ISSN: 2834-4928

Research Article

Journal of Electrical Electronics Engineering

Examining and Evaluating Classification Algorithms Based on Decision Trees

Mehr Ali Qasimi*

IT Engineering Department, Technology Faculty, Selçuk University, Konya, Turkey.

***Corresponding Author**
Mehr Ali Qasimi, IT Engineering Department, Technology Faculty, Selçuk University, Konya, Turkey

Submitted: 2024, Feb 08; **Accepted:** 2024, Mar 12; **Published:** 2024, Apr 18

Citation: Ali Qasimi, M. (2024). Examining and Evaluating Classification Algorithms Based on Decision Trees. *J Electrical Electron Eng, 3*(2), 01-07.

Abstract
Machine learning learns everything from the data we provide it, uses that data to predict future outcomes, and more. Machine Learning is the process by which computer work more accurately as it learns from the given data. The adoption of machine learning techniques is beneficial in all fields of research. There are many types of machine learning include unsupervised learning, enforcement learning, and supervised learning. Classification is a part or type of supervised machine learning technique. These algorithms are used to identify and predict data in a variety of domains, including replacement statistical methods, search engine fields, and fields with medical certifications. A supervised learning technique called classification assigns a data item to one of several pre-established classifications. One of the most popular approaches for representing classifiers in data classification is the use of decision tree classifiers. A decision tree classification is a machine learning technique that predicts or determines the classes of future data sets when the class labels are unknown by using the predetermined labels from previous known sets. Decision tree classifiers have been suggested for usage in many different disciplines, including medical disease analysis, text categorization, user smartphone classification, pictures, and many more. Numerous decision tree algorithms exist, and they are categorized according to how accurate and costly they are to use.

Keywords: Machine Learning, Decision Trees, Classification, Algorithm.

1. Introduction

Today's technology is very advanced, particularly in the area of machine learning (ML), which helps to reduce the amount of labor that needs to be done by humans. Over the past three decades, machine learning has become more and more important in a number of areas [1]. Machine learning (ML) combines computer science and statistics to create algorithms for artificial intelligence that become increasingly effective when exposed to pertinent data instead of being given explicit instructions. ML is the study of computing methods that are automatically improved by experience, in addition to speech recognition, picture identification, text localization, etc. As a subset of artificial intelligence, it is recognized. ML algorithms build a model population based on a sample, known as "training data," which is ordered to generate prediction or judgment without being expressly designed to do so. in a variety of fields, including computer vision and email screening [2].

The purpose of classification is to make the most accurate prediction of the target class. The classification algorithm determines the relationship between the training process's input and output [3]. Huge volumes of data are collected in data mining environments. Using the decision tree method is best if the data set is correctly classified and has the fewest possible nodes. Assigning objects to categories with a wide range of applications is the work of classification.

Classification is a machine learning method which places objects in a collection into desired classes or categories. Predicting the target class for each occurrence in the data is the goal of classification. A classification model, for example can be used to categorize bank loan applications as safe or dangerous [4].

Classification uses attributes to forecast data instances. The process of classifying future data into known classes is known as classification. Typically, this methodology employs a training dataset to construct a model and a test dataset to verify its accuracy. Decision trees, Naïve Bayes, logistic regression, and other methods are commonly used in classification [5].

One of the most popular and useful techniques for inductive inference over supervised data is decision tree learning. A decision tree is a procedural representation of a classification process for categorical data based on many attributes. In addition, decision trees are useful in machine learning because they can process vast amounts of data [6]. There is no need for subject expertise or parameter setup when building decision trees. Hence, decision trees are both adequate and suitable for the exploratory process of discovering new information, and their treebased representation of learned information is clear and simple to comprehend [4].

2. Decision Tree

Decision trees are tree structures that resemble flowcharts, with each internal node representing a test on an attribute, each branch representing a test result, and each leaf node (or terminal node) representing a class label. The decision tree is used to compare the attribute values of a given tuple, X. From the root node to a leaf node that contains the tuple's class prediction, a path is traced [7]. Translating decision trees into categorization rules is a simple process. Using a decision tree as a predictive model, decision tree learning links observations about an item to inferences about its intended value. It is among the methods for predictive modeling that are applied in data mining, machine learning, and statistics [8].

A distinct advantage of using decision trees (DT) is that provides the availability to the employer to conduct both supervised and unsupervised learning. Therefore, they are commonly used for knowledge discovery [9].

A typical tree has leaves, branches, and roots. Decision Tree follows the same structure. It is made up of leaf, branch, and root nodes. Every internal node is used for attribute testing; the branch is used for the test's conclusion, and the leaf node is used for the class label [10]. As its name implies, a root node is the highest node in a tree and is the parent of all other nodes. A decision tree is a tree in which every leaf represents an outcome (continuous or categorical value), every link (branch) represents a choice (rule), and every node represents a feature (attribute). Decision trees make it easy to gather data and produce insightful interpretations since they closely resemble human thought processes [11].

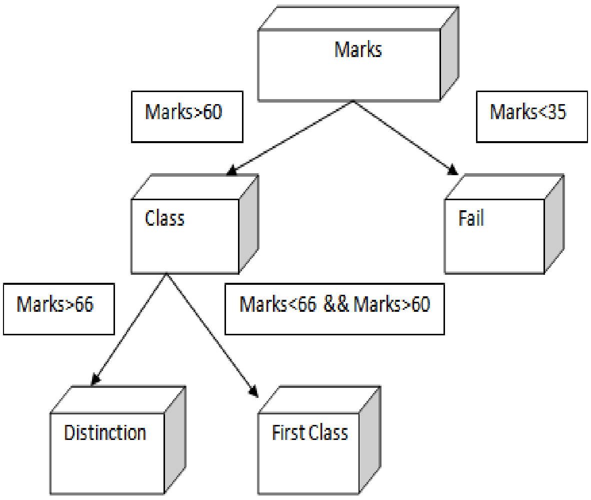

Figure 1: Example of Decision Tree [12].

In above example simple decision tree is used in student database. Here we can easily classify the different categories of student based on their result. Hence we obtain different classes of student and also can easily get the count for the number of students in each class.

3. Decision Tree Algorithms

The attributes are divided using decision tree algorithms so that they may be tested at any node to see if splitting is "Best" for each class. Since the splitting criterion for each branch must be the same, the resulting partitioned is as PURE as possible [13].

There are various decision trees algorithms namely ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and Regression Tree), CHAID (CHi- squared Automatic Interaction Detector), MARS. Out of these, we will be discussing the more popular ones which are ID3, C4.5, CART [4].

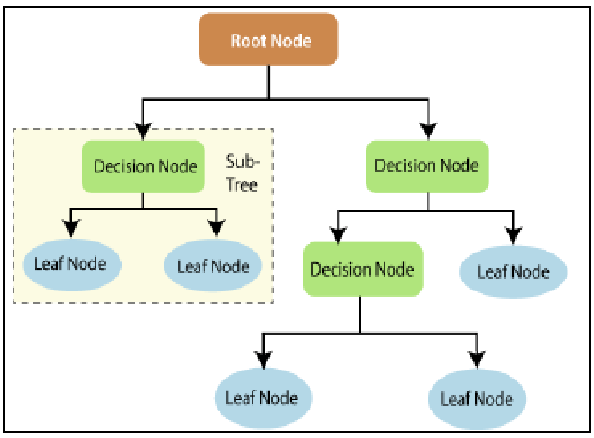

Figure 2: Structure of Decision Tree [14].

3.1. ID3 Algorithm

Quinlan Ross first presented the straightforward decision tree learning algorithm Iterative Dichotomiser 3 in 1986. It utilizes Hunt's algorithm and is serially implemented. The fundamental idea behind the ID3 approach is to build the decision tree by testing each characteristic at each node of the tree using a top-down, greedy search over the supplied sets. The information gain technique is typically employed in the decision tree method to identify appropriate properties for each node of a decision tree that is constructed. Consequently, we can designate as the current node's test attribute the attribute with the biggest information gain (entropy reduction at the maximum level) [15]. This will result in the least amount of information required to classify the training sample subset that was obtained through subsequent partitioning. Therefore, the mixture degree of various types for all generated sample subsets will be minimized when this property is used to divide the sample set included in the present node. Therefore, the number of divisions needed for object categorization will be effectively decreased by using an information theory technique [11].

ID3 is an algorithm for supervised learning. Through a series of training examples from multiple classes, it is explicitly explained. Based on the hypothesis it develops, it forecasts an item's class. ID3 looks for characteristics (or attributes) that set one class of samples apart from another. ID3 demands that every feature be well-behaved, meaning that all potential values are known ahead of time, and that every feature be known beforehand [16].

ID3 only takes categorical attributes in order to construct a decision tree model. When there is noise and when ID3 is implemented serially, accurate results are not obtained. Thus, prior to building a decision tree, data is preprocessed. Information gain is determined for each attribute in order to construct a decision tree, with the attribute with the largest information gain serving as the root node. Arcs indicate the remaining possible values. Next, all potential outcome instances are analyzed to determine whether or not they are members of the same class. Instances of the same class are identified by a single name class; instances of different classes are categorized using splitting attributes [4].

3.2. C4.5

The decision tree-generating algorithm C4.5 was created by Ross Quinlan. The older ID3 algorithm by Quinlan is expanded upon in C4.5. Because C4.5 can produce decision trees that are useful for classification, the program is frequently referred to as a statistical classifier [11]. The C4.5 algorithm uses information gain as a splitting criterion. Both numerical and category data can be entered into it. In order to manage continuous values, a threshold is created. Attributes with values above the threshold and values equal to or below the threshold are then divided. Since the C4.5 methods does not use missing attribute values in gain calculations, it can handle missing values with ease [17].

It uses an improved technique for pruning trees, which lowers misclassification errors caused by noise and excessive information in the training set. To find the ideal splitting attribute, the data is sorted at each node of the tree, just like in ID3. The splitting attribute is assessed using the gain ratio impurity approach [18].

In C4.5, a tree grows in three stages [19].

- C4.5 uses a technique akin to ID3 algorithms for dividing categorical attributes. Binary splits are always produced by continuous attributes.
- Picking the attribute that has the maximum gain ratio.
- These procedures are repeatedly applied to newly formed tree branches, and the tree's growth is halted upon verification of the stop criterion. Increased information biases the property with more values. Therefore, C4.5 use the less biased selection criterion of Gain Ratio.

3.3. CART

The acronym CART represents Classification and Regression Trees. Breiman introduced it in 1984. It creates regression trees as well as classifications. Binary splitting of the characteristics forms the foundation of the CART classification tree generation process. CART is a serializable algorithm that is also based on Hunt's algorithm [20]. When choosing the dividing attribute, the splitting measure utilized is the Gini index. In contrast to other Hunt-based algorithms, CART is capable of doing regression analysis using regression trees. When predicting a dependent variable over a specified time period, a collection of predictor variables is provided. This is done using the regression analysis function. CARTS has an average processing speed and can handle nominal and continuous attribute data [11].

Unlike other Hunt-based algorithms, CART utilizes regression trees to facilitate regression analysis. Regression analysis is a tool that helps predict a dependent variable over a certain time period given a set of predictor factors [18].

The CART method functions as a binary tree, with each internal node having precisely two outbound edges. The produced tree is pruned using Cost-Complexity Pruning, and the splits are chosen according on the Towing Condition. The capability of CART to produce regression trees is a crucial feature [12].

Tree Algorithms [11]

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

Table 1: Comparisons between different Decision

4. Metrics

The training data are divided into multiple subgroups based on the values of the splitting property. The algorithm continues iteratively until every instance in a subset in every Decision Tree belongs to the same class [21].

Metrics	Equation
Information Gain	$Information\ Gain = I(p,n) = \left(\frac{-p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{n+p}\right)\log_2\left(\frac{n}{p+n}\right)$
Gain Ratio	Gain Ratio=I(p,n)-E(A) I(p,n)= Information before splitting E(A)= Information after splitting
Gini Index	$Gini\ Index, G = \left(\frac{1}{2n^2\mu}\right)\sum_{j=1}^{m}\sum_{k=1}^{m} n_j n_k \lvert y_j - y_k \rvert$

Table 2. Splitting Criteria [13]

The primary disadvantage of information gain is that it is skewed toward multivariate characteristics. When data is split unevenly and one of the child nodes has more entries than the other, the gain ratio typically favors that. When there are more than two categories in the data set, the Gini Index produces negative results. These are the disadvantages of dividing standards [13].

5. Evaluation Mechanism

A set is considered precise if the values are near to one another. The set is considered accurate if its average closely matches the actual value of the quantity being measured. One can only measure more than two terms if they are provided with a set of data points from many measurements of the same quantity [22].

$$\boxed{\textbf{Accuracy} =} \ \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$\boxed{\textbf{Precision} =} \ \frac{(TP)}{(TP + FP)}$$

TP = True positive, TN = True Negative

FP = False Positive, FN = False Negative

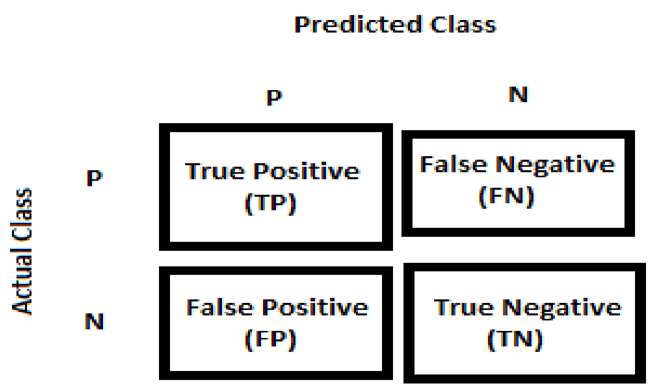

Figure 3. Confusion Matrix sample in Decision Tree

6. Dataset Description

The automobile dataset is the one utilized in this experiment. By using this dataset with the ID3, C4.5, and CART decision tree algorithms. The description of the dataset is as follows.

There are two sections to the automobile dataset. Car acceptability is one, while technical characteristic is the other. Two aspects of car acceptability are total cost (purchasing) and maintenance cost (maintenance). The quantity of doors (doors), the number of people the automobile can hold (people), the size of the luggage boot (lug boot), and an assessment of the safety of the vehicle (safety).

Number of Instances: 1728
Number of Attributes: 6
Missing Attributes Value: None

Attributes Value:

Attribute	Attribute Values
Buying	v-high, high, med, low
Maint	v-high, high, med, low
Doors	2, 3, 4, 5-more
Persons	2, 4, more
Lug-boot	small, med, big
Safety	low, med, high

Class Distribution (Number of instances per class):

Class	N	N [%]
Unacc	1210	70.023%
Acc	384	22.222%
good	69	3.993%
v–good	65	3.762%

Experiment

With the WEKA tool, the experiment is simulated. WEKA is a collection of machine learning algorithms for data mining jobs. Weka has tools for preprocessing data, classifying data, regressing data, clustering data, associative rules, and visualizing data [23]. Weka is distributed under the GNU General Public License, making it open source software. It works well for creating novel machine learning systems as well. The algorithms can be called from your own Java code or applied straight to a dataset [11].

Algorithm	Attribute Type	Missing Value	Pruning Strategy	Outlier Detection
ID3	Only categorical values	No	No	Susceptible to outlier
CART	Categorical and Numerical both	Yes	Cost complexity pruning is used	Can handle
C4.5	Categorical and Numerical both	Yes	Error based pruning is used	Susceptible to outlier

Table 3. Theoretical results [11]

This paper distributes the identical data sets on three distinct decision tree methods, such as ID3, C4.5, and CART, for the experiment. The results of all three algorithms in the terms time and accuracy with the help of the outcome from the below table [24]. The algorithm's division to get a better result is described in the splitting Criteria column. Details regarding the kinds of values the algorithm can handle are provided in the attribute type column. The algorithm's accuracy is determined by the result obtained from the Missing Value column, which indicates whether or not the algorithm detects the missing value.

Algorithm	Time Taken (Seconds)	Accuracy (%)	Precision
ID3	0.02	89.35	0.964
CART	0.5	97.11	0.972
C4.5	0.06	92.36	0.924

Table 4. Practical results

As we can see the above table is the practical result of three algorithms ID3, C4.5, and CART. One can notice that CART takes 0.5 seconds to execute an algorithm, ID3 takes 0.02 seconds and C4.5 takes 0.06 seconds. The slowest execution is of CART and fastest is ID3 [11].

Though CART takes too much time or we can say it is the slowest one among them, accuracy is highest and it gives very precise result than the other algorithms which are ID3 and C4.5. So, we can conclude from the above table that if we do the comparative study of all three algorithms, the CART is best to choose [11].

7. Conclusion

An overview of machine learning, regression, and classification methods is given in this paper. Our attention has mostly been on and decision trees and its different algorithms. The ID3 technique is the simplest and most effective for classifying large datasets. The dataset was subjected to the Decision Tree algorithms ID3 C4.5 and CART. When it comes to precision, accuracy, and time, decision trees perform better than others. The recommendation system is mostly responsible for identifying engaging resources. After a thorough investigation into decision tree algorithms, this research comes to the conclusion that CART is the most exact and accurate algorithm for the given dataset.

Future Work

This will be put on the Apache server and made available online in the future [25]. The online rating will be used for the forecast, and the datasets are updated on a regular basis. To evaluate the system's performance, the prediction methods can also be tested on other datasets [11].

References

1. Sathiyanarayanan, P., Pavithra, S., Saranya, M. S., & Makeswari, M. (2019). Identification of breast cancer using the decision tree algorithm. In *2019 IEEE International conference on system, computation, automation and networking (ICSCAN)* (pp. 1-6). IEEE.
2. 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.
3. Ahmed, N. S., & Sadiq, M. H. (2018). Clarify of the random forest algorithm in an educational field. In *2018 international conference on advanced science and engineering (ICOASE)* (pp. 179-184). IEEE.
4. Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications, 163*(8), 15-19.

5. Prakash, A., & Solanki, V. K. (2023, February). Performance Analysis of Classification Algorithms. In *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022* (pp. 647-656). Singapore: Springer Nature Singapore.
6. Havrda, J., & Charvát, F. (1967). Quantification method of classification processes. Concept of structural $ a $-entropy. *Kybernetika, 3*(1), 30-35.
7. Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification.
8. Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications (Vol. 81). World scientific.
9. Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications, 4*(2), 13-19.
10. Gershman, A., Meisels, A., Lüke, K. H., Rokach, L., Schclar, A., & Sturm, A. (2010). A decision tree based recommender system. 10th International Conferenceon Innovative Internet Community Systems (I2CS)–Jubilee Edition 2010.
11. Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR), 5*(4), 2094-2097
12. M. Somvanshi. (2016). “Iccubea”.
13. Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering, 6*(10), 74-78.
14. X. D. (2016). Crystallography.
15. Chahal, H. (2013). ID3 modification and implementation in data mining. *International Journal of Computer Applications, 975*, 8887.
16. Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. In *2011 IEEE control and system graduate research colloquium* (pp. 37-42). IEEE.
17. Talaat, A. A. (2017). “Comparison between Traditional and Modern Models of Data Mining,” *Curresweb.Com*, pp. 243–258.
18. 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.
19. Mahajan, N., & Kaur, B. P. (2016). Analysis of factors of road traffic accidents using enhanced decision tree algorithm. *International Journal of Computer Applications, 135*(6), 1-3.
20. Bengio, Y., Delalleau, O., & Simard, C. (2010). Decision trees do not generalize to new variations. *Computational Intelligence, 26*(4), 449-467.
21. Zhang, X., & Jiang, S. (2012). A Splitting Criteria Based on Similarity in Decision Tree Learning. *J. Softw., 7*(8), 1775-1782.
22. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1,* 81-106.
23. Venkatesan, E. V., & Velmurugan, T. (2015). Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology, 8*(29), 1-8.
24. D. T. Larose. (2005). “Discovering Knowledge in Data: An Introduction to Data Mining,” *Discov. Knowl.Data An Introd. to Data Min.,* pp. 1–222.
25. Apostolakis, J. (2010). An introduction to data mining. *Data Mining in Crystallography*, 1-35.