ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

REVIEW ON DIFFERENT CLASSIFICATION TECHNIQUES IN DATA MINING

Er. Ekta[*1] AND Er. Rama Rani[2]
[*1,2]Department of Computer Science and Engineering, DAV University, Jalandhar, Punjab India

ABSTRACT

Data mining is the process of discovering or extracting new patterns from large data sets involving methods from statistics and artificial intelligence. The Decision Tree is an important classification method in data mining classification. The classification algorithm of the decision tree ID3, C4.5,CART algorithms have the merits of high classifying speed, strong learning ability and simple construction. The classification accuracy of CART is higher when compared to ID3 and C4.5. all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory This paper provides focus on the a variety of algorithms of Decision tree their characteristic, Merits and demerits.

Keywords- *Classification and regression trees (CART), ID3(Iterative Dichotomiser 3),C4.5.*

I. INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Decision tree algorithms have been widely used to establish classification models which are close to human reasoning and comprehensible. Now-a-days the data stored in a database and which is used for application is huge. This explosive growth in data and database has generated an urgent need for new techniques and tools that can intelligently automatically transform the processed data into useful information and knowledge. Hence data mining has become a research are with increasing importance. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classis or to predict future data trends. Such analysis can help provide us with better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. Many classification and prediction methods have been proposed by researcher in machine learning pattern recognition and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data.
The Decision trees are classified using the two phases:
1. Tree Building Phase
2. Tree Pruning Phase

In classification, the cases are placed in differing groups. The procedures behind this methodology create rules as per training and testing individual cases. A number of algorithms have been developed for classification based data mining. Some of them include decision tree, k-Nearest Neighbor, Bayesian and Neural-Net based classifiers. At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure. Quinlan in 1979 put forward a well-known Iterative Dichotomiser 3 algorithm, which is the most widely used algorithm in decision tree. But that algorithm has a defect of tending to select attributes with many values. It has also problem of over classification which leads to have less accuracy .

II. RELATED WORK

In 2008, Xindong Wuet al. [1] This paper presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, *k*-Means, SVM, Apriori, EM, Page Rank, , KNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, we provide a description of the algorithm, discuss the impact of the algorithm, and
Review current and further research on the algorithm. These 10 algorithms cover classification,

ISSN 2348 – 8034
Impact Factor- 3.155

In 2013, Leonard Gordon et al.[2] Classification and regression trees (CART) - a non-parametric methodology- were first introduced by Breiman and colleagues in 1984. CARTs are underused (especially in public health) and they have the ability to divide populations into meaningful subgroups which will allow the identification of groups of interest and enhance the provision of products and services accordingly. They provide a simple yet powerful analysis. It is hoped that their value is demonstrated and this will enhance their increased use in data analysis.

In 2011, S.L. Ting et al.[3] the Naïve Bayes content classifier has been broadly utilized because of its effortlessness as a part of both the preparation and characterizing stage. The purpose of this paper is to highlight the execution of utilizing Naïve Bayes in document classification.

In 2012, Mary Slocum et al.[4] examined diverse calculations utilized for creating a decision making (or predictive analysis) system. There are calculations for making decision trees, for example, J48 alongside algorithms for deciding known nearest neighbor (KNN) or grouping when working on pattern recognition. The objective of this paper is to take a one specific decision tree algorithm called Iterative Dichotomiser 3 (ID3)

In 2012, Jaimin N. Undavia et al. [5] Decision Tree is the most prominent classification algorithm in Data Mining. It can be implemented through various algorithms like J48, Simple CART / CART, Random Tree, ID3 etc depending on the types of data and requirements of applications. Here in this paper we have compared different algorithms of decision tree induction on the basis of different parameters. In this paper, we have selected three algorithms for comparison which are theoretically better as compare to other decision tree induction algorithm. We compare these three algorithms using open source data mining tool Weka & present the result. We found that J48 classification algorithm works better for the Prediction of Post-Graduation Course than other Decision Tree Induction classification algorithms.

In 2013, Anuja Priyam et al.[6] At the present time, the amount of data stored in educational database is increasing swiftly. These databases contain hidden information for improvement of student's performance. Classification of data objects is a data mining and knowledge management technique used in grouping similar data objects together. There are many classification algorithms available in literature but decision tree is the most commonly used because of its ease of execution and easier to understand compared to other classification algorithms. The ID3, C4.5 and CART decision tree algorithms former applied on the data of students to predict their performance. But all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory.

III. CLASSIFICATION TECHNIQUES

a. ID3 (Iterative Dichotomiser 3)

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to theC4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) or information gain IG(A) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute (e.g. age < 50, 50 <= age < 100, age >= 100) to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

STEPS:

- Calculate the entropy of every attribute using the data set S.
- Split the set S into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum).
- Make a decision tree node containing that attribute
- Recurse on subsets using remaining attributes.

ISSN 2348 – 8034
Impact Factor- 3.155

b. C4.5 Technique

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S= s1,s2……. of already classified samples. Each sample S_i consists of a p-dimensional vecto$(x_{1,i}, x_{2,i}, ..., x_{p,i})$, where the x_j represent attributes or features of the sample, as well as the class in which S_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

- Permit numeric attributes.
- Deal sensibly with missing values.
- Pruning to deal with for noisy data

ID3 and C4.5 ar algorithms introduced by Quinlan for causation Classification Models, additionally referred to as decision Trees, from data. ID3 works on separate values solely

Table 1: ID3 Uses only Discrete range

Attributes	Possible Values
Age	New, Middle, Old
Competition	Yes, No
Type	Hardware, Software

Table 2: C4.5 uses different attribute range.

Attributes	Possible Values
Outlook	Sunny, overcast, Rain
Temperature	Continuous
Humidity	Continuous
Windy	True, False

C4.5 Steps:

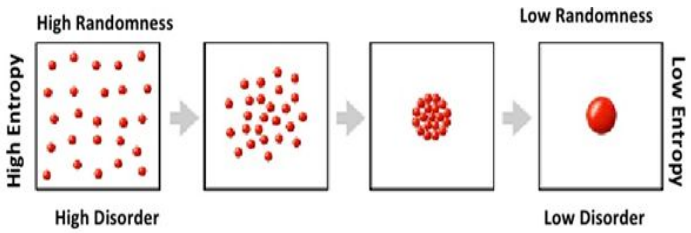
- Choose attribute for root node
- Create branch for each value of that attribute
- Split cases according to branches
- Repeat process for each branch until all cases in the branch have the same class

With the highest gain ratio root node will be selected.
Gain ratio can be calculated with the help of this formula.

Gain ratio(A)

The subsequent formula for calculating the entrophy

ISSN 2348 – 8034
Impact Factor- 3.155



Entrophy is measure of impurity. The formula for gain attribute is:

c. CART Technigue

CARTs are not as popular compared to traditional statistical methods because of the lack of tests to evaluate the goodness of fit of the tree produced and the relatively short span that they have been around. They are typically model free in their implementation. Howbeit, a model based statistic is sometimes used for a splitting criterion. The main idea of a classification tree is a statistician's version of the popular twenty questions game. Several questions are asked with the aim of answering a particular research question at hand. However, they are advantageous because of their non-parametric and non-linear nature. They do not make any distribution assumptions and treat the data generation process as unknown and do not require a functional form for the predictors. They also do not assume additivity of the predictors which allows them to identify complex interactions

STEPS:

- Tests in CART are always binary, but C4.5 allows two or more outcomes.
- CART uses the Gini diversity index to rank tests, whereas C4.5 uses information-based Criteria.
- CART prunes trees using a cost-complexity model whose parameters are estimated by Cross-validation; C4.5 uses a single-pass algorithm derived from binomial confidence limits.
- CART looks for surrogate tests that approximate the outcomes when the tested attribute has an unknown value, but C4.5 apportions the case probabilistically among the outcomes.

IV. COMPARISON OF DIFFERENT CLASSIFICATION ALGORITHMS

	ID3	C4.5	C5.0	CART
Type of data	Categorical	Continuous and Categorical	Continuous and Categorical, dates, times, timestamps	continuous and nominal attributes data
Speed	Low	Faster than ID3	Highest	Average
Pruning	No	Pre-pruning	Pre-pruning	Post pruning
Boosting	Not supported	Not supported	Supported	Supported
Missing Values	Can't deal with	Can't deal with	Can deal with	Can deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Same as C4.5	Use Gini diversity index

V. CONCLUSION

This paper utilizes classification and regression trees (CART) and demonstrates their usefulness for data analysis.C4.5 is the best algorithm for small datasets among all the three because it provides better accuracy and efficiency than the other algorithms. The main disadvantages of serial decision tree algorithm (ID3, C4.5 and CART) are low classification accuracy when the training data is large. But all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory.

REFERNCES

1. *Leonard Gordon,"Using Classification and Regression Trees (CART) in SAS® Enterprise MinerTM For Applications in Public Health",2013*
2. *S. B. Kotsiantis," Decision trees: a recent overview", © Springer Science+Business Media B.V. 2011*
3. *Jon Crowcroft , Michael Segal , Liron Levin," Improved structures for data collection in static and mobile wireless sensor Networks", © Springer Science+Business Media New York 2014*
4. *Rupali Bhardwaj , Sonia Vatta," Implementation of ID3 Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering.*
5. *Deyi Sun, Wing Cheong Lau," Social Relationship Classification based on Interaction Data from Smartphones",IEEE 2nd international workshop on hot topics in pervasive computing 2013.*
6. *Aziz Nasridinov & Sun-Young Ihm & Young-Ho Park," A hybrid construction of a decision tree for multimedia contents", Springer Science+Business Media New York 2013*
7. *H. Trevor, T. Robert, and F. Jerome, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer, Second Edition, 2009.*
8. *N. Matthew, and G. Sajjan, "Comparative Analysis of Serial Decision Tree Classification Algorithms.", IJCSS,vol. 3(3), pp.230-240, 2009.*
9. *W.-Y. L. Y.-S. S. TJEN-SIEN LIM, "A Comparison of Prediction Accuracy,Complexity, and Training Time of Thirty-three Old and New Classi," Kluwer Academic Publishers, Boston., 2000.*
10. J.R Quinlan, "Induction of Decision Trees Machine Learning". Vol.1, pp81-106, 1986.