ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 5, May 2024

Role of Decision Tree in Data Mining: A Qualitative Study of Decision Tree Based Classification Algorithms

[1] Nikhil Somani, [2] Sunita Choudhary
[1] M.Tech Student, Department of CSE, BTU, Bikaner, Rajasthan, India
[2] Assistant Professor, Department of CSE, UCET, Bikaner, Rajasthan, India
Corresponding Author Email: [1] nikhilsomani805@gmail.com, [2] sunitadangi@gmail.com

Abstract*— After the Internet Era, today we have a billions of data generating everyday by millions of users and thus being having this amount of data, it is neccessary to make useful outcome from this. This can be done by making data analysis and data extraction. Data mining is the method of finding out some useful outcome from an incomplete, random or noisy data records. Decision Tree algorithms make data mining easier by calculating useful outcome from tree including its root, branches, and nodes. Nodes represents attribute, a branch defines an outcome while leaf nodes defines a class label. This research paper provides a qualitative comparison of commonly used decsion tree algorithms used for data mining with its tools and applications.

Keywords: Decision Tree Algorithm, c4.5, c5.0, id3, cart, weka.

I. INTRODUCTION

A decision maker makes use of data mining algorithms to obtain certain useful outcomes or information from a dataset. But these data mining operations generate two folded privacy issues. Data miners can observe personal information in the data and thus compromise the privacy of individual data owners. Sometimes, some very sensitive information can be observed by applying various data mining operations and thus the privacy of data owners are kept in compromise. Thus the main idea behind this research paper is to provide a qualitative performance comparison of various classification methods.

II. DECISION TREE

A regular tree has its roots, branches and leaves. This same is applied on decision trees as well. A decision tree has a root node, connecting branches followed by its leaves node. The topmost node of tree, called as root node is basically a parent to all the nodes of the tree. In a decision tree, every node indicates an attribute or feature, a branch indicates a rule or decision and each leaf defines an outcome of the decision tree. The outcome from the leaf node can be categorical or continuous. Decision trees can be defined as the strategic impression of human thinking over a large data set to make some useful data interpretations. The main aim of using decision tree is to represent the entire data over a tree with the leaf indicating the outcome.

Formally, a decision tree is designed as a flowchart like tree shaped structure, formed by its internal nodes, braches, and leaves node representing an attribute, an outcome, and class label respectively. For classification of data (say for a given tuple T), the decision tree is used as a predictive model for mapping an item's observations to its target value. Path tracing is made from the root node of the tree to terminal nodes.

Decision trees are relatively faster and efficient compared to other classification methods. Decision trees can be used to access databases using SQL statements constructed from the tree. A comparative study of common decision tree are explained here:

III. DECISION TREE ALGORITHM

3.1 ID3

Iterative Dichotomiser 3 is based on the famous Hunt's algorithm and was introduced by Ross Quijnlan. It is a very simple decision tree algorithm and applied serially. This algorithm works in a top-down fashion with a greedy search approach within the given data sets and checks each attribute at node points of the tree. Identification of individual property of a node is more essential in decision trees. This is used as a testing measure for a node leading to develop a classification for highest information gain at top level levelling a degree of subset from maximum to minimum. Thus the information gain approach is useful in proper classification of data nodes.

3.2 C4.5

This algorithm was also introduced by Ross Quinlan. It is an enhancement to the ID3 algorithm. This algorithm is used to obtain a decision tree which is used for classification and thus also called a statistical classifier. Information gain is used in this algorithm for splitting of node points. This algorithm works on categorical or numerical values with a threshold value. Threshold value is used for splitting and thus handles the missing values easily as it does not consider it.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 5, May 2024

3.3 C5.0

This algorithm too was introduced by Ross Quinlan. It is an enhanced version of C 4.5 and works on categorical data as well as numerical data. It provides a binary tree or multi branches tree as an outcome. Information gain approach is used for splitting and for handling the missing values, it allows whether to use missing values or not.

3.4 CART 5.0

Breiman had introduced "Classification and Regression Trees" abbreviated as CART. This algorithm makes use of both classification and regression trees. It makes use of binary splitting and provides a binary tree as an outcome. It is implemented serially and can also be used for regression analysis and it makes use of Gini index for splitting attributes. Both continuous and nominal attributes can be used in CART. Regression analysis technique is used to perform the predictive analysis for dependent attributes.

Table 1. Splitting Metric and Formulas

Splitting Metric	Equation
İnformation Gain	$I(p,n) = \left(\frac{-p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{n+p}\right)\log_2\left(\frac{n}{p+n}\right)$
Gain Index	G = $Gini\ Index, G = \left(\frac{1}{2n^2\mu}\right)\sum_{j=1}^{m}\sum_{k=1}^{m} n_j n_k \lvert y_j - y_k \rvert$
Gini İndex	= Subtraction of Infomation data before splitting operation and informaton data after splitting operation = I(p,n) - E(A)

IV. DECISION TREE LEARNING DATASET AND TOOLS

There are various tools and learning dataset, some of the major dataset and tools are defined here:

4.1 WEKA

"Waikato Environment for Knowledge Analysis" or "WEKA" can be defined as a collection of various data mining tools. This is basically a workbench providing access to various visualization tools, predictive modelling and data analysis tools with GUI. WEKA supports almost all operating systems including linux, windows and MAC. Collection of algorithms in this workbench includes classification, clustering and association algorithms. WEKA assumes that the input file is a file in which data points are represented by fixed attributes (numerical or nominal attributes).

Table 2. Qualitative Comparison of Decision Tree Algorithms

Decision Tree Algorithm	Attribute Type	Splitting Metric	Splitting Method	Missing Value Handling	Outlier Detection	Speed and Performance	Common Tool
ID3	Only Categorical	Information Gain	No Restriction	NO	Susceptible	low	WEKA
C4.5	Categorical & Numerical Data	Information Gain	No Restriction	YES	Susceptible	Faster than ID3	WEKA
C5.0	Categorical & Numerical Data	Information Gain	No Restriction	YES	Susceptible	Faster than ID3	Sec5/C5.0
CART	Continuous & Nominal Data	Gini Index	Binary Split	YES	Can handle	Average	CART 5.0

4.2 See5 / C5.0

This tool is used for analysis of larger databases with nominal or numerical records. This tool makes use of the cpu core for speedy analysis. It is very easy to use with GUI buttons. It dont pre-requisite any knowledge of statistics or maths or machine learning. It supports both windows and linux operating systems. Outcome of this software will be in C programming language and thus can be easily integrated with any machine.

V. CONCLUSION

The purpose of this research was to provide a qualitative comparison of various decision tree algorithms. It can be concluded that selection of a right decision tree algorithm depends on various factors like data size, processor availability, data type and others. Two major things - accuracy and time taken, can be considered for their efficiency. This research paper presents a basic fundamental accumulation of information about various decision tree algorithms, their qualitative comparison, along with its applications and tools used to perform analysis.

REFERENCES

[1] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991

[2] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D. (2004), An introduction to decision tree modeling. J. Chemometrics, 18: 275-285.

[3] Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, *3*(2), 334-337.

[4] Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. In *2009 4th International Conference on Computer Science & Education* (pp. 127-130). IEEE.

[5] Li, L., & Zhang, X. (2010, June). Study of data mining algorithm based on decision tree. In *2010 International Conference On Computer Design and Applications* (Vol. 1, pp. V1-155). IEEE.

[6] Brijain, M., Patel, R., Kushik, M. R., & Rana, K. (2014). A survey on decision tree algorithm for classification.

[7] Upadhayay, A., Shukla, S., & Kumar, S. (2013). Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set. *International Journal of Computer Science & Communication Networks*, *3*(1), 64.

[8] Yang, J., Zhang, N. N., LI, J., LIU, Y. M., & LIANG, M. H. (2010). Research and application of decision tree algorithm. *Computer Technology and Development*, *2*.

[9] Yang, C. C., Prasher, S. O., Enright, P., Madramootoo, C., Burgess, M., Goel, P. K., & Callum, I. (2003). Application of decision tree technology for image classification using remote sensing data. *Agricultural Systems*, *76*(3), 1101-1117

[10] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.