A Survey on Decision Tree Algorithm For Classification

[1]Mr. Brijain R Patel, [2]Mr. Kushik K Rana
[1]Department of computer engineering, GEC Modasa, India
[2]Assistant Professor, Department of computer engineering, GEC Modasa, India
[1]patelbrijain808@gmail.com

***Abstract* - Data mining is the process of discovering or extracting new patterns from large data sets involving methods from statistics and artificial intelligence. Classification and prediction are the techniques used to make out important data classes and predict probable trend .The Decision Tree is an important classification method in data mining classification. It is commonly used in marketing, surveillance, fraud detection, scientific discovery. As the classical algorithm of the decision tree ID3, C4.5, C5.0 algorithms have the merits of high classifying speed, strong learning ability and simple construction. However, these algorithms are also unsatisfactory in practical application. When using it to classify, there does exists the problem of inclining to choose attribute which have more values, and overlooking attributes which have less values. This paper provides focus on the various algorithms of Decision tree their characteristic, challenges, advantage and disadvantage.**

***Keywords*- Decision tree algorithms, ID3, C4.5, C5.0, classification techniques**

I. INTRODUCTION

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.

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 [1].

Several challenges are facing researchers and developer. Therefore, several papers and articles have tried to cover these issues. Such as Improvement on ID3 Algorithm[2], Re optimization of ID3, C4.5 and C5.0 algorithm [3].

II. DECISION TREE ALGORITHMS

Researchers have developed various decision tree algorithms over a period of time with enhancement in performance and ability to handle various types of data. Some important algorithms are discussed below.

CHID: CHAID (CHi-squared Automatic Interaction Detector) is a fundamental decision tree learning algorithm. It was developed by Gordon V Kass [4] in 1980. CHAID is easy to interpret, easy to handle and can be used for classification and detection of interaction between variables. CHID is an extension of the AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures. It works on principal of adjusted significance testing. After detection of interaction between variables it selects the best attribute for splitting the node which made a child node as a collection of homogeneous values of the selected attribute. The method can handle missing values. It does not imply any pruning method.

CART: Classification and regression tree (CART) proposed by Breiman *et al.* [5] constructs binary trees which is also refer as Hierarchical Optimal Discriminate Analysis (HODA). CART is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. The word binary implies that a node in a decision tree can only be split into two groups. CART uses gini index as impurity measure for selecting attribute. The attribute with the largest reduction in impurity is used for splitting the node's records. CART accepts data with numerical or categorical values and also handles missing attribute values. It uses cost-complexity pruning and also generate regression trees.

ID3: ID3 (Iterative Dichotomiser 3) decision tree algorithm is developed by Quinlan [6]. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this

way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification.

C4.5: 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 [7]. C4.5 algorithm uses information gain as splitting criteria. It can accept data with categorical or numerical values. To handle continuous values it generates threshold and then divides attributes with values above the threshold and values equal to or below the threshold. C4.5algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5.

C5.0/Sec 5: C5.0 algorithm is an extension of C4.5 algorithm which is also extension of ID3. It is the classification algorithm which applies in big data set. It is better than C4.5 on the speed, memory and the efficiency. C5.0 model works by splitting the sample based on the field that provides the maximum information gain. The C5.0 model can split samples on basis of the biggest information gain field. The sample subset that is get from the former split will be split afterward. The process will continue until the sample subset cannot be split and is usually according to another field. Finally, examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be rejected. C5.0 is easily handled the multi value attribute and missing attribute from data set [8].

Hunt's Algorithm: Hunt's algorithm generates a Decision tree by top-down or divides and conquers approach. The sample/row data contains more than one class, use an attribute test to split the data into smaller subsets. Hunt's algorithm maintains optimal split for every stage according to some threshold value as greedy fashion [9].

III. THE APPLICATIONS OF DECISION TREES IN VARIOUS AREAS

The decision tree algorithms are largely used in all area of real life. The application areas are listed below

- **Business:** Decision trees are use in visualization of probabilistic business models. It also use in customer relationship management and uses for credit scoring for credit card users.
- **Intrusion Detection**: Decision trees use for generate genetic algorithms to automatically generate rules for an intrusion detection expert system. Abbes et al. proposed protocol analysis in intrusion detection using decision tree.
- **Energy Modeling**: Energy modeling for buildings is one of the important tasks in building design. Decision tree is use for it.
- **E-Commerce**: Decision tree is widely use in e-commerce. It is use for generated online catalog which is essence for the success of an e-commerce web site.
- **Image Processing**: Park et al. [10] proposed perceptual grouping of 3-D features in aerial image using decision tree classifier. Macarthur et al. [11] proposed use of decision tree in content-based image retrieval.
- **Medicine**: Medical research and practice are the important areas of application for decision tree techniques. Decision tree is most useful in diagnostics of various diseases. It also use for Heart sound diagnosis.

Table 1 Comparisons between different Decision Tree Algorithm

	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

- **Industry**: Production quality control (faults identification), non-destructive tests are areas where decision tree algorithm is useful.
- **Intelligent Vehicles:** The job of finding the lane boundaries of the road is important task in development of intelligent vehicles. Gonzalez and Ozguner [12] proposed lane detection for intelligent vehicles by using decision tree.
- **Remote Sensing**: Remote sensing is a strong application area for pattern recognition work with decision trees. Some researcher proposed algorithm for classification for land cover categories in remote sensing. And binary tree with genetic algorithm for land cover classification.
- **Web Applications** Chen et al. [13] presented a decision tree learning approach to diagnosing failures in large Internet sites. Bonchi et al. [14] proposed decision trees for intelligent web caching.

IV. DECISION TREE LEARNING SOFTWARE AND COMMONLY USED DATASET

Thousand of decision tree software are available for researchers to work in data mining. Some software are use for the analysis of data and some are use for commonly used data sets for decision tree learning are discussed below.

- **WEKA:** WEKA (Waikato Environment for Knowledge Analysis) workbench is set of different data mining tools developed by machine learning group at University of Waikato, New Zealand [15]. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces for easy access to this functionality. WEKA supported versions are windows, Linux and MAC operating systems. It provides various associations, classification and clustering algorithms. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes. WEKA also provides pre-processors like filters and attributes selection algorithms. WEKA provides J48. In J48 we can construct trees with EBP, REP and unpruned trees.
- **GATree:** GATree (Genetically Evolved Decision Trees) is use genetic algorithms to directly evolve classification decision trees [16]. Instead of using binary strings, it adopts a natural representation of the problem using binary tree structures. The evaluation version of GATree is available on request to the authors. Here we can set various parameters like generations, populations, crossover and mutation probability etc. to generate decision trees.
- **Alice d'ISoft:** Alice d'ISoft software for Data Mining by decision tree is a powerful and inviting tool that allows the creation of segmentation models. It makes it possible for the business user to explore data on line interactively and directly. It works on windows operating system. The evaluation version of Alice d'ISoft is available on request to the authors [17].
- **See5/C5.0:** See5/C5.0 has been designed to analyze substantial databases containing thousands to millions of records and tens to hundreds of numeric, time, date, or nominal fields. See5/C5.0 also takes advantage of computers with up to eight cores in one or more CPUs (including Intel Hyper-Threading) to speed up the analysis. It is easy to use and does not presume any special knowledge of Statistics or Machine Learning. It is available for Windows Xp/Vista/7/8 and Linux [18].

Table 2 Widely Used datasets in decision tree research

No.	Data Set Name	No. of Instances	No. of Attributes	Description
1	Balance Scale	625	4	psychological experiments Data set
2	Kr vs Kp	3196	36	Chess data
3	Glass	24	10	Glass identification data
4	Abalone	4177	8	Marine Resources Dataset
5	Heart Disease	303	75	Health Resource Dataset
6	Image segmentation	2310	19	Image related data
7	Breast Cancer	286	9	Health Resource Dataset
8	Protein Secondary Structure	128	58	Study for protein sequence
9	Labor	57	16	Labor agreement and negotiation
10	Bank Marketing	45211	17	Finance Related dataset
11	Credit Approval	690	15	Finance credit card related dataset
12	Segment	2310	19	Hand segment image data

- **OC1:** OC1 (Oblique Classifier 1) is a decision tree induction system originally designed by S. K. Murthy [19], designed for applications where the instances have numeric (continuous) feature values. OC1 builds decision trees that contain linear combinations of one or more attributes at each internal node; these trees then partition the space of examples with both oblique and axis-parallel hyper planes [20]. OC1 has been used for classification of data representing diverse problem domains, including astronomy, gene.

Commonly Used Data set

Some of the largely used datasets by researchers are shown in Table 2 with number of records and attributes in each data set and short description of data. Some of the data sets may not be available at present.

V. SOME OF THE RECENT ISSUES RELATED TO DECISION TREE

- **Fragmentation problem [21, 22]**
  The fragmentation problem exists if data is gradually partitioned into smaller segments. Replication and repetition leads to fragmentation but it can occur without both of them if many features need be tested.
- **Replication problem [21, 22]**
  The replication problem can be observed if sub-trees are replicated in decision tree. It causes the data to be partitioned into the smaller segments which leads to fragments problem.

- **Partitioning in continuous data [23]**
  The partitioning is the considerable issue in decision tree algorithm. Attributes having discrete values can be easily partitioned but continuous attributes like age have problem in partitioning while decision tree construction.
- **Repetition problem [21, 22]**
  The repetition (or repeated testing) problems is present if the features are repeatedly (more than once) tested along a path in a decision tree. These repetitions split data into smaller and smaller segments, hence result in fragmentation.
- **ID3, 4.5 tends to take multi valued attributes [24, 25]**
  It always selected the attribute with many values. The algorithm with many values wasn't the correct one, and it created wrong classification. Sometimes gain ratio of some non-valuable attribute which is having so many values (because of the formula of the information gain and entropy) becomes highest and some valuable attribute's gain ratio becomes lower. So difficulty would come of making the root of the tree.
- **ID3 algorithm does not backtrack in searching [24]**
  Whenever certain layer of the tree chooses a property to test, it will not backtrack to reconsider this choice. In this way, algorithm could easily converged local optimal answer, but not global optimal answer.
- **Decision tree algorithm does not provide incremental learning**
  ID3 algorithm is a kind of greedy algorithm. For incremental learning task, ID3 algorithm could not accept training sample incrementally, so the each increase of example requires abandoning original decision tree, to restructure new decision tree, and to cause lots of overhead.
- **Handling of range inputs [23]**
  Sometimes dataset may have range input attributes and present decision tree building methods, namely mean substitute, min-max substitute and mean-extent substitute, may not be suitable. If attributes like income, the values would be range inputs so feature selection problem would be there. We need to improve the membership grade and entropy calculation method.
- **XOR Parity and Multiplexer Problem [26]**
  Decision trees do not express them easily in some mathematical relations such as XOR, Parity Check and multiplexer problem. In such cases, the decision tree becomes prohibitively large. Approaches to solve the problem involve either changing the representation of the problem domain (known as propositionalisation) or using learning algorithms based on more expressive representations such as statistical relation or inductive logic programming.
- **Over fitting Decision Tree [27]**
  Decision-tree learners can create over-complex trees that do not generalize the data well, which is called over fitting. Mechanisms such as pruning setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.

VI. CONCLUSION

According to our observations, the performances of the algorithms are strongly depends on the entropy, information gain and the features of the data sets. There are various work has been done using the Decision tree Algorithm. But they all are like Static in Nature. Some recent improve algorithm reduce problem like replication, handle continuous data, biased to multi value attribute. This paper provides students and researcher some basic fundamental information about decision tree, tools and recent issues.

REFERENCES

[1] Jiawei Han And Micheline Kamber,Data Mining Concept and Techniques, Copyright 2006, Second Edition.
[2] Chen Jin, Luo De-lin and mu Fen-xiang An improve ID3 Decision tree algorithm. IEEE 4th International Conference on computer Science & Education.
[3] Devashish Thaku, Nisarga Makandaiah and Sharan Raj D (2010). Re Optimization of ID3 and C4.5 Decision tree. IEEE Computer & Communication Technology.
[4] Gordan.V.Kass(1980). An exploratory Technique for inverstigation large quantities of categorical dataApplied Statics, vol 29, No .2, pp. 119-127.
[5] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.
[6] Quinlan J. R. (1986). Induction of decision trees. *Machine Learning,* Vol.1-1, pp. 81-106.
[7] Zhu Xiaoliang, Wang Jian YanHongcan and Wu Shangzhuo(2009) Research and application of the improved algorithm C4.5 on decision tree.
[8] Prof. Nilima Patil and Prof. Rekha Lathi(2012), Comparison of C5.0 & CART Classification algorithms using pruning technique.
[9] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
[10] Kyu Park, Kyoung Mu Lee and Sang Uk Lee (1999). Perceptual grouping of 3D features in aerial image using decision tree classifier. In Proc. of 1999 International Conference on Image Processing, Vol. 1, pp. 31 - 35.
[11] Sean D. MacArthur, Carla E. Brodley, Avinash C. Kak and Lynn S. Broderick (2002). Interactive content-based image retrieval using relevance feedback. Computer Vision and Image Understanding, pp. 55-75.
[12] Juan Pablo Gonzalez and U. Ozguner (2000). Lane detection using histogram-based segmentation and decision trees. Proc. of IEEE Intelligent Transportation Systems.

[13] M. Chen, A. Zheng, J. Lloyd, M. Jordan and E. Brewer (2004). Failure diagnosis using decision trees. *Proc. of the* International Conference on Autonomic Computing.
[14] Francesco Bonchi, Giannotti, G. Manco, C. Renso, M. Nanni, D. Pedreschi and S. Ruggieri (2001). Data mining for intelligent web caching. Proc. of International Conference on Information Technology: Coding and computing, 2001, pp. 599 - 603.
[15] Ian H. Witten; Eibe Frank, Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition".
[16] A. Papagelis and D. Kalles (2000). GATree: Genetically evolved decision trees. Proc. 12th International Conference On Tools With Artificial Intelligence, pp. 203-206.
[17] ELOMAA, T. (1996) Tools and Techniques for Decision Tree Learning.
[18] R. Quinlan (2004). Data Mining Tools See5 and C5.0 Rulequest Research (1997).
[19] S. K. Murthy, S. Salzberg, S. Kasif And R. Beigel (1993). OC1: Randomized induction of oblique decision trees. In Proc. Eleventh National Conference on Artificial Intelligence, Washington, DC, 11-15th, July 1993. AAAI Press, pp. 322-327.
[20] Dipak V. Patil and R. S. Bichkar (2012). Issues in Optimization of Decision Tree Learning:
[21] Rudy Setiono and Huan Liu. Fragmentation problem and Automated Feature Constructions.
[22] Zheng Yao, Peng Liu, Lei and Junjie Yin (2005) R-C4.5 Decision Tree Model and its Applications to Health Care Dataset.
[23] Han Jing-ti and Gu Yu-jia (2009) Study on Handling Range Inputs Methods On C4.5 algorithm. IEEE International Forum on Computer Science – Technology and Application.
[24] Devashish Thakur, Nisarga Makandaiah and Sharan Raj D (2010). Re Optimization of ID3 and C4.5 Decision tree. IEEE Computer & Communication Technology.
[25] C Rui min, Wang Mio(2010) A more efficient Classification scheme for ID3. IEEE 2th International Conference on computer Science & Education.
[26] Horvath, Tamas; Yamamoto, Akihiro, eds. (2003). Inductive Logic Programming Lecture Notes in Computer Science 2835.
[27] Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan et al. "Top 10 algorithms in data mining." Knowledge and Information Systems 14, no. 1 (2008): 1-37.