ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

A Survey on Comparative Study of Decision Tree Methods in Data Mining

Maria Nithi M[1], Jeya Priya U[2]

PG Scholar, Department of Computer Science, Stella Maris College (Autonomous), Chennai, India[1]

Assistant Professor, Department of Computer Science, Stella Maris College (Autonomous), Chennai, India[2]

ABSTRACT: Data mining is the process of identifying hidden information patterns of data based on different aspects for arranging and to make helpful data. These useful data are stored in common house which is known as data warehouse, for efficient analysis and data mining algorithms, decision makings and other data requirements to eliminate cost and increase revenue. There are several major data mining techniques that includes association, classification, clustering, prediction, sequential patterns and decision tree. A decision tree is a graphical depiction of a decision and every potential outcome of making that decision. Decision trees give people an effective and easy way to understand the potential options of a decision and its range of possible outcomes. The tree is structured to show how and why one choice may lead to the next, with the use of the branches indicating each option is mutually exclusive. In this paper we have discussed about various decision tree algorithms like CART, ID3, C4.5, CHAID algorithms and the performance analysis of algorithms.

KEYWORDS: data mining, decision tree, performance of CART, ID3, C4.5, CHAID.

I. INTRODUCTION

Data mining is a process used to turn noisy raw data into useful data. Data mining is based on effective collection of data warehousing and computer processing. Decision tree algorithm is a kind of data mining model to make induction learning algorithm based on examples. A decision tree is a representation of graph of solutions to decisions which are based on certain conditions. Hence it is known as decision tree which starts with single idea and end up with number of decisions or solutions like branches of tree. Decision tree is very helpful because it makes us to take decision. Only limitation is we can able to select the solution with in the alternatives. Decision trees are used for formalizing the brainstorming process so that we can identify more potential solutions.

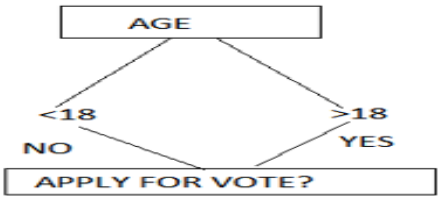

Fig1: Decision tree

II. DECISION TREE ALGORITHMS

2.1 CART:

The abbreviation of CART is Classification and regression tree which was proposed by Bremen. It is used to construct binary trees which are also referred to as Hierarchical Optimal Discriminate Analysis (HODA). CART is a nonparametric decision tree technique which produces classification or regression trees depending on whether the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

variable dependent is numeric or categorical respectively. The word binary tree shows that the node in the decision tree is split into two groups. For selecting the attributes CART uses Gini index as the impurity measure. CART handles missing attributes values and also accepts numerical and categorical values. CART uses cost-complexity pruning and also generate regression trees. [2].It is based on Hunt's Algorithm.

2.1.2 CART ALGORITHM

1. Find each feature's best split.
2. For each feature with K different values there exist K-1 possible splits.
3. Find the split, which maximizes the splitting criterion. The resulting set of splits contains best splits.
4. Find the node's best split. Among the best splits from Step i find the one, which maximizes the splitting criterion.
5. Split the node using best node split from Step ii and repeat from Step i until stopping criterion is satisfied. [13]

2.1.3 ADVANTAGE

- Both numerical and categorical variables can be easily handled by CART
- CART is nonparametric and therefore does not rely on data belonging to a particular type of distribution.
- CART can easily handle outliers. [6]

2.1.4 DISADVANTAGE

- CART may have unstable decision tree.
- Splitting variables and values may change **[6]**
- CART splits into only one variable

2.2 ID3

ID3 is the expansion of Iterative Dichotomies 3 .It is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. It is implemented based on Hunt's algorithm. The ID3 algorithm is used to construct the decision tree by a top-down, and greedy search through the sets to test each attribute at every tree node. Metric information gain is introduced to classify a given set. By minimizing the question asked we can find an optimal way of classifying learning set. Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function. ID3 uses information gain measure to choose the splitting attribute. It accepts only categorical attributes in building a tree model. It does not give accurate result. Thus an intensive pre-processing of data is carried out before building a decision tree model with ID3. [4]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

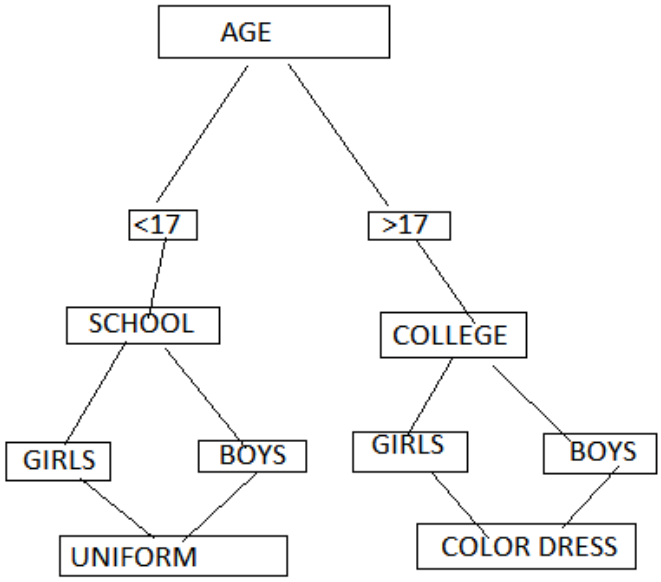

FIG 2: ID3 decision tree

2.2.1 Main steps in ID3 Algorithm:

1) Select the attributes from different levels of decision tree nodes
2) Calculate the information gain for every attribute.
3) Using the information gain as the attribute selection criteria, select the attribute with the largest information gain to decide the root node of the decision tree
4) The different information gain values of the node are used to calculate Branches of the decision tree.
5) Build the decision tree nodes and branches recursively until a particular dataset the instances belong to the same group [11].

2.2.2 ADVANTAGES

- The training data create the understandable prediction rules
- Builds the fastest tree.
- Builds a short tree.
- Only to test enough attributes until all data is classified.
- Finding leaf nodes makes test data to be pruned, and reduces number of tests.
- To create data complete full dataset is searched [6]

2.2.3 DISADVANTAGES

- Even if the small sample is tested data may be over fitted or over classified.
- To make a decision only one attribute is tested at the time.
- Does not handle numeric attributes and missing values. [6]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

2.3 C4.5

This algorithm was proposed in 1993 by Ross Quinlan.The improved version of ID3 is C4.5.the algorithm C4.5 uses the gain ratio for splitting criteria in growth phase. Hence C4.5 is an evolution of ID3. Both continuous and discrete attributes handles by C4.5. C4.5 develop threshold and split it into two whose attributes value is above value of threshold or equal to it then it will handle continuous attributes. Similar to ID3 the data is sorted at every nodes of the tree to determine the best splitting attributes [9].C4.5 is used with datasets The C4.5 can be used for classification hence C4.5 is often referred to as a statistical classifier. Whose patterns have unknown attribute values this algorithm deals with attributes values which can be used even if they are marked as missing but theses attribute values are not used in gain and entropy calculations it also has an enhanced method of tree pruning which decreases misclassification errors caused by noise or abundant details available in training dataset.

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.

2.3.1 C4.5 ALGORITHM

1. Check for the common value cases.
2. Identify the more common values and split it
3. Label it the more common values
4. Set that label as decision node
5. Repeat the steps from 1 to 4 and create the decision tree.

2.3.2 ADVANTAGES

1. Easily interpreted model is build
2. Implementation is easy
3. Categorical and continues values can be used
4. Deals with noise

2.3.3DISADVANTAGES

1. The decision tree goes different if there is a small variation.
2. It does not work well and small training set.

2.4 CHAID

The expansion of CHAID is Chi-squared automatic interaction detector. It was proposed by kass in 1980 and it is the oldest classification method. CHAID algorithm build non binary tree which is based on simple algorithm and suited for large dataset and also CHAID algorithm gives multiple tables. CHAID is the recursive partitioning method. CHAID algorithm is applied for both classification and regression type. CHAID is mainly used for the prediction and for classification and interaction between variables. The response rate is low in CHAID algorithm. 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.

2.4.1 CHAID ALGORITHM

1. Cross tabulate the response variable with each of the explanatory variables.
2. This is applied to each table with more than 2 columns. Compute Pearson X2 tests for independence for each allowable suitable.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

3. Allows categories combined at step 2 to be broken apart.
4. You have now completed the optimal" combining of categories for each explanatory variable.
5. Use the "most signicant"variable in step 4 to split the node with respect to the merged" categories for that variable

2.4.2 ADVANTAGES

1. CHAID approach is very fast.
2. CHAID is popular in market research it builds "wider" decision trees because it is not constrained to make binary splits.
3. CHAID gives many terminal nodes which is connected to a single branch, which can be summarized in a simple two-way contingency table, with multiple categories for each variable.

2.4.3 DISADVANTAGES

1. The algorithm requires large qualities of data to split fragments of variables into small sub ranges and get expected results.
2. The multiple splits are hard to relate to real business conditions because the CHAID tree may be unrealistically short and uninteresting.
3. Real variables are forced into categorical bins before analysis, which may not be helpful, particularly if the order in the values should be preserved.

III. REAL WORLD APPLICATIONS OF DECISION TREE

- **Agriculture:** Application of a range of machine learning methods to problems in agriculture and horticulture is described in [11].
- **Biomedical Engineering:** Use of decision trees for identifying features to be used in implantable devices can be found in [12].
- **Control Systems:** Automatic induction of decision tree was recently used for control of nonlinear dynamical systems [13].
- **Pharmacology:** Use of tree based classification for drug analysis can be found in [14].
- **Physics:** Decision trees have been used for the detection of physical particles [15].
- **Energy Modeling**: Energy modeling is used to build designs it is very important task while building designs using decision trees.
- **Business:** Decision trees are used in the visualization of probabilistic business models. It is also used in the customer relationship management and for credit scoring for credit card users.
- **Intrusion Detection**: Decision trees are used to generate genetic algorithms to automatically generate rules for an intrusion detection expert system.
- **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. It is used to generate online catalog which is essential for the success of an e-commerce web site[2]

IV. COMPARATIVE STUDY ON ID3 AND CART

Table 1: comparison between different decision tree algorithm

Algorithm(↓)	Developed by	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Techniques	Tree types	Speed	Outlier Detection
ID3	Quinlan	Information	Handles	Do not	No pruning is	Top down	Classification	Low	Susceptible

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

	Ross	Gain	only Categorical value	handle missing values.	done		and regression		on outliers
CART	Bremen .	Towing Criteria	Handles both Categorical and Numeric value	Handle missing values.	Cost-Complexity pruning is used	Non parametric	Classification and regression	Average	Can handle outliers
C4.5	Ross Quinlan	Make use of split information and gain ratio	Handle continuous and categorical values	Do not deals with missing value	Post and Pre pruning	Top down decision tree	Classification	Faster than ID3	Susceptible on outliers
CHAID	Kass	No splitting	Both continuous and categorical values	Handles missing values	No pruning method	Non parametric	Regression analysis	Low	-

The above table shows the difference between the two algorithms ID3 and CART It shows the splitting criteria for two algorithms, the attribute type, missing values, pruning strategy and the outlier detection.[6]CART can easily handle both numerical and categorical variables. CART algorithm will itself identify the most significant variables and eliminate no significant ones. CART can easily handle outliers. The ID3 algorithm is considered as a very simple decision tree algorithm (Quinlan, 1983). The ID3 algorithm is a decision-tree building algorithm. It determines the classification of objects by testing the values of the properties. It builds a decision tree for the given data in a top-down fashion, starting from a set of objects. The algorithm c4.5 was developed by Ross Quinlan and it makes use of split information and gain ratio and it also handles continuous and categorical values will not deals with missing values. The CHAID algorithm was by kass it also handles missing values here the speed is low.

V. CONCLUSION

In this paper we have survived the various basic properties of the decision tree algorithms .We can apply them on different types of data sets having different types of values and properties and can attain a best result by knowing that which algorithm will give the better result on a specific type of data set.

REFERENCES

[1] David Bowser-Chao and Debra L. Dzialo, "Comparison of the Use of Binary Decision Trees and Neural Networks in Top Quark Detection", Physical Review D: Particles And Fields, 47(5):1900, 1993
[2] K. J. Hunt, "Classification by Induction: Applications to Modeling and Control of Non-Linear Dynamic Systems", Intelligent Systems Engineering, 2(4):231--245, 1993.
[3] W. J. Gibb, D. M. Auslander, And J. C. Griffin. Selection Of Myocardial Electro Gram Features For Use By Implantable Devices. Ieee Transactions On Biomedical Engineering, 40(8):727--735, August 1993.
[4] K.T. Dago, R. Luthringer, R. Lengelle, G. Rinaudo, and J. P. Matcher," Statistical Decision Tree: A Tool for Studying Pharmaco-Eeg Effects of Cns-Active Drugs", Neuropsychobiology, 29(2):91--96, 1994.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

[5] R.J. Mcqueen, S. R. Garner, C.G. Nevill-Manning, and I.H. Witten, "Applying Machine Learning to Agricultural Data", Computers and Electronics in Agriculture, 12(4):275--293, 1995
[6] Priyama , Abhijeet , Rahul Gupta , Anju Ratheeb and Saurabh Srivastava, " Comparative Analysis of Decision Tree Classification Algorithms", Computer Science & Engineering, Kanpur Institute Of Technology, Vol.3, No.2, 2013.
[7] Brijain R Patel, Kushik K Rana, "A Survey On Decision Tree Algorithm For Classification Of Computer Engineering", Ijedr, Vol 2, Issue 1, ISSN: 2321-9939, 2014.
[8] Venkatadri.M, Lokanatha C. Reddy, " A Comparative Study On Decision Tree Classification Algorithms In Data Mining", Dravidian University, Vol 2, Issue 2, Issn: 0974-3596, 2010.
[9] S. Naga Parameshwara Chary, Dr .B .Rama, "A Survey on Comparative Analysis Of Decision Tree Algorithms In Data Mining", Kakatiya University, Vol 3, Issue.1, 2017.
[10] Purva Sewaiwar, Kamal Kant Verma , "Comparative Study Of Various Decision Tree Classification Algorithm Using Weka",International Journal Of Emerging Research In Management &Technology, Issn: 2278-9359, Vol-4, Issue-10, 2015.
[11] Brinjain R Patel , Kushik K Rana," A Survey On Decision Tree For Classification ",Issn:2321-9939,Vol 2,Issue:1,2014
[12] Sonia Singh, Manoj Giri, " Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey", International Journal Of Advanced Information Science And Technology, ISSN: 2319:2682, Vol.3, 2014.
[13] Breiman L, "Classification and regression trees" The Wadsworth and Brooks-Cole statistics probability series. Chapman & Hall , 1984.