International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining

S.Nagaparameshwara chary, Dr.B.Rama
Research Scholor,Department of Computer Science, Kakatiya University,TS India
Assistant Professor,Department of Computer Science , Kakatiya University,TS India

Abstract : ***Data mining is an active area of Research. Data mining is the process of extracting knowledge from the large amount of data. A large amount of data can be exploration and analysis by using data mining to discover potentially useful and understandable patterns in data. In this Data mining there are various techniques available to mine the large amount of data. The data mining technique are Classification, Association rule mining, cluster analysis, outlier analysis. In this data mining techniques classification technique is the most powerful technique to classify our large amount of data. In this classification, the Decision trees are the most prominent classification technique in data mining. Decision tree induction is the learning of decision trees from class labeled training tuples. A decision tree is flow chart like tree structure. Decision tree has used to solve a wide variety of implications in data mining process. This is used in statistics, machine learning. Data mining also a predictive model. Decision trees are classified by using the two phases, first phase is tree building phase and the second phase is tree pruning phase. In this paper discuss about various data mining decision tree algorithms like ID3.C4.5,J48,CART and the performance analysis of the algorithms.***

Keywords: Data mining, Decision tree, classification, ID3,C4.5,J48,CART.

I . INTRODUCTION

Decision Trees are the one of the most powerful classification technique in Data mining. By using this technique, it builds the models in the form of tree structure. In datasets breaks in small sets and concurrently an associated decision tree is formed. The Decision trees are handles both numerical data and categorical data[1].The concept Decision tree classification is based on Decision tree Induction, which is defined as the learning of decision trees from class-labeled training tuples. A Decision tree is represented as a Graphical tree structure in which each outcome of the test, and each leaf node(or terminal node) holds a class label. The topmost node is considered as the root node[2].

II. DATA MINING

Data mining is an active area of research. A large amount of data can be exploration and analysis by using data mining to discover valid, novel potentially useful and ultimately understandable patterns in data. Data mining is the essential step in discovery the knowledge. Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge. There are different data mining techniques available ,in this decision trees are most useful concept for classifying our data. In the process of mining the data, it takes data as input and it provides knowledge as output. It is also called as knowledge mining. This is the part of knowledge discovery process. The advanced computation methods are applied in data mining for extracting the information from data[3].

III. THE CLASSIFICATION TECHNIQUE

The Classification technique is most widely used technique in Data mining. Data mining that use classifier to predict categorical class labels, it is the one of the major issues in the field of Data mining Research[8].

Classification is one of the most important data mining problem. The input is a dataset of training record, where in each record has several Attributes. Attributes with numerical domains are called numerical attributes and attributes whose domain is non-numerical are called categorical Attributes. There is also a distinguished attribute called the Class label. This classification aims at building a console model that can be used to predict the class label future, unlabeled records. Many classification models including Decision trees, Naive Bayes, k-Nearest Neighbor.[9]

Classification is a classic data mining technique based on machine learning .Basically classification is used for classify each item in a set of data into one of the predefined set of classes or groups. Decision tree induction is the learning of decision trees these are a flow chart-like tree structure, where each internal node(non-leaf node)denotes a test on attribute, each branch represents an outcome of the test and each leaf node(or terminal node)

holds a class label. The topmost node in a tree is the root node[10].

The Classification Steps

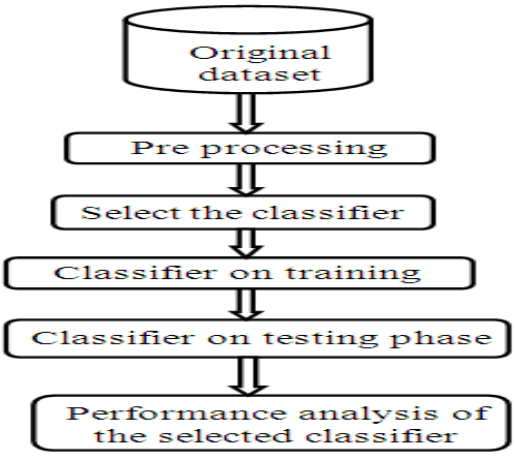

Figure1: The Classification Steps

IV. LITERATURE REVIEW

The Literature review concentrates about the various Decision Tree algorithms are based on different Domains. In this the performance of each algorithm is analyzed based on various different parameters.

Aman Kumar sharma, Suruchi, Sahani in their research they are conducted an experiment in weka environment by using four Decision tree algorithms namely ID3,C4.5,Simple CART and alternating decision tree on the student dataset and later these four algorithms were compared in terms of the classification accuracy. According to the simulation results, the C4.5 algorithm out performs the ID3,CART and ADTree in terms of Classification accuracy[5].

Manpreet Sing, Sonam Sharma,Avinash Kaur are present in their Research, they discuss about Decision Tree Algorithm.

ID3 Algorithm:

TheID3(IterativeDichotomiser)Algorithm is introduced in 1986 by Quinlan Ross. The ID3 is a Greedy approached learning decision tree algorithm. This algorithm is recursively selects the best attribute as the current node using top down induction. Then the child nodes are generated for the selected Attribute.

It uses on information as entropy based measure to select the best splitting attribute and the attribute with the highest information gain is select as best splitting attribute. The level of accuracy is not maintained by the algorithm when there is two much noise in the training dataset. The disadvantage of this algorithm is that it accepts only categorical attributes and only one attribute is tested at a time for making decisions. this concept of pruning is not present in ID3 algorithm[3].

Vaithyanathan.V,K.Rajeshwari, Kapil Tajane, Rahul pitale are discussed in their research about J48 Algorithm. J48 algorithm is called as optimized implementation of the C4.5 or improved version of the C4.5. Output given by J48 is the Decision tree. Decision tree is same as that of the tree structure having different nodes, such as root node, intermediate nodes and leaf node. Each node in that tree contains a decision and the decision leads to our result as name is decision tree. Decision tree divide the input space of a dataset into mutually exclusive areas, where each are having a label, a value to describe or elaborate its data point. The splitting criterion is used in decision tree to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node.

Kasra Madadipouya said in the research C4.5 is the enhanced version of ID3.It is based on numeric input attributes. It builds trees in 3 phases.

a) creating splits for categorical attributes like ID3 algorithm. This algorithm considers all probable binary splits for numerical attributes. Splits of numeric attributes are always binary.

b) Evaluation of the greatest split according to gain ratio metric

c) testing the stop criterion and repeating the steps in a recursive manner for new subdivisions. These three steps are done iteratively for all of nodes.

Asli calis, Ahmet Boyaci, Kasim Baynal are discussed in their research about the CART Algorithm has the nature of being the continuation of the decision tree of Morgan and sonquist titled AIDand was proposed by Breiman and his team in 1984. The CART algorithm accepts the numeric and the nominal data types as input and predicted variables, it can be used as a solution in classification and regression problems. CART decision tree algorithm As a unique dual from divided into a structure as branching criteria, CART tree is benefits from GiniIndex, without any stopping rule at the phase of its structuring, it is continually divided and grows. In the state where a new branching will not be realized, a cutting out from top in the direction of root is started. The most successful decision tree is subjected to assessment with test data independently selected after each cutting offs and efforts are made to make determinations [7].

V. METHODOLOGY

The following figure illustrates the Methodology uses for Performance Analysis of Decision Tree Algorithms.

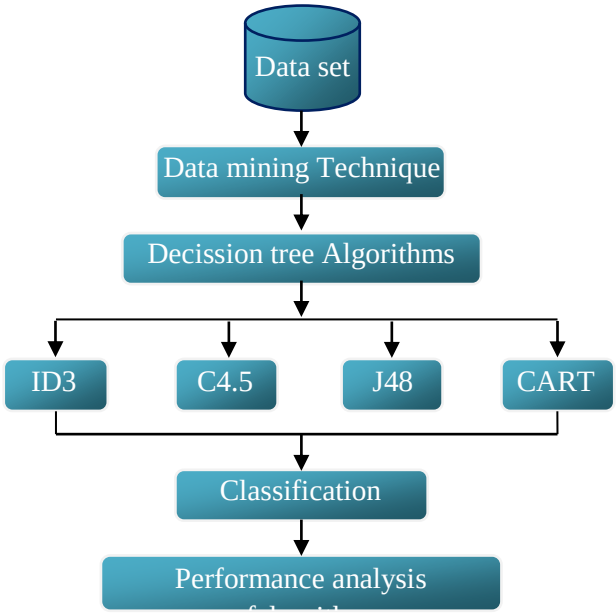

Figure 2: Methodology

VI. DECISION TREE INDUCTION

Decision tree is a tree in this each branch node represents a choice between number of alternatives and each leaf node represent a Decision. Decision trees are commonly used for gaining information for the purpose of Decision making. Every Decision tree starts with a root node which is for user to take actions. From this node, user split each node recursively according to Decision tree learning algorithms. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. The widely used decision tree learning algorithms are ID3,C4.5 and CART.

The Decision tree induction is a step of segmentation methodology. This acts as a tool for analyzing the large datasets, The response of analysis is predicted in the form of Tree structure. The Decision tree are classified using the two phases first one is Tree building phase and second one is tree pruning phase.

Tree building phase is the phase the training data is repeatedly partitioned until all the data in each partition belong to one class or the partition is sufficiently small. The form of the split depends on the type of attribute. Splitting for numeric attribute are of the form a<=S, where 'S' is a real number and the split for categorical attributes are of the form A€C, where C is a subset of all possible values of A. Alternate splits for attribute are compared using split index.

Different tests for choosing the best split are:

- Gini Index (Population Diversity)
- Entropy(Information Gain)
- Information Ratios

VII. DECISION TREE ALGORITHMS

Generate_decision Tree:

Generate a decision tree from the training tuples of data partition D.
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 lables with the class C;
4) if attribute_list is empty then
5) return N as a leaf node labled with the majority class in D;
6)apply attribute_selection_method(D,attribute_list) to find the “best” splitting _criterion;
7) lable 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
11) let Dj be the set of data tuples in D satisfying outcome j;
12) if Dj is empty then
13) attach a leaf labled with the majority class in D to node N;
14) else attach the node returned by Generate_decision_tree(Dj,attribute_list)to node N;
endfor
15) return N;

ID3 Algorithm

ID3 is simple Decision Tree learning algorithm, it is developed by Ross Quinlan Basic idea of ID3 algorithm is to construct the decision tree by employing a top down, The Greedy search through the given sets to test each attribute at every tree node in order to select the attribute which is most useful for classifying a given sets. statistical property called information gain is defined to measure the worth of the attribute.

Main steps in ID3 Algorithm:

1) Select the all attributes from different levels of decision tree nodes
2) Calculate the information gain for each and every attribute.

3) Use the information gain as the attribute selection criteria and select the attribute with the largest information gain to decide it the root node of the decision tree
4) Branches of the decision tree are calculate by the different information gain values of the nodes.
5) Build the decision tree nodes and branches recursively until a particular dataset of the instances belongs to the same group.

An Attribute Selection Measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of class-labeled training tuples into individuals classes. If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure. Conceptually, the "best" splitting criterion is the one that most closely results in such a scenario, Attribute selection measures are also known as splitting rules because they determines how the tuples at a given node are to be split.

Information Gain:

ID3 uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or information content of messages.

The expect information needed to classify a tuple in D is given by

$$\text{Info(D)} = -\sum_{i=1}^{m} pi \log 2(pi)$$

how much more information would we still need(after the partitioning) to arrive at an exact classification. This amount is measured by

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{Dj}{D} * \text{Info}(D_j)$$

Information gain is defined as the difference between the original information requirement and the new requirement.

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$

Gain Ratio:

The information gain measure is biased towards tests with many outcomes. that is it prefers to select attributes having a large number of values.

C4.5 as a successor of ID3 uses an extension to information gain known as Gain Ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a "split information" value defined analogously with Info(D).

The Gain Ratio is defined as

$$\text{Gain Ratio(A)} = \frac{Gain(A)}{SplitInfoA(D)}$$

Gini Index:

The Gini Index is used in CART. The Gini Index measures the impurity of D, a data partition or set of training tuples, as

$$\text{Gini(D)} = 1 - \sum_{i=1}^{m} p2i$$

The C4.5 Decision Tree Algorithm

The C4.5 Algorithm is a successor of ID3 that uses gain ratio as splitting criterion to partition the dataset. This algorithm applies a kind of normalization to information gain as a "split information" value.

C4.5 algorithm Pseudo code:

1) Check for the base case
2) Construct a DT using random training data
3) Find the attribute with the highest info gain(A_Best)
4) A_Best is assigned with entropy minimization
5) Partition S into S1,S2,S3...
6) According to the value of A_Best
7) Repeat the steps for S1,S2,S3
8) For each ti€D, apply the DT

The CART Decision Tree Algorithm

The algorithm CART stands for Classification And Regression Tree introduced by Brieman. it also based on Hunt's Algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values.

The CART algorithm uses Gini Index as an attribute selection measure to build a decision tree. Unlike ID3 and C4.5 algorithms, This CART algorithm produces binary spits. Gini Index measure does not use probabilistic assumptions like ID3,C4.5, CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

VIII. COMPARATIVE ANALYSIS OF DECISION TREE ALGORITHMS

Comparison parameter	ID3	C4.5	CART
Developed by	J.R.Quinlan	J.R.Quinlan	L.Breiman and team
Advantages	Easy to understand	Memory efficient than ID3	Handles missing values automatically
Dis advantages	Can suffer from over fitting	High training samples are needed	Poor modeling in a linear structure
Measure	Entropy information Gain	Entropy Iinformation Gain	Gini Diversity Index
Procedure	Topdown Decision tree construction	Topdown decision tree construction	Constructs binary decision tree
Pruning	Pre pruning using a single pass algorithm	Pre pruning using a single pass algorithm	Post pruning based on cost complexity measure
Approach	Greedy	Greedy	Greedy

IX. CONCLUSION & FUTURE WORK

Data mining is the process of extracting the knowledge from large amount of data. Data mining has various techniques. In this classification technique is one of the most important technique to classify the data. In this paper discuss about various Decision tree algorithm like ID3,C4.5,J48,CART, and also analyze these algorithms compare using various parameters like Advantages, Dis advantages, Measure, Procedure, Pruning, Approach. Based on this analysis, use the algorithms in coming research, and apply these algorithms on real datasets for finding useful patterns.

X. REFERENCES

[1] Pooja sharma, Divakar singh, Anju Singh "Classification Algorithms on A Large continuous Random Dataset using Rapid Miner Tool" IEEE 2nd International Conference on Electronics and Communication System-2015.

[2] Neelam Singhal, Mohd.Ashraf "Investigationof Effect of Reducing Dataset's size on Classification Algorithms" IEEE 2nd International Conference on Computing for Sustainable Global Development-2015.

[3] Neelam Singhal, Mohd.Ashraf "Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm" IEEE International Conference on Computing, Communication and Automation-2015 ISBN:978-1-4799-8890-7/15.

[4] Monika Gandhi, Dr.Shailendra , Narayan Singh "Prediction in Heart Disease Using Techniques of Data Mining" IEEE 1st International conference on futuristic trend in Computational analysis and Knowledge Management-2015

[5] Thiptanawat Phonghwattana, Worrawat Engchuan "Clustering-based Multi-class Classification of Complex Disease" IEEE-978-1-4799-6049-1/15-2015

[6] AlbertoRos, Mahdad Davari, Stefanos Kaxiras "Hierarchical Private/shared Classification: The Key to Simple and Efficient Coherence for Clustered Cache Hierachies"IEEE-978-1-4799-8930-0/15-2015.

[7] Kuizhi Mei, Jinye Peng, Ling Gao, Naiquan Zheng Jianping Fan "Hierarchical Classification of Large-scale patient Records for automatic Treatment Stratification"IEEE Journal of Biomedical and Health Informatics., VOL.19.NO.4,JULY 2015.

[8] Sudeep D.Thepade , Madhura M.Kalbhor "Extended Performance appraise of Bayes,Function,Lazy,Rule,Tree Data mining Classiifier in Novel Transformed Fractional Content Based Image Classification" IEEE International Conference on Pervasive Computing(ICPC)-2015.

[9] Asli Calis, Ahmet Boyaci "Data Mining Applications in banking sector with clustering and classification methods"IEEE International Conference on Industrial Engineering and Operations Management, March3-5,2015.

[10] Sneha Chandra, Maneet Kaur "Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the field of Medical Data Mining." IEEE 2nd International Conference on Computing for sutainable Global Development 2015