International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

# A Survey on Decision Tree Algorithms of Classification in Data Mining

**Himani Sharma[1], Sunil Kumar[2]**

[1] M.Tech Student, Department of computer Science, SRM University, Chennai, India

[2] Assistant Professor, Department of computer Science, SRM University, Chennai, India

**Abstract**: ***As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.***

**Keywords:** Decision Tree Learning, classification, C4.5, CART, ID3

## 1. Introduction

In order to discover useful knowledge which is desired by the decision maker, the data miner applies data mining algorithms to the data obtained from data collector. The privacy issues coming with the data mining operations are twofold. If personal information can be directly observed in the data, privacy of the original data owner (i.e. the data provider) will be compromised. On the other hand, equipping with the many powerful data mining techniques, the data miner is able to find out various kinds of information underlying the data. Sometimes the data mining results reveals sensitive information about the data owners. As the data miner gets the already modified data so here the objective was to show the comparative performance between already used classification method and the new method introduced. As previous studies shows that the ensemble techniques provide better results than the decision tree method thus the desired result was inspired thru this concern.

### 1.1 Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch represents an outcome of the test, class label is represented by each leaf node (or terminal node) . Given a tuple X, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for the tuple. It is easy to convert decision trees into classification rules. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees, in this tree structure, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision tree can be constructed relatively fast compared to other methods of classification. SQL statements can be constructed from tree that can be used to access databases efficiently. Decision tree classifiers obtain similar or better accuracy when compared with other classification methods.

A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Classification is one of the most frequently. The C4.5, ID3, CART decision tree are applied on the data of students to predict their performance. These algorithms are explained below-

## 2. ID3 Algorithm

Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. It is serially implemented and based on Hunt‟s algorithm. The basic idea of ID3 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 generally used to determine suitable property for each node of a generated decision tree. Therefore, 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. So, the use of this property for partitioning the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduced to a minimum. Hence, the use of an information theory approach will effectively reduce the required dividing number of object classification.

## 3. C4.5 Algorithm

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. As splitting criteria, C4.5 algorithm uses information gain. It can accept data with categorical or numerical values. Threshold is generated to handle continuous values and then divide attributes with values above the threshold and values equal to or below the threshold. C4.5 algorithm can easily handle missing values, as missing attribute values are not utilized in gain calculations by C4.5.

### 3.1 The algorithm C4.5 has following advantages:

- Handling each attribute with different cost.
- Handling training data set with missing attribute values- C4.5 allows attribute values to be marked as „?" for missing. Missing attribute values are not used in gain and entropy calculations.
- Handling both continuous and discrete attributes- to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Pruning trees after creation- C4.5 goes back through the tree once it has been created, and attempts to remove branches that are not needed, by replacing them with leaf nodes.

### 3.2 C4.5's tree-construction algorithm differs in several respects from CART, for instance

- Tests in CART are always binary, but C4.5 allows two or more outcomes.
- CART uses Gini index to rank tests, whereas C4.5 uses information-based criteria.
- CART prunes trees with a cost-complexity model whose parameters are estimated by cross-validation, whereas C4.5 uses a single-pass algorithm derived from binomial confidence limits.
- This brief discussion has not mentioned what happens when some of a case's values are unknown.

CART looks for surrogate tests that approximate the outcomes when the tested attribute has an unknown value, on the other hand C4.5 apportions the case probabilistically among the outcomes.

## 4. CART Algorithm

It stands for Classification And Regression Trees. It was introduced by Breiman in 1984. It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. CART also based on Hunt"s algorithm and can be implemented serially. Gini index is used as splitting measure in selecting the splitting attribute. CART is different from other Hunt"s based algorithm because it is also use for regression analysis with the help of the regression trees. The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. CARTS supports continuous and nominal attribute data and have average speed of processing.

### 4.1 CART Advantages

1) Non parametric (no probabilistic assumptions)
2) Automatically perform variable selection
3) Use any combination of continuous or discrete variables
   - Very nice feature: ability to automatically bin massively categorical variables into a few categories.
4) Zip code, business class, make/model.
5) Establish "interactions" among variables
   - Good for "rules" search
   - Hybrid GLM-CART models

**Table 1:** Comparisons between different Decision Tree Algorithms

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

## 5. Decision Tree Learning Software

Some softwares are used for the analysis of data and some are used 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. For easy access to this functionality, it contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces. WEKA supported versions are windows, Linux and MAC operating systems and it providens 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). It also provides pre-processors like attributes selection algorithms and filters. WEKA provides J48. In J48 we can construct trees with EBP, REP and unpruned trees.

**GATree:** GATree (Genetically Evolved Decision Trees) use genetic algorithms to directly evolve classification decision trees. Instead of using binary strings, it adopts a natural representation of the problem by using binary tree structure. On request to the authors, the evaluation version of GATree is now available. To generate decision trees, we can set various parameters like generations, populations, crossover and mutation probability etc.

**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. For the business user, this software makes it possible to explore data on line interactively and directly. Alice d'ISoft software works on windows operating system. And the evaluation version of Alice d'ISoft is available on request to the authors.

**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. It takes advantage of computers with up to eight cores in one or more CPUs (including Intel Hyper-Threading) to speed up the analysis. See5/C5.0 is easy to use and does not presume any special knowledge of Statistics /Machine Learning. It is available for Windows Xp/Vista/7/8 and Linux.

## 6. Applications Of Decision Trees In Different Areas Of Data Mining

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, used in customer relationship management and used for credit scoring for credit card users.

**Intrusion Detection**: Decision trees is used to 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**: Decision tree is used for energy modeling. Energy modeling for buildings is one of the important tasks in building design.

**E-Commerce**: Decision tree is widely use in e-commerce, use to generate online catalog which is essence for the success of an e-commerce web site.

**Image Processing**: Perceptual grouping of 3-D features in aerial image using decision tree classifier.

**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.and also use for Heart sound diagnosis.

**Industry**: decision tree algorithm is useful in production quality control (faults identification), non-destructive tests areas.

**Intelligent Vehicles:** The job of finding the lane boundaries of the road is important task in development of intelligent vehicles. Gonzalez and Ozguner 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, binary tree with genetic algorithm for land cover classification.

**Web Applications** Chen et al presented a decision tree learning approach to diagnosing failures in large Internet sites. Bonchi et al proposed decision trees for intelligent web caching.

## 7. Conclusion

This paper studied various decision tree algorithms. Each algorithm has got its own pros and cons as given in this paper. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. This paper provides students and researcher some basic fundamental information about decision tree algorithms, tools and applications.

## References

[1] Anju Rathee, Robin prakash mathur, "Survey on Decision Tree classification algorithm for the evaluation of student performance" International Journal of Computers & Technology, Volume 4 No. 2, March-April, 2013, ISSN 2277-3061
[2] S.Anupama Kumar and Dr. Vijayalakshmi M.N. (2011) "Efficiency of decision trees in predicting student"s academic performance", D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335-343, 2011.
[3] Devi Prasad bhukya and S. Ramachandram " Decision tree induction- An Approach for data classification using AVL –Tree", International journal of computer and electrical engineering, Vol. 2, no. 4, August, 2010.
[4] Jiawei Han and Micheline Kamber Data Mining: Concepts and Techniques, 2ndedition.
[5] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
[6] Quinlan, J.R., C4.5 -- Programs For Machine Learning.Morgan Kaufmann Publishers, San Francisco, Ca, 1993.
[7] Introdution To Data Mining By Tan, Steinbach, Kumar.
[8] Mr. Brijain R Patel, Mr. Kushik K Rana, "ASurvey on Decision Tree Algorithm for Classification", © 2014 IJEDR, Volume 2, Issue 1.
[9] Prof. Nilima Patil and Prof. Rekha Lathi(2012), Comparison of C5.0 & CART Classification algorithms using pruning technique.
[10] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
[11] Neha Midha and Dr. Vikram Singh, "A Survey on Classification Techniques in Data Minng", IJCSMS (International Journal of Computer Science & Management Studies) Vol. 16, Issue 01, Publishing Month: July 2015.
[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: