Decision Tree Classifier for Mining Data Stream: A Survey

[1]Pranali T. Sawant, [2]Mahesh C. Ahire, [3]Ragini P. Mahale, [4]Ms. T. S. Pawar

[1,2,3]Student, [4]Professor
Department of Information Technology,
Karmaveer Adv. Baburao Ganpatrao Thakare College of Engineering, Nashik, India

***Abstract*: Big data is raising many technical challenges in today's world like the amount of data is getting higher and higher, streaming data and the data dimensionality which affects academic researches and IT sectors. This was found that streaming data accumulates exponentially making traditional methods to become infeasible during real-time data mining and also to extract useful knowledge from it. Generally, for the purpose of decision making, decision trees are used for information gaining. But, due to certain drawbacks, some new approach needs to be used. A learning algorithm is to be implemented for mining streaming data which is more precise. By using Random Forest algorithm for classification which achieves enhanced analytical accuracy within reasonable processing time against streaming data. In this survey, we had explained most commonly used decision tree algorithms like CHAID, CART, C4.5, ID3 and also Random Forest algorithm.**

***Index Terms*: Decision tree, ID3, C4.5, CART, CHAID, and Random Forest.**

I. INTRODUCTION

This research is based on the domain of Data Mining and Data Analysis. Traditional data mining algorithm lack in precision, accuracy, requires large computational power and consumes lots of time for mining. We can mine the data using regular data mining algorithm, but due to the above problems, it becomes infeasible to the user to use them. Data mining algorithm such as Option Tree, Decision Tree, Hoeffding Tree, etc., successfully mines the data and the result are also obtained but the accuracy of the result is not precise as per the prediction of the user. This is so because of the uncertainty in the data sets. Normally, when we use a decision tree for data mining on large datasets, the result obtained is the combined result of the large set of data. But if the similar data sets are mined using the number of decision trees along with some different combination of feature subsets, then individual decision tree will have an output which might be different from outputs of other trees. The best result which has maximum number of occurrences is used as final result of data mining. The group of decision tree used can be collectively called as forest of trees. And this whole method is termed as Random Forest.

The motive of the project is to upgrade the data mining technique with some recent technologies which enhances the efficiency of mining. Various studies and practical implementations are done on similar types of project. Various decision tree algorithms were used. But, decision tree is an traditional method and the result obtained from it is not so much precise. While the recent data mining techniques like random forest can be used and the output gained will be more precise while it will consume same computational power and time.

II. DECISION TREE

The decision tree have their benefits which are like: It makes straight forward visualizations and as the internal workings are capable of being observed which makes it possible to regenerate the work. Decision trees can handle numerical as well as categorical data and also it performs well on large datasets. They are extremely fast. Beside it also have some drawbacks like small changes in datasets may produce meaningfully different models which can be unstable. It also needs careful adjustment through pruning.

Some previous researchers encountered problem with decision trees because of its low accuracy and therefore random forest was introduced to enhance both performance and accuracy to achieve higher efficiency. As random forest is easy to understand and also more effective and accurate, it was applied from decision trees. Random forest is popular and widely used in research area.

Decision trees are used to construct a structure with a bunch of instances which can be used to classify the new instances. Bunch of attributes or features which can have numeric or symbolic values are described by each instance. Therefore, Decision Tree model is upgrade into very efficient Random Forest, which is a data mining technique and it has obtained increasing popularity [1].

A. CHAID

CHAID stands for Chi-squared Automatic Interaction Detector choice tree method was produced in South Africa and is utilized for expectation and grouping. The systems depend on balanced criticalness the testing and utilized for recognition of communication between factors. CHAID is an expansion of the Automatic Interaction Detection (AID) and Theta Automatic Interaction Detection (THAID) systems. This method is generally utilized as a part of the light of coordinating and database promoting exploration and makes a forecast how unique gathering of clients reactions influence the variable. CHAID utilizes multiway parts by defaults, for bigger example sizes of clients aggregate it work viably and the unwavering quality investigation is more than the littler example sizes.

CHAID (Chi-square Automatic Interaction Detector) investigation is a calculation utilized for finding connections between a downright reaction variable and other clear cut indicator factors. It is helpful when searching for designs in datasets with loads of straight out factors and is an advantageous method for abridging the information as the connections can be effortlessly envisioned.

Practically speaking, CHAID is frequently utilized as a part of direct promoting to see how extraordinary gatherings of clients may react to a crusade in light of their attributes.

B. CART

A Decision tree is a type of supervised machine learning algorithm that is mostly used in classification problem. CART is one of the decision tree algorithms. CART was introduced by Breiman in 1984 and it
is stands for classification and regression trees. It also provides users to provide prior probability distribution. CART builds both classifications and regression trees.
Inside the most recent 10 years, there has been expanding enthusiasm for the utilization of characterization and relapse tree (CART) investigation. Truck examination is a tree-building strategy which is not at all like conventional information investigation strategies. It is preferably suited to the age of clinical choice tenets. Since CART investigation is not at all like different examination techniques it has been acknowledged generally gradually. Besides, most by far of analysts have almost no involvement with the system. Different variables which constrain CART's general adequacy are the unpredictability of the investigation and, as of not long ago, the product required to perform CART examination was hard to utilize. Fortunately, it is currently conceivable to play out a CART investigation without a profound comprehension of every one of the different advances being finished by the product. Furthermore, CART is frequently ready to reveal complex associations between indicators which might be troublesome or difficult to reveal utilizing customary multivariate strategies. The motivation behind this address is to give an outline of CART approach, accentuating down to earth utilize instead of the basic measurable hypothesis.

C. C4.5

C4.5 is a new way of classifying the decision trees generated through various sources, whereas, the decision trees created by C4.5 are used to simplify classifications, it is frequently used as an easiest way for the simplification of statistical classifiers. From a set of training data C4.5 builds decision tree like ID3, using concept of information entropy:

$$Entropy(S) = \sum_i^c \; - p_i \, log2 \, p_i \qquad (1)$$

where c = number of values in the target attribute (number of classes)
pi = number of samples in class i

$$Gain(S, A) \; = \; Entropy(S) - \sum_{Vevaluation(A)} \frac{|S_v|}{S} \, entropy(s_v) \qquad (2)$$

Where A = attribute
V = probable value for attribute A

At every junction of decision tree, C4.5 selects the attributes of the data that are most effectively splits set of data into subsets improve in single class. The attributes which are having highest normalized information gain is selected to make decision.C4.5 handles both continuous as well as discrete(categorical) attributes, it can also handle training data with missing attribute.C4.5 builds models that can be easily explained and it is easy to implement. It performs a tree pruning process, which generates smaller trees and more simpler rules.C4.5 algorithm is very efficient algorithm.C4.5 is appropriate for real world problems as it deals with numeric attributes as well as missing values. This C4.5 algorithm can be used for constructing smaller or larger and correct decision trees.

D. ID3

ID3 is a decision tree learning algorithm which is introduced by Quinlan Ross in 1986. Decision tree classifies data using the attributes. Tree consists of decision nodes and decision leafs. The nodes can have two or more branches which represents the value for the attribute tested. ID3 stands for Iterative Dichotomiser 3. ID3 is implemented and based on Hunt's algorithm [7]. The basic idea of Iterative Dichotomiser 3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In the decision tree method, information gain approach is used to determine suitable property for each node of a generated decision tree [8]. So, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. ID3 has advantages and disadvantages as has the following [9]:

Advantages

1. Prediction rules on training data are easy to understand.
2. ID3 can construct the fastest tree.
3. ID3 can construct a short tree.
4. Only got to take a look at enough values till all data is classify.
5. Finding leaf nodes allows test data to be pruned, reducing number of tests.
6. Whole dataset is searched to create tree.

Disadvantages

1. Data can be over-classified or over-fitted, if a small element is tested.
2. It does not handle missing values and numeric data attributes

III. RANDOM FOREST

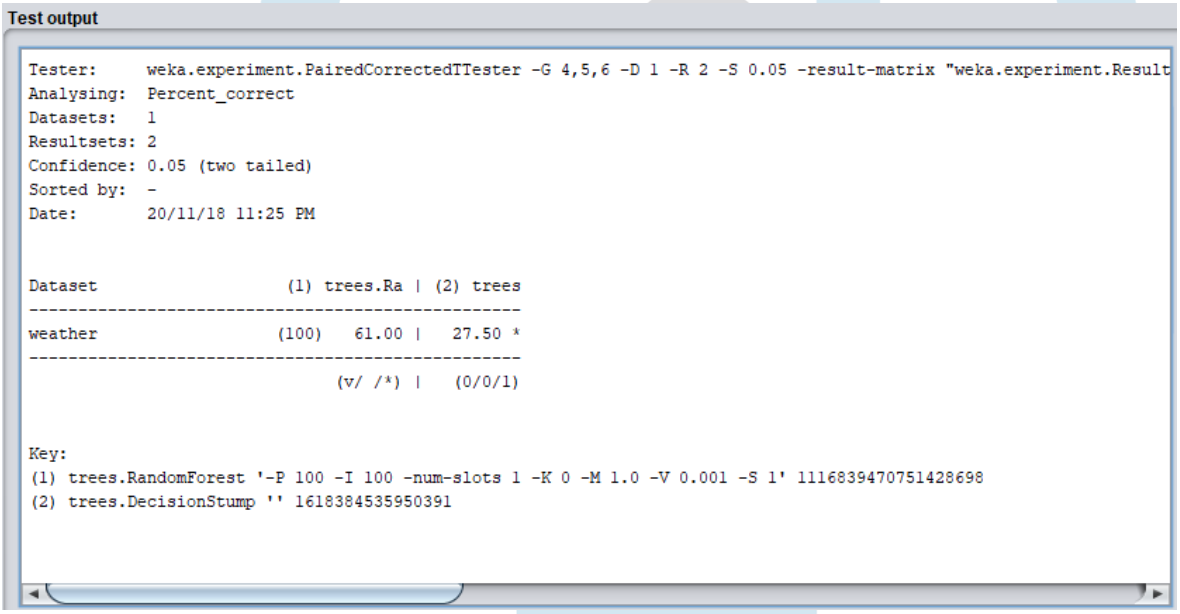
Random Forest algorithm is most favored and most strong supervised machine learning algorithm. It is efficient of performing Regression as well as Classification tasks. This algorithm creates a forest having number of decision trees. In general more trees in forest more robust the prediction and thus higher accuracy [2].In random forest we grow multiple trees as opposed to a single tree to classify a new object. Based on attributes each tree gives a classification and we save the tree votes for that class. The forest chooses the classification have maximum votes (overall trees in forest) and in case of regression it takes average of the outputs by different trees.

Some of the pros of random forest that it can handle the missing values and maintains accuracy for missing data. Also, when there are more trees in forest random classifier wont overfit the model. It also has power to handle large datasets with higher dimensionality. But, random forests also have some drawbacks which are: It does good job at classification but it is not good for regression as it does not give exact continuous nature predictions. For statistical models Random Forest can be like black box approaches which have very small control of what model does.

Random Forest can be used in banking sector for finding fraud customers and loyal customers. In Medicine sector to identify correct combination of components to validate and also for identifying disease by analyzing patients medical records. In Stock market to identify stock behavior as well as expected loss or profit by purchasing a particular task. In computer vision, Random forest is used for image classification. Microsoft has used it for body parts classification for Xbox kinect. It can be also applied for other applications like lip reading, voice classification.

Experiment:

To show the execution of our study we used an open source data mining tool called as Weka (Waikato Environment for Knowledge Analysis). For our experimentation we used weather dataset which is readily available in Weka software. We uses two classification algorithm i.e. decision tree and random forest. After the experiment we found that random forest performs much better than decision tree and thus accuracy of the mining gets increase.



The Characteristic of these algorithms are explained in Table 1 given below [7]:

TABLE I. CHARACTERISTICS OF DECISION TREE ALGORITHMS

Algorithm	Splitting Criteria	Attribute Type	Missing Values	Pruning Strategy	Outlier Detection
ID3	Splitting Criteria in ID3 is Information Gain	It handles Categorical value.	It does not handle missing values.	Pruning is not done in ID3.	It is Susceptible on outliers.
CART	Towing criteria is used in CART	It handles categorical as well as numeric values.	It handles missing values.	Cost complexity pruning is used in CART.	CART can handle outliers.

C4.5	Gain Ratio is the splitting criteria in C4.5	It handles categorical as well as Numeric values.	It handles missing values.	In C4.5 Error Based pruning is used.	It is Susceptible on outliers
Random Forest	Splitting Criteria in Random Forest is Information Gain	It handles categorical as well as numeric values	It handles missing values	Pruning is not done in Random Forest	Bootstrapping is susceptible to outliers

CONCLUSION

The whole study in this paper was to study various decision tree algorithms and to understand the usage of each algorithm wisely. But the efficiency of various algorithms can be analyzed using their accuracy. Along with them the study also reached to random forest algorithm which showed various advantages over the decision tree. The overall analysis and theoretical study of random forest and decision tree showed that the random forest would always generate a better result as compare to decision tree.

REFERENCES

[1] ”Data Mining: Classification and Prediction Encyclopedia of Bioinformatics and Computational Biology”, Volume 1, 2019, Pages 384-402 Alfonso Urso,Antonino Fiannaca, Massimo La Rosa, Valentina Ravì, Riccardo Rizzo.

[2] “Advanced Algorithms for Data Mining Handbook of Statistical Analysis and Data Mining Applications” (Second Edition), 2018, Pages 149-167 Robert Nisbet, Gary Miner, Ken Yale.

[3] “Random Forest for Salary Prediction System to Improve Students Motivation”.

[4] L. Breiman, “Random Forests”. Machine Learning, vol. 45, no. 1,pp. 5-32.2001.

[5] “Implementation of Decision Tree Using C4.5 Algorithm in Decision Making of Loan Application by Debtor” (Case Study: Bank Pasar of Yogyakarta Special Region).

[6] Himani Sharma1, Sunil Kumar2 “A Survey on Decision Tree Algorithms of Classification in Data Mining” 2015.

[7] [7] Sonia Singh, Priyanka Gupta “Comparative Study ID3, CART AND C4.5 Decision Tree Algorithm: A Survey”July 2014.

[8] [8] 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.

[9] [9] 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.