, Issue-2, March-2016

Performance Analysis of Cart & Id3 Algorithm for Heart Prediction System

Prof. Anil Hingmire
Computer Engineering Department.,VCET, Vasai,Maharashtra India

Nikita M. Chaudahri
Computer Engineering Department,VCET, Vasai,Maharashtra India

Utkarsha P. Patil
Computer Engineering Department,VCET, Vasai,Maharashtra India

Lalita S. Mahajan
Computer Engineering Department,VCET, Vasai,Maharashtra India

ABSTRACT
Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems issuing inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules .It is one of the most effective forms to represent and evaluate the performance of algorithms, due to its various eye catching features: simplicity, comprehensibility, no parameters, and being able to handle mixed-type data. There are many decision two algorithm available named ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE, and CTREE. We have explained three most commonly used decision tree algorithm in this paper to understand their use and scalability on different types of attributes and feature. ID3(Iterative Dichotomizer 3) developed by J.R Quinlan in 1986, CART stands for Classification and Regression Trees developed by Breiman et al.in 1984).

Keywords
Regression, Node, Prediction etc.

1. INTRODUCTION

Keeping in view the goal of this study to predict heart disease using classification techniques, we used different supervised machine learning algorithms i.e., Decision Tree, Classification & Regression .The performances of the Algorithms in this study were evaluated using the standard metrics of accuracy, precision.[4]

On the first scenario the algorithm was run on a full training set containing 7,339 instances with 15 attributes. It took 0.89 second to build the model and the model generated a tree and leaves. On the second scenario the algorithm was run on a full training set containing 7,339 instances with [4]

Selected 8 attributes. It took 0.36 second to build the model and the model generated smaller and less complex tree.[1]

This research has found that the CART algorithm performs better than ID3 and C4.5 algorithm, in terms of classifier accuracy. The advantage of CART algorithm is to look at all possible splits for all attributes. Once a best split is found, CART repeats the search process for each node, continuing the recursive process until further splitting is impossible or stopped, for that the CART algorithm has been used to improve the accuracy of classifying the data.[6].

2. ANALYSIS

2.1 CART

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees. Decision trees are then used to classify new data. In order to use CART we need to know number of classes a priori. CART methodology was developed in 80s by Breiman, Freidman, Olshen. For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations. For example, learning sample for credit scoring system would be fundamental information about previous borrows (variables) matched with actual payoff results (classes).[6]

CART methodology consists of three parts:

1. Construction of maximum tree .
2. Choice of the right tree size.
3. Classification of new data using constructed tree.

2.1.1 Steps for developing tree

- Rules based on variables values are selected to get the best split to differentiate observations based on the dependent variable e.g.: **Age, blood pressure, sugar**, etc..
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node .The rule is like that **Age>60?**
- Splitting stops when CART detects no further gain can be made. [6]

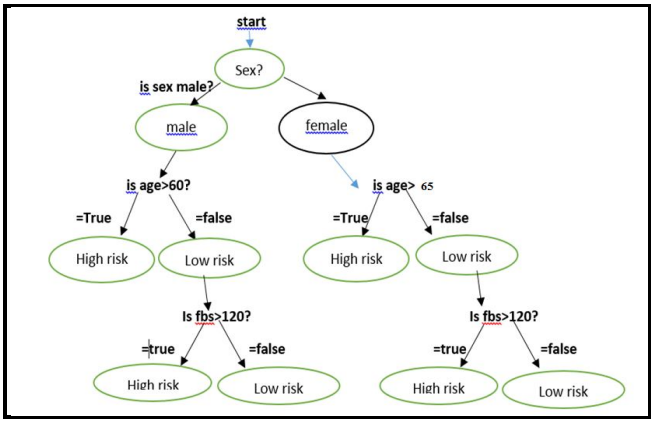

Figure 1: Implementation of CART

2.1.2 Advantages of cart:

- CART is nonparametric. Therefore this method does not require specification of any functional form.
- CART does not require variables to be selected in advance.
- CART algorithm will itself identify the most significant variables and eliminate non-significant ones.
- CART results are invariant to monotone transformations of its independent variables. Changing one or several variables to its logarithm or square root will not change the structure of the tree.
- CART has no assumptions and computationally fast. There are plenty of models that can not be applied to real life due to its complexity or strict assumptions.
- CART is flexible and has an ability to adjust in time. The main idea is that learning sample is consistently replenished with new observations. It means that CART tree has an important ability to adjust to current situation in the market.

2.1.3 Disadvantages of CART:

- CART may have unstable decision trees.
- Insignificant modification of learning sample, such as eliminating several observations, could lead to radical changes in decision tree: increase or decrease of tree complexity, changes in splitting variables and values.
- CART splits only by one variable.

2.2 ID3

"Iterative Dichotomizer 3" was Invented by Ross Quinlan in 1979.Algorithm used to generate a decision tree. It Classifies data using the attributes&Tree consists of decision nodes and decision leafs. Nodes can have two or more branches which represents the value for the attribute tested. Leafs nodes produces a homogeneous result. It Attempts to create the smallest possible decision tree.[5]

2.2.1 ID3 Algorithm:

Process:-

a) Take all attributes and calculates their entropies.
b) Chooses attribute that has the lowest entropy is minimum or when information gain is maximum
c) Makes a node containing that attribute.

Splitting Criteria Based on Entropy

- Entropy = - p+log2 (p+) -p-log2 (p-) for a sample of negative and positive elements.
- EG:-Calculate the entropy

Given: Set S contains14 examples

- 9 Positive values
- 5 Negative values
- Entropy(S) = - (9/14) Log_2 (9/14) - (5/14) Log_2 (5/14) = 0.940

2.2.2 Advantages of ID3:

- Predict new data
- Training set is used to create rules for predicting.
- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests. [5]

2.2.3 Disadvantages of ID3:

- Considers only one attribute to create nodes
- Numerous trees needed for continuous data
- Over classification for small data.
- Data may be over-fitted or over-classified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.
- Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.[8].

2.3. Decision Table

Decision tables (DTs) provide an alternative way of representing rule-based classification models which is known as tabular representation used to describe and analyze decision situations.[1]

In the analysis phase we perform the analysis on CART and ID3 analysis based on following some criteria

- Timing to build model (in Sec)
- Correctly classified instances
- Accuracy (%)
- Efficiency
- Structure

Table 1: Analysis Table

Evaluation Criteria	**Classifiers**	
	CART	ID3
Timing to build model (in Sec)	0.24	0.02
Correctly classified instances	253	222
Accuracy (%)	82.59%	72.93%
efficiency	excellent	good
structure	simple	complex

3. CONCLUSION

In this paper we have studied the various basic properties of the decision tree algorithms which provides as a better understanding of these algorithms .We can apply them on different types of data sets having different types of values and properties and can attain a best result by knowing that which algorithm will give the best result on a specific type of data set. This research work compares the performance of ID3 and CART algorithms. The experimentation result shows that the CART has the best classification accuracy when compared to ID3.

REFERENCES

[1] Wei Peng, Juhua Chen and Haiping Zhou, of ID3,' An Implementation Decision Tree Learning Algorithm', University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia .

[2] LiorRokach and OdedMaiman, "Chapter 9 Decision Trees "Department of InduatrialEngineering,"Tel-Aviv University"

[3] Vipin Kumar, Pang-Ning Tan and Michael Steinbach, "Introduction to Data Mining" Pearson.

[4]AjanthanRajalingam,DamandeepMatharu,KobiVinayagamoorthy,NarinderpalGhoman, "Data Mining Course Project: Income Analysis" SFWR ENG/COM SCI 4TF3, "December 10, 2002".

[5] Ahmed Bahgat El Seddawy, Prof. Dr Turkey Sultan, Dr. Ayman Khedr, "Applying Classification Technique Using DID3 Algorithm to Improve Decision Support Under Uncertain Situations". 'Department of Business Information System, Arab Academy for Science and Technology and Department of Information System, Helwan University, Egypt. "International Journal of Modern Engineering Research ", Vol 3, Issue 4, July-Aug 2013 pp-2139-2146.

[6]Roman Timofeev to Prof. Dr. Wolfgang Hardle"Classification and Regression Trees (CART). Theory and Applications," CASE- Center of Applied Statistics and Economics, Humboldt University, Berlin Dec 20, 2004.

[7] Mohammad M Mazid,A B M Shawkat Ali, Kevin Tickle, "Improved C4.5 Algorithm for Rule Based Classification" School of Computing Science, Central Queensland University, Australia.

[8] http://www.wikipedia.com

