ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

*International Journal of Mechanical Engineering*

# DETERMINE THE ENTROPY IN ID3 ALGORITHM TO SPLITTING A DECISION NODES AND PREDICT THE DECISION TREE

**P. Pavithra[1], Dr. M. Kavitha[2]**

[1]Research Scholar, Department of Mathematics, St.Peter's Institute of Higher Education and Research is a Deemed to be University, Avadi, Chennai, Tamil Nadu, India.

[2]Associative professor, Department of Mathematics, St.Peter's Institute of Higher Education and Research is a Deemed to be University, Avadi, Chennai, Tamil Nadu, India.

**Abstract:**

Decision tree is a powerful learning and classification process and also uncomplicated for induction. From the database, the decision tree learning presents a material in a data for discovery of relationships, patterns and knowledge. Both the number of attributes and instances are quite large in size when the volume of data is increased in the database. In databases, sets of records are large and quite complex tasks for decision tree learning. In the paper, Classification and prediction of decision problems can be solved with the use of decision trees, based on entropy and information gain. Here we have chosen an example; the bunch of women selecting dresses in a shopping mall for a college farewell party. Using the data, we calculate the entropy and information gain and make a possible decision from the formation of the decision tree.

**Keywords:** Decision tree, Data mining, Decision tree algorithm, IterativeDichotomiser3, Entropy.

## Introduction:

The formation of tree structure can be built into classification or regression models to form a decision tree. The algorithm of the structure forms a tree, hence known as a decision tree analysis. A decision tree technique was a coincidence with both quantitative and qualitative effects. The dataset breaks down into smaller subsets and develops a depth of the tree and shows their tree structure. The result or final node of the tree is known as decision nodes and leaf nodes. In a decision node they have two or more branches then it denotes classification or decision. Each class label and branches are defined as terminal nodes and the conjunction features to lead the class. The header of the decision node is a good predictor node for the tree and it is known as the root node of the tree. Processing of decision tree both categorical and numerical data. In decision trees, Decision tree algorithms belong to the group of supervised learning algorithms. Decision tree algorithms can resolve a classification and regression problem in supervised learning algorithms. To create a training model and apply decision rules to form a prior data (training data) and the decision tree is to design the prior model to predict the class or value to point. In decision trees, we start from the root node from a record to predicting a class label of the tree. From the records attribute contract the value of the root attribute. On comparing the value, we follow the branch corresponding to the value and move to the next node.

A training set contains a set of attributes and a class label be partitioned recursively to construct a decision tree using a decision tree algorithm. Real, Boolean or Ordinal values are input attributes. On a particular value of attribute be sustained at the state of the test for the decision node. In the test branch they expose each probable output for the tree. To identify the class of decision tree is traversed from the root to a leaf (terminal) of the tree. Classification of class at the leaf is defined by the decision tree. In test data, the classification accuracy defined as a percentage is correctly classified instances and specifies the performance of the decision tree.

## Literary review:

In 2018, Harsh patel et al., "Study and analysis of Decision Tree Based Classification Algorithms", performed work, classification of algorithms can be used to find the data for procedures of statistical replacement, to extract text, medical certified fields and also in search engines. According to various decisions tree algorithms have been constructed for their accuracy and cost of effectiveness. The ID3, C4.5 and CART are three different algorithms of the decision tree that the author included in the section.

In 2016, Amita Sharma et al., "Theoretical Study of decision tree algorithms to Identify pivotal factors for performance improvement: A Review" performed work to identify the factor of decision tree algorithms. The review of the contribution is providing a way to select an individual factor of improvement on the Decision tree algorithm.

In 2016, Banu GR et al., "A role of decision tree classification data mining Technique in diagnosing thyroid disease" performed work, comparing classification techniques accuracy through the confusion matrix by evaluating. It has been concluded that the algorithm gives better accuracy than the decision stump technique.

In year 2014, Brijain R patel., "A survey on decision tree algorithm for classification" performed work, the decision tree of ID3, C4.5, C5.0 as the classical algorithm, it merits the speed of classifying is high, ability if learning is and constructing a simple. They concentrate on the different algorithms of decision trees, their characteristics, challenges, advantages and disadvantages.

In year2014, Sonia singh et al., “Comparative Study ID3, CART and C4.5 decision tree algorithm: A survey “, performed work, these systems inductive methods to determine the appropriate classification of unknown objects to the given values of attributes according to decision tree rules.

In year 2014, Hssina, B., et al., “A comparative study of decision tree ID$_3$ and C4.5”, performed work, present a ID3 algorithm that is classical algorithm, here discussing the study of more about C4.5, it is one of the natural extensions of ID3 algorithm. And they make a comparison between these two algorithms and others algorithms.

In year 2014, jayakameswaraiah et al., “Implementation of an improved ID3 decision tree algorithm in data mining system”, performed work, discussing a shortcoming ID3 is chosen and many values in the attribute, and combining an ID3 and association function are presenting a new decision tree algorithm. These effectively reasonable and effective rules are shown in the proposed algorithm that can overcome shortcoming ID3.

In 2013, Pandey M et al., “A decision tree algorithm pertaining to the student performance analysis and prediction” performed work; the analysis related to improvement in quality education applying essential requirements and predicting a student's academic performance in higher education. While constructing a decision tree they considered some significant factors and formed a decision tree according to their grades.

**The tree construction algorithm:**

The tree builds algorithms use a divide and conquer approach to construct a decision tree. It evolves the decision tree for a given training set T consisting of set training instances. The values for a set of attributes and a class denote as instance. Let the class are denoted by $\{C_1, C_2 \ldots\ldots\ldots\ldots C_n\}$. Initially, the class frequency is computed for occurrence in training set T. If all instances belong to the same class, node K with that class is built. However, if set T contains instances depending on more than one class, for selecting a test attribute and splitting is executed then satisfying attribute and splitting criteria were chosen for the test at the node. The training set T is then partitioned into K exclusive subsets $\{T_1, T_2 \ldots\ldots\ldots\ldots T_K\}$ on the basis of this test and the algorithm is recursively applied on each non-empty partition.

**Construction an algorithm for decision tree:**

Construct (T)

**Step1**: Evaluate freq (C $_{i,}$ T)

**Step2**: If (all cases belong to same class labels), return terminal node or leaf.

**Step3**: For each attribute A tested for splitting criteria.

Attribute ‘A’ satisfying test is K (test node).

**Step4**: Again construct (T$_i$) on every partition T$_i$.

Include those nodes as children of parent node K.

**Step5**: Stop.

**Algorithm for the decision tree:**

**Step1**: In the dataset, the given attributes to identify the information gain.

**Step2**: various values of information gain for the datasets in descending order.

**Step3**: After evaluating the information gain choose the best attribute of the dataset as the root node of the tree.

**Step4**: Then using the same formula determines the information gain.

**Step5**: Based on the value of highest information gain for splitting the nodes as sub nodes of the tree.

**Step6**: Repeat the process until each attribute is set as leaf nodes in all the branches of the tree.

**Steps in ID3 algorithm:**

**Steps1**: The original set S as the root node of tree

**Steps2**: From the algorithm at each iteration, it iterates the attribute of the original set S and evaluates Entropy (S) and information gain (IG) of this attribute.

$$\text{Entropy(s)} = -2 \sum_{i=1}^{c} {p_i}^2 \log p_i$$

Where, S is collection of purity and impurity values

C is a number of attributes in a data set.

P is probability that the sample example belongs to class i.

Base 2 is measure of the information available

Information gain = entropy (parent) - [weighted average] * entropy (children)

**Steps3**: Attribute is selected which has the smallest entropy or largest information gain

**Step4**: Selected attribute split from original set S and produce a subset (branches) of the tree in dataset.

**Step5**: Repeat the algorithm to recur on every subset, considering only attributes never selected before.

- To construct a decision tree the ID3 algorithm used training dataset S, which is stored in memory. At runtime, this decision tree is used to categorize new test cases by traversing the decision tree using the features of the datum to arrive at a leaf node.
- When the probability value of entropy is maximum 0.5 then the projects become perfect randomness in data and there is no change if perfectly determining the outcome.
- Reducing entropy is related to increasing information gain and after a data –set is split on an attribute.
- Finding the attribute to constructing a decision tree that returns the highest information gain
- When entropy is zero then the tree turns leaf node and entropy more than zero needs further splitting for the tree.

**Example for calculating an Entropy for decision Tree algorithm**

Here we calculate the entropy for the selection of chuddar i.e., the possibility of selecting chuddar for that we compute the new entropy and information gain.

| No. | Chui Dhār Name | Colour of cloth | Cost | Quality | Buy a cloth |
|---|---|---|---|---|---|
| 1 | Palazzo | Red | High | High | No |
| 2 | Antalkali | Yellow | High | High | No |
| 3 | Sal war | Blue | High | High | Yes |
| 4 | Palazzo | Red | Medium | Low | Yes |
| 5 | Antalkali | White | Low | Low | No |
| 6 | Palazzo | Red | Low | Low | No |
| 7 | Sal war | Blue | Low | Low | Yes |
| 8 | Antalkali | Yellow | Medium | Medium | Yes |
| 9 | Antalkali | Yellow | Low | Low | No |
| 10 | Palazzo | White | Medium | Low | No |

**Calculating entropy using below formula**:

$$\text{Entropy(s)} = -2\sum_{i=1}^{c} {p_i}^2 \log p_i$$

Total number of data: 10, Total number of selections: 04, Total number of non- selections: 06

$$\text{Entropy (10)} = -2\left[\left(\left(\frac{4}{10}\right)^2\right) log\left(\frac{4}{10}\right) + \left(\left(\frac{6}{10}\right)^2\right) log\left(\frac{6}{10}\right)\ \right]$$

$$= -2[-0211508 – 0.265307]$$

$$= -2\ [-0.47681]$$

Entropy (10) = 0.9536

**Entropy for colour of cloth**

a) yellow colour of cloth

Total number of yellow colour cloth: 03, No. of yellow cloth chosen: 01, No. of yellow cloth not chosen: 02

$$\text{Entropy (3)} =-2\left[\left(\left(\frac{1}{03}\right)^2\right) log\left(\frac{1}{03}\right) + \left(\left(\frac{2}{03}\right)^2\right) log\left(\frac{2}{03}\right)\ \right]$$

=0.87218

b) Red colour of cloth

Total number of red colour cloth: 03, No. of red cloth chosen: 01, No. of red cloth not chosen: 02

$$\text{Entropy (3)} =-2\left[\left(\left(\frac{1}{3}\right)^2\right) log\left(\frac{1}{3}\right) + \left(\left(\frac{2}{3}\right)^2\right) log\left(\frac{2}{3}\right)\ \right]$$

=0.87218

c) Blue colour of cloth

Total number of blue cloths: 2, No. of blue cloth chosen: 2, No. of blue cloth not chosen: 0

$$\text{Entropy (2)} =-2\left[\left(\left(\frac{2}{2}\right)^2\right) log\left(\frac{2}{2}\right) + \left(\left(\frac{0}{2}\right)^2\right) log\left(\frac{0}{2}\right)\ \right]$$

=0

d) white colour of cloth

Total number of white cloths: 2, No. of white cloth chosen: 2, No. of white cloth not chosen: 0

$$\text{Entropy (2)} =-2\left[\left(\left(\frac{2}{2}\right)^2\right) log\left(\frac{2}{2}\right) + \left(\left(\frac{0}{2}\right)^2\right) log\left(\frac{0}{2}\right)\ \right]$$

=0

Similarly, we calculate entropy for cost, quality, chuddar

Now, calculating Information gain

Information gain = entropy (parent) - [weighted average] * entropy (children)

Gain (<Colour>, <White>, <Blue>, <Yellow>, <Red>)

$$=0.95363-\left[\left[\frac{2}{10}\right](0) + \left[\frac{2}{10}\right](0) + \left[\frac{3}{10}\right](0.87218) + \left[\frac{3}{10}\right](0.87218)\right]$$

=0.95363-0.523308

=0.430322

Gain (<Chuddar>, <Antalkali>, <Palazzo>, <Sal war>)

$$=0.95363-\left[\left[\frac{3}{10}\right](0.87218) + \left[\frac{4}{10}\right](0.716917) + \left[\frac{2}{10}\right](0)\right]$$

=0.95363-0.54842

=0.4052

Gain (<Quality>, <High>, <Medium>, <Low>)

$$=0.95363-\left[\left[\frac{3}{10}\right](0.87218) + \left[\frac{6}{10}\right](0.87218) + \left[\frac{1}{10}\right](0)\right]$$

=0.95363-0.78496

=0.16867

Gain (<Cost >, <High>, <Medium>, <Low>)

$$=0.95363-\left[\left[\frac{3}{10}\right](0.87218) + \left[\frac{3}{10}\right](0.87218) + \left[\frac{4}{10}\right](0.716917)\right]$$

=0.95363-0.8100748

=0.14355

Now, above the calculation of highest information gain is root node of the tree. For splitting the branches again have to follow the above same procedure.

By the process of entropy and information gain, we form a decision tree

**DECISION TREE:**

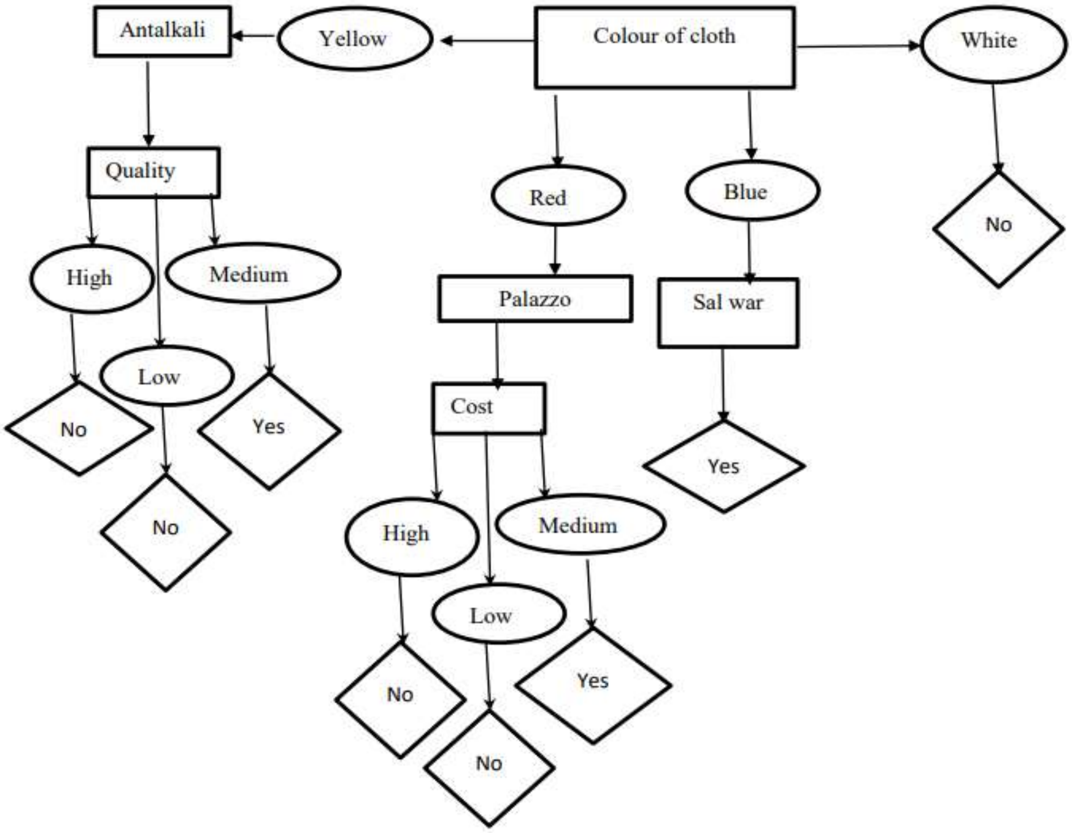

Figure: Decision Tree

Finally, we form a decision tree by applying new entropy.

**Conclusion:**

Decision tree build classification or regression models in the structure of a tree making it simple to debug and handle. The algorithm works by finding the information gain of the attributes and new entropy taking out the attributes for splitting the branches in trees. Here students select dresses in the mall for our farewell party. By using a dataset, construct a decision tree through the new entropy and information gain. The model can be made to produce an impact in the accuracy of the decision tree.

**References:**

[1] Ms. Priti Phalak, et al., "Analysis of Decision Tree – A Survey", International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-08181, volume 3, Issue 3, March -2014.

[2] Hssina, B., Merbouha,A., et al., " A comparative study of decision tree ID$_3$ and C4.5", International Journal of Advanced computer science and application, 4(2), 13-19, (2014).

[3] Josip Mesaric and Dario Sebalj, "Decision trees for predicting the academic success of students", Croatian Operation Research Review, 367-388, 2016.

[4] Himani Sharma, Sunil Kumar et al., "A survey on Decision tree Algorithms of classification in Data mining", International Journal of Science and Research, ISSN – 2319-7064, April 2016.

[5] He Zhang, Runnjing Zhou, "The analysis and optimization of decision tree based on ID3 algorithm", IEEE 2017; ISBN: 978-1-5090-6575-2.

[6] Harsh patel, purvi prajapati, "Study and Analysis of decision tree-based classification Algorithm", International Journal of computer sciences and engineering, ISSN:2347-2693, Vol-6, Issue -10, Oct 2018.

[7] Jayakameshwaraiah M, Ramakrishna S. "Implementation of an Improved ID3 Decision tree algorithm in data mining system", International Journal of computer Science and engineering, Volume-2 Issue -3 E-ISSN-2014.

[8] Banu GR. "A role of decision tree classification data mining technique in diagnosing Thyroid disease," international journal of computer sciences and engineering, 2016;4(11):111-5.

[9] Pooja Gulati et al., "Theoretical Study of Decision Tree Algorithms to identify pivotal Factors for Performance Improvement: A Review", International Journal of computer Applications, ISSN: 0975-8887, Volume 141 –No.14, May2016.

[10] Brijain R Patel et al., "Survey on Decision tree Algorithm for classification", IJEDR, 2014, Volume 2, Issue 1, ISSN:2321-9939.

[11] Sonia Singh, Priyanka Gupta, "Comparative study ID3, CART and C4.5 Decision tree algorithm: A Survey", International Journal Of Advanced Information Science and Technology, ISSN: 2319-2682, Vol.27, No.27, July 2014.

[12] Mrinal Pandey, Vivek Sharma, "A Decision tree algorithm pertaining to the student performance analysis and prediction", International Journal of Computer Applications, ISSN: 0975-8887, Vol61-No.13, Jan 2013.