IJIRMET
ISSN (Online): 2456-0448
International Journal Of Innovative Research
In Management, Engineering And Technology
Available online

COMPARISON OF VARIOUS DECISION TREE ALGORITHMS

[1] Anjali K K, [2] Anusha K R, [3] Muhammed Ali, [4] Shamily George
[1] [2] [3] [4] Computer Science & Engineering, St.Thomas College of Engineering & Technology, kerala, India.

To access & cite this article
Website: www.ijirmet.com

ABSTRACT

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. Classification is a supervised machine learning algorithm which deals with identifying, to which of the set of categories a new observation belong on the basis of training set of data containing observations whose category membership is known. Decision tree builds classification or regression models in the form of tree structure. This paper focuses on comparison of ID3, CART, C4.5, C5.0 and random forest.

Keywords: Machine learning, Data mining, Classifier, ID3, CART, C4.5, C5.0, Random forest.

I. INTRODUCTION :

Classification is a task that occurs very frequently which involves dividing an object so that each is assigned to one of a number of exhaustive and exclusive categories known as classes. Decision tree is an important model to realize the classification. Data mining is extraction of hidden predictive information from large database. Classification methods aim to identify the classes that belong from some descriptive traits. They find utility in a wide range of human activities particularly in automated decision making. Decision trees are very effective methods of supervised learning. It aims in the partition of dataset into groups as homogeneous terms of variable to be predicted. It takes as input, a set of classified data, and outputs a tree that resemble an orientation diagram where each end node is a decision and non-final node represents a test. Each leaf represents the decision of belonging to a class of data verifying all test paths from the root to the leaf. Growing a tree involves deciding on features to choose and conditions to use along with knowing when to stop. Performance of a tree can be increased by pruning (removing branches that make use of features having low importance). Advantage of using decision tree includes transparency, easy to use, comprehensive nature and flexibility.

II. A COMPARATIVE STUDY OF ID3 AND CART :

In decision tree learning, ID3 (iterative dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate decision tree from a dataset. ID3 can over fit the training data and is also harder to use on continuous data[2]. It uses attributes without missing data. ID3 algorithm selects the best attribute based on concept of entropy and information gain for developing the tree[5].
Shannon Entropy H(S) is a measure of the amount of uncertainty in set S.

$$H(S) = \sum_{x \in X} -p(x)\log_2 p(x)$$

Where, H(s) is a measure of amount of uncertainty in the data.

S – The current dataset for which the entropy is being calculated.

X- it is the set of classes in S.

P(x)- it is the proportion of the number of elements in class x to the number of elements in class S.

If H(S)=0, the set S is perfectly classified.

Information gain is the measure of the difference in entropy from before to after the set S is split on an attribute A

$$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A)$$

Where, H(S)- Entropy of set S.

T- The subset created from splitting set S by attribute A.

P(t)- The proportion of the number of elements in t to the number of elements in set S.

H(t)- Entropy of subset t.

The attribute with largest information gain is used to split the set S first.

Classification and Regression tree (CART) is method to describe how the variable Y distributes after assigning the forecast vector X. The model uses binary tree to divide the forecast space into certain subsets[1]. Tree's leaf nodes correspond to different division areas which are determined by Splitting Rules relating to each internal node.

CART uses GINI index to determine in which attribute the branch should be.

Let S be the sample, a the target attribute S1, S2….. Sk starting from S according to the classes of a,

$$Gini(S) = \sum_{i=1}^{K} \frac{|S_i|}{|S|}\left(1 - \frac{|S_i|}{|S|}\right) = \sum_{i \neq j} \frac{|S_i| \times |S_j|}{|S|^2}$$

The strategy is to choose the attribute whose GINI index is minimal after splitting.

- Attribute type of ID3 is categorical whereas CART supports both categorical and discrete types.

- ID3 does not handle missing values whereas CART handles missing values.
- There is no pruning strategy used in ID3 whereas CART supports pruning based on cost complexity.
- ID3 is susceptible to outliers whereas CART can handle outliers.

III. A COMPARATIVE STUDY BETWEEN CART AND C4.5 :

The CART implementation is similar to C4.5. CART uses Gini Index to find the best attribute that splits data set S whereas C4.5 uses Gain ratio. C4.5 is an extension of CART[1].

$$GainRatio(p,T) = \frac{Gain(p,T)}{SplitInfo(p,T)}$$

Where SplitInfo is:

$$SplitInfo(p,test) = -\sum_{j=1}^{n} P'\left(\frac{j}{p}\right) * log\left(P'\left(\frac{j}{p}\right)\right)$$

Where, P'(j/p) – proportion of elements present at the position p, taking the value j-th test.

Both CART and C4.5 can handle missing values in attribute.

Both CART and C4.5 can handle categorical and discrete attribute types.

In CART pruning is done by cost-complexity method and error based pruning Is done in C4.5.

CART can handle outliers whereas C4.5 is susceptible to outliers.

CART constructs the tree based on numerical splitting criteria recursively applied to the data, whereas in C4.5 includes the intermediate step of constructing rule sets.

IV. A COMPARATIVE STUDY OF C4.5 AND C5.0 :

C4.5 was superseded in 1997 and C5.0 is an extension of C4.5[2]. C5.0 splits based on information gain similar to ID3 whereas C4.5 splits using Gain Ration criteria. The changes encompass are capabilities as well as much improved efficiency[8]. New data types have been formulated such as “not applicable” and unordered rule set are defined which improves both interpretability of rule set and predictive accuracy and also improved scalability of both decision trees and rule sets. Scalability is enhanced by multi-threading. Unlike C4.5, C5.0 handles all types of data like continuous, dates, times and timestamps. It also supports boosting to improve classifier accuracy.

Speed of C5.0 algorithm is significantly faster and more accurate than C4.5.

The pruning strategy used in C4.5 is error based pruning whereas C5.0 use binomial confidence limit method.

C5.0 does not work well with small training samples. Both supports discrete and categorical attribute types.

C5.0 is better than C4.5 on efficiency and memory.

V. RANDOM FOREST :

Random forest is an ensemble classifier which uses many decision tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. The basic difference is that Random Forest (RF) is a collection of numerous decision trees. Random Forest can be used to rank the importance of variables in regression or classification problem in a natural way. In random forest each decision tree gives a vote for the prediction of target variable[4].

ADVANTAGE OF RANDOM FOREST:

- High predictive accuracy.
- Efficiency on large datasets.
- Ability to handle multiple input features

without need for feature deletion.

- Prediction is based on input features considered important for classification.
- Works well with missing data still giving a better predictive accuracy.

DISADVANTAGES OF RANDOM FOREST:

- Not easily interpretable.
- Random Forest over fit with noisy classification or regression.

Random Forest is better than decision tree in the sense that deep decision trees might suffer from over fitting. Random Forest prevent over fitting most of the time by creating random subsets of the features and building smaller trees using these subsets. Afterwards, it combines the sub tree. It doesn't work every time and it also makes computation slower depending on how many trees random forest builds.

VI. CONCLUSION :

In this paper, we considered four classification algorithm of decision tree, ID3, CART, C4.5, C5.0. According to the comparison it is inferred that C5.0 has better performance when compared to other algorithms. This paper also deals with a brief study of Random Forest, which is an extended version of decision tree and its advantage over decision tree.

VII. REFERENCE :

1. Prof. Nilima Patil, prof. Rekha Lathi and prof. Vidya Chithre, Comparison of C5.0 and CART classification algorithm using pruning technique, Vol .1 Issue 4, June-2012.
2. Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI and Mohammed ERRITALI, A comparative study of decision tree ID3 and C4.5.
3. R. Revathy and R. Lawrence, Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data, Vol. 5, special issue 1, March 2017.
4. Prajwala T R, A Comparative Study on Decision Tree and Random Forest Using R Tool. Vol. 4, Issue 1, January 2015.
5. Wei Peng,Juhua Chen, and Haiping Zhou. An Implementation of ID3 – Decision Tree Learning Algorithm. Retrieved March 10.
6. Miao Li, Application of CART decision tree combined with PCA algorithm in intrusion detection(IEEE), 23 April 2018.
7. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, 2019. "The Influence of 'Adventure Tourism Activities' in promoting tourism business in mountain stations", African Journal of Hospitality, Tourism and Leisure, Volume 8 (2).
8. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, Nov 2018. "A Study On consumer satisfaction and preference of colour TV brands in Chennai city", International Research Journal of Management and Commerce, Volume4, Issue 10.
9. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, "A Study on Attrition: Turnover intentions of employees", Jan 2019. International Journal of Civil Engineering and Technology (IJCIET), Volume 10, Issue 9.
10. Dr. S. Rabiyathul Basariya, and Dr. Nabaz Nawzad Abdullah, Dec 2018. "A STUDY ON CUSTOMER'S SATISFACTION TOWARDS E-BANKING", International Research Journal of Management and Commerce, Volume 5, Issue 12,
11. R. Sudhakar and Dr. S. Rabiyathul Basariya, "IMPACT REGARDING THE TRAINING PROVIDED FOR THE EMPLOYEES AND ITS EFFECTIVENESS BETWEEN THE PUBLIC SECTOR COMPANIES VS. NEW GENERATION COMPANIES ON SELECTED MIDDLE AGED EMPLOYEES", International Journal of Mechanical Engineering and Technology (IJMET) Volume 9, Issue 2, February 2018, pp. 300–306.
12. Huang Ming, Niu Wenying and Liang Xu, An improved Decision Tree classification algorithm based on ID3 and the application in score analysis(IEEE), 07 August 2009.
13. Saba Bashir, Usman Qamar, Farhan Hassan khan and M. Younus Javed, An efficient rule based classification of Diabetes using ID3, C4.5, &amp; CART Ensembles(IEEE), 08 June 2015.
14. A C Tsoi and R A Pearson, Comparison of three classification techniques, CART, C4.5 and Multi-Layer Perceptrons.