In-Depth Study of Decision Tree Model

Anas Muhammad Sani[1], Ahmad Salihu Ben-Musa[1], Muhammad Haladu[1]

[1, 2, 3] Department of Computer Science and Statistics, Federal College of Agricultural Produce Technology, Kano, Nigeria

Corresponding Author Email: *ahmadbnsalihu[at]fcapt.edu.ng*

Abstract: ***Machine Learning is the science and technology of teaching machines how to learn by them through training and testing and later determines the result for every condition without being programmed. Machine learning achieves this using different machine learning technique or Algorithms. Decision tree is one of the best classification algorithms that work both on categorical and continuous input and output variables. Different types of decision tree Algorithm exist with different accuracy and cost effectiveness. It is used by different professional in various fields. It can be used to classify emails weather it a spam or not, to find data as a replacement of statistical procedure, to extract text, to find missing data in a class and to improve search engines. Decision tree can as well be used in medical field for disease diagnosis. These documents described different types of Decision tree, how the algorithm work as well as its application. Their advantages and disadvantages with respect to other machine learning algorithms have also been discussed.***

Keywords: ID3, C4.5, IDA, CART, C5.0, MARS

1. Introduction

In our daily lives we come across situations where by we find ourselves in a state of presented with different choices or options which we have to make a selection among the choices through some form of judgment. Hence, we can say that a process or procedure through which we select the best out of the available choices is known as decision making. However, in order to make a good decision one should be able to weigh the pros and cons of each and every choice available and also take into consideration all other available choices. human beings have the capability and understanding to make decisions using their natural ways of thinking. However computers can mimic human ways of think and that enable them to learn, observe and make informed decisions just like humans.

According to [3] machine learning is one of the many branch of artificial intelligence (AI) which focuses on the use of available data and specialised algorithms to imitate or copy the way that humans learn, and adapt. A machine or Model that exhibit this behaviour is said to have learn.

Machine learning achieves this using different machine learning technique or Algorithms. Decision tree is among the best classification algorithms that work both on categorical and continuous input and output variables. Different types of decision tree Algorithm exist with different accuracy and cost effectiveness. Machine learning; however is an important part of the growing field of data science that is carried through the use of statistical methods. Algorithms are trained to make some prediction or classification based on the given data set thereby uncovering key insights within data mining work, and Hence the these insights consequently initiate decision making within applications, businesses, and other related domain.

Machine learning, can be categorised into three different categories or methods, supervised machine learning, unsupervised machine learning and semi-supervised machine learning [5].

- Supervised machine learning: A supervised machine learning also called supervised learning is a method of learning in which a labeled data set is used to train in the training algorithm that predict outcome or classify data accurately. As input data is fed into the model during the training, it changes its weights until the model has been fitted properly. This situation occurs as part of the cross validation method to ensure that the model escapes from being either over-fitting or under-fitting.
- Supervised learning are useful in solving real-world problem such as, classifying emails as either spam or not, disease prediction, and so on. supervised leaning algorithms includes, neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and Adaboost.
- Unsupervised Machine Learning: unsupervised Machine learning Known as unsupervised learning is another type of leaning that uses machine learning algorithms to examine and cluster unlabeled data sets, the algorithms in this type of learning usually discover hidden patterns within the data set without the need for human intervention. Its ability to discover similarities and differences within information make it a perfect solution for customer segmentation, cross-selling strategies, exploratory data analysis, pattern and image recognition. unsupervised Algorithms include: neural networks, k-means clustering, probabilistic clustering methods, component analysis (PCA) and singular value decomposition (SVD).
- Semi- supervised machine learning: also called Semi-supervised learning is a method that is between supervised and unsupervised learning. It uses a smaller portion of the labeled data set during training in order to guide classification and feature extraction from a larger unlabeled data set. Because of its nature, Semi-supervised learning can tackle issue of having not enough labeled data to train a supervised learning algorithm.

Apart from the above mentioned methods, another method of machine learning called Reinforcement machine learning also existed; this is a machine learning model that is analogous to supervised learning. However, this algorithm isn't trained using sample data, but rather, it learns as it goes by using trial and error. During this process a sequence of successful outcomes will be reinforced to develop the best policy for a given problem [3].

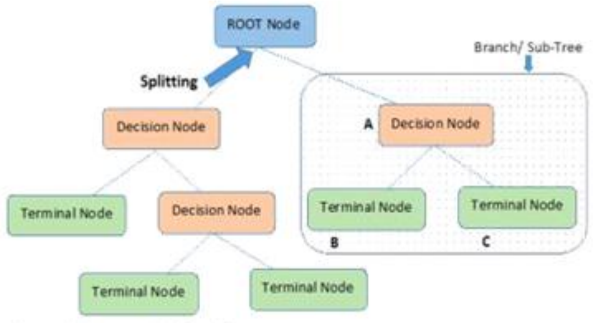
2. Decision Tree

According to [1] a decision tree is the most popular used tool for decision making. To accomplish this task one should draw a decision tree with different branches and leaves, the branches and leafs should point to all the various factors concerning a particular situation. A decision is almost similar to a decision support tool. It uses a tree-like graph of decisions and their possible outcomes which include resource costs, event outcomes, and utility. However, according to a survey carried out by [6], states a decision tree as the most popular approach for presenting classifiers. Tree based algorithms empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).Moreover, a decision tree is a classifier presented as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves or terminal nodes and sometimes also called decision nodes. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to as range. Figure 1 illustrates the structure of a decision tree.

3. How Decision Tree Evolved

Decision Tree was evolved from a model known as decision stump, a decision stump is a machine learning model that has one level decision tree, in the essence it is a decision tree with only one internal node that is directly connected to the leaf node. In decision stump, prediction is base on the value of just a single input feature. Decision stump can be expressed using the following notation or decision rule:

$$\text{If } x_1 > t \text{ then } \hat{y} = 0 \text{ else } \hat{y} = 0 \qquad (1)$$



Note:- A is parent node of B and C.

Figure 1: Structure of Decision Tree

```
Stepsize ← 1, minError ← 99999
For y = min (x) to max (x)  by stepsize do
    numError = numberOfErrors (y)
    If numErrs ≤ minErr  Then
    minErr ← numErrs
    $t_{beat} \leftarrow y$
    End If
End For
Return $t_{beat}$
```

$$\text{If } x_1 > 80 \text{ then } \hat{y} = 0 \text{ else } \hat{y} = 0 \qquad (2)$$

From equation 1 above the notation y^ is used to indicate a prediction of the variable y. Hence if we imagine a function: f(x) = (x_1 t), with t=80 then an equivalent rule is:

$$\text{If } f(x) > 0 \text{ then } \hat{y} = 0 \text{ else } \hat{y} = 0 \qquad (3)$$

From equation 2 above, the function f(x) is now a complete model with a parameter, t and is called a Decision Stump. Where a parameter t is a threshold that is required to change or switch decision from 1 to 0 and it is called decision boundary, hence the model has linear decision boundary.

However, the model above is just the function f(x) = (x1 - t), and the decision rule for the model is a simple if-then" rule. When the parameter is set correctly, the model f(x) should give good predictions. The only parameter to set is the t, so we do a simple line-search to find the optimal value, measuring the number of errors made on the data for each possible threshold. Finding the best parameter setting is what is usually referred to as learning.

Based on the above explanation on decision stump, we can deduced that a decision tree was formed by splitting a decision stump into two parts, left sub-tree and right-sub-tree, on the left sub-tree, $\hat{y} = 0$ and on the right sub-tree, $\hat{y} = 1$. Based on the above explanation on decision stump, we can deduced that a decision tree was formed by splitting a decision stump into two parts, left sub-tree and right-sub-tree, on the left sub-tree, y^ = 0 and on the right sub-tree, y^ = 1. For a given threshold t, we get:

Set y right to the most common label in the (> t) sub-sample set y left to the most common label in the (< t) sub-sample.

4. Types of Decision Tree

Basically, there are two types of decision tree, these types are based on the type of target variable used, we have categorical variable decision tree and continues variable decision tree.

1) Categorical Variable Decision Tree: This is the type of decision tree that has categorical variable type, example, yes or no. The categories mean that every stage of the decision process falls into one category, and there are no in-between.
2) Continuous Variable Decision Tree: This is a type that has has a continuous target variable, example an individual income can be predicted base on some information such as his age, occupation, qualification and so on

1. **Function** DecisionTree(depth, subsample)

```
If (depth==0) OR (all examples have same level) then //Best Case
    Return most common label in the subsample
End If
For each feature do //Recursive Case
    Try splitting data (build decision stump)
    Calculate the cost for this stump
End For
Pick feature with minimum cost
Find left/right subsamples
Add left branch ← BUILDTREE (leftSubSample, depth – 1)
Add right branch ← BUILDTREE (rightSubSample, depth – 1)
Return tree
End Function
```

Figure 3: Decision Tree Algorithm

5. Some Important Terminologies Used in Decision Tree

1) **Root Node:** It represents the entire population or sample and its further gets divided into two or more homogeneous sets.
2) **Splitting:** This is a process of dividing a node into two or more sub-nodes.
3) **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
4) **Leaf of Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5) **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
6) **Branch or Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.
7) **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

However, Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every sub-tree rooted at the new node [2].

6. How Decision Tree Works?

The criteria of making decision in decision tree depend on the type of task at hand for both classification and regression. Hence, the decision of making strategic splits heavily affects a tree's accuracy.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes [2]. With respect to the above, there are about five (5) or four different types of algorithm used in this regard.

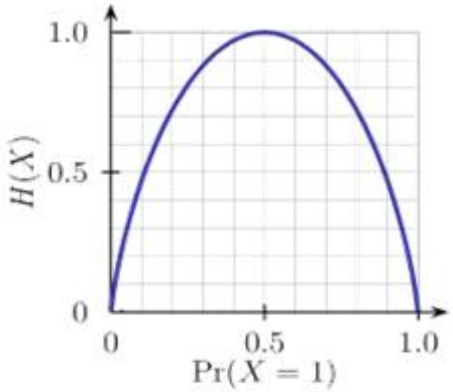


Figure 4: Entropy

7. Types of Decision Tree Algorithms

1) ID3 Algorithm: The ID3 algorithm is an extension of D3, it builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking. A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment. Steps in ID3 algorithm:
    a) It begins with the original set S as the root node. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy (H) and Information gain (IG) of this attribute.
    b) It then selects the attribute which has the smallest Entropy or Largest Information gain.
    c) The set S is then split by the selected attribute to produce a subset of the data.
    d) The algorithm continues to recur on each subset, considering only attributes never selected before.
2) C4.5 Algorithm: C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an expansion of Quinlan's earlier ID3 algorithm. The decision trees generate by C4.5 can be used for classification, and hence it is often referred to as a statistical classifier. It can allow data with categorical or numerical values, to knob 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 simply handle missing value. As missing attribute values are not utilize in gain calculations by C4.5 [4].
3) IDA (Intelligent Decision Tree Algorithm): is a better and more computational efficient than ID3. It uses a diversion measure of two attributes instead of entropy function. [8]
4) CART (Classification And Regression Tree Algorithm) uses Gini index classification. Numeric splitting used to construct trees [7].
5) MARS (Multi-Adaptive Regression Splines) is used find best splines. It uses regression trees. [7]
6) C5.0: is an improved version of C4.5. it allows whether to estimate missing value as a functions of other attributes or statistically used among other results.[7]

8. Attribute Selection

If the data-set consists of N attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. For solving this attribute selection problem, researchers worked and devised some solutions [2]. They suggested using some criteria like:

1) Entropy: This is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. From figure 2 above, it is quite evident that the entropy H(X) is zero when the probability is either 0 or 1. The Entropy is maximum when the probability is 0.5 because it projects perfect randomness in the data and there is no chance if perfectly determining the outcome.

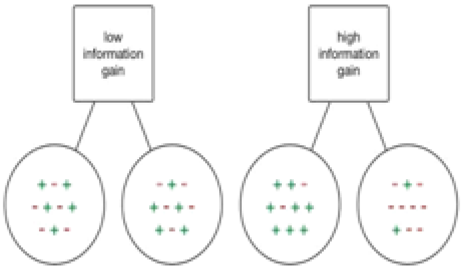


Figure 5: Information Gain

Mathematically, entropy with one attribute is given as:

$$\text{E(S)} = \sum_{i=1}^{c} - pilog_2pi \qquad (4)$$

2) Information Gain: Information gain or IG is a statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy. Information gain is nothing but decrease in entropy, hence the higher the information the lower the entropy. Figure 3 deficits information gain.

$$\textit{InformationGain}(\text{T, X}) = \textit{Entropy}(\text{T}) - \textit{Entropy}(\text{T, X}) \quad (5)$$

3) Gini Index: Gini Index can be seen as a cost function used to evaluate splits in the data-set. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values. Gini index is given as:

$$Gini = 1 - \sum_{i=1}^{c}(pi)_2 \qquad (6)$$

Gini Index works with the categorical target variable "Success" or "Failure". It performs only Binary splits.

Higher value of Gini index implies higher inequality, higher heterogeneity.

4) Gain Ratio: As Information gain is biased towards choosing attributes with a large number of values as root nodes. It means it prefers the attribute with a large number of distinct values. Below is the ratio gain formula:

$$Gain\ Ratio = \frac{InformationGain}{SplitInfor} = \frac{Entrophy(Before) - \sum_{j=1}^{c}(Entrophy(j,after)}{\sum_{j=1}^{k} w_j \log_2 w_j} \qquad (7)$$

5) Reduction in variance: This is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population, this algorithm uses the formula below:

$$Variance = \frac{\sum(X-\hat{X})^2}{N} \qquad (8)$$

6) CHAID: CHAID is an acronym which stand for Chi-squared Automatic Interaction Detector. It is one of the oldest tree classification methods, it use finds out the statistical significance between the differences between sub-nodes and parent node. We measure it by the sum of squares of standardized differences between observed and expected frequencies of the target variable.

$$\text{x}^2 = \sum \frac{(O-E)^2}{E} \qquad (9)$$

where:
x^2 = Chi square obtained
$\sum$ = The sum of
O = Observed Score
E = Expected Score

9. Application of Decision Tree

Decision tree has a diverse application in machine learning, it helps in providing solutions for different professionals in their various field, some of the applications are:

1) It is used in Assessing prospective growth opportunities on businesses based on the historical data
2) Using demographic data to find prospective clients that can help to streamline a marketing budget and make informed decision.
3) Serving as a support tool in several fields: Banks and other loan providers Lenders also use decision trees to predict the probability of a customer defaulting on a loan by applying predictive model generation using the client's past data. The use of a decision tree support tool can help lenders evaluate a customer's creditworthiness to prevent losses.

10. Advantages and Disadvantages of Decision Tree

10.1 Advantages

1) Easy to understand: Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
2) Less data cleaning required: Decision tree requires less data cleaning compared to other machine learning modeling techniques. It is not influenced by outliers and missing values to a fair degree.

3) Data type is not a constraint: Decision can handle both continues and categorical variables.
4) Non Parametric Method: Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

10.2 Disadvantages

1) Less effective in predicting the outcome of a continuous variable
2) Over fitting: Over fitting is one of the most practical difficulties for decision tree models.
3) Not fit for continuous variables: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories. Some of the problems associated with decision tree model can be tacked by using Pruning, bagging and boosting algorithm.

11. Conclusion

We can be able to conclude that, decision tree model is happens to be the best classification algorithm that work both on continues and categorical variables. However, Decision uses multiple algorithms to decide to split a node into two or more sub-nodes with respect to the type of target variable in use.

References

[1] Arundhati Navada, Aamir Nizam Ansari, S. P. and A.Sonkamble, B. (2011). Overview of use of decision tree algorithms in machine learning. IEEE Control and System Graduate Research Colloquium, pages 1–50.
[2] Chauhan, N. S. (2020). Decision tree algorithm, explained. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.
[3] Cloud Education, I. (2020). Machine learning. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.
[4] Patel, N. and Singh, D. (2015). An algorithm to construct decision tree for machine learning based n similarity factor. International Journal of Computer Applications (0975 – 8887).
[5] Ray, S. (2017). Commonly used machine learning algorithms (with python and r codes). https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.
[6] Rokach, L. and Maimon, O. (2015). Decision Tree. Research Gate.
[7] Harsh H. Patel, Purvi Prajapati (2018), Study and Analysis of Decision Tree Based Classification Algorithms, International Journal of Computer Sciences and Engineering, Vol 6, Issues – 10, Oct 2018.
[8] Pei-Lei Tu, Jen –Yao Chung (1992), A New Decision – Tree Classification Algorithm for Machine Learning, IEEE International Conference on Tools with AI, Arlington, Nov 1992.