COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY

Sonia Singh
Assistant Professor
Department of computer science
University of Delhi
New Delhi, India
14sonia.singh@gmail.com

Priyanka Gupta
Assistant Professor
Department of computer science
University of Delhi
New Delhi, India
shinepriyanka@gmail.com

Abstract: Decision tree learning algorithm has been successfully used in expertsystems in capturing knowledge. The main task performed in these systems isusing inductive methods to the given values of attributes of an unknown objectto determine appropriate classification according to decision tree rules.It is oneof the most effective forms to represent and evaluate the performance of algorithms, due to its various eye catchingfeatures: simplicity, comprehensibility, no parameters, and being able to handle mixed-type data. There are many decision tree algorithm available named ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE, and CTREE. We have explained three most commonly used decision tree algorithm in this paper to understand their use and scalability on different types of attributes and feature. ID3(Iterative Dichotomizer 3) developed by J.R Quinlan in 1986, C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993).CART stands for Classification and Regression Trees developed by Breiman *et al*.in 1984).

Keywords: Decision tree, ID3, C4.5, CART, Regression, Information Gain, Gini Index, Gain Ratio, Pruning,

I INTRODUCTION

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or *decision [8]*.Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in whicheach branch represents a possible scenario of decision and its outcome[1].ID3,CART and C4.5 are basically most common decision tree algorithms in data mining which use different splitting criteria for splitting the node at each level to form a homogeneous(i.e. it contains objectsbelonging to the same category) node.

II DECISION TREE

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a *rooted tree*, meaning 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 (also known as terminal or 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. Each leaf is assigned to one class representing the most appropriate targetvalue. Instances are classified bynavigating them from the root of the tree down to a leaf, according to theoutcome of the tests along the path [2].

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or *decision*is shown in Fig: 1 below in which each attribute name is represented in rectangles, attribute values are represented in ovals and the class (decision) is represented in diamonds.Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in whicheach branch represents a possible scenario of decision and its outcome [2].

Decision Tree Learning Algorithm

Decision tree learning is a method for approximating discrete-valued targetfunctions, in which the learned function is represented by a decision tree.Decision tree learning is one of the most widely used and practical methodsfor inductive inference'. (Tom M Mitchell, 1997, p52) Decision tree learning algorithm has been successfully used in expertsystems in capturing knowledge. The main task performed in these systems isusing inductive methods to the given values of attributes of an unknown objectto determine appropriate classification according to decision tree rules.Decision trees classify instances by traverse from root node to leaf node. We start from root node of decision tree, testing the attribute specified by thisnode, and then moving down the tree branch according to the attribute value in thegiven set. This process is the repeated at the sub-tree level. Decision Tree Algorithm is suited because of the following [1]:

1. Instance is represented as attribute-value pairs. For example, attribute 'Temperature' and its value 'hot', 'mild', 'cool'. Represented in Fig: 1 below.

Day	Outlook	Temperature	Humidity	Wind	Play (CLASS)
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

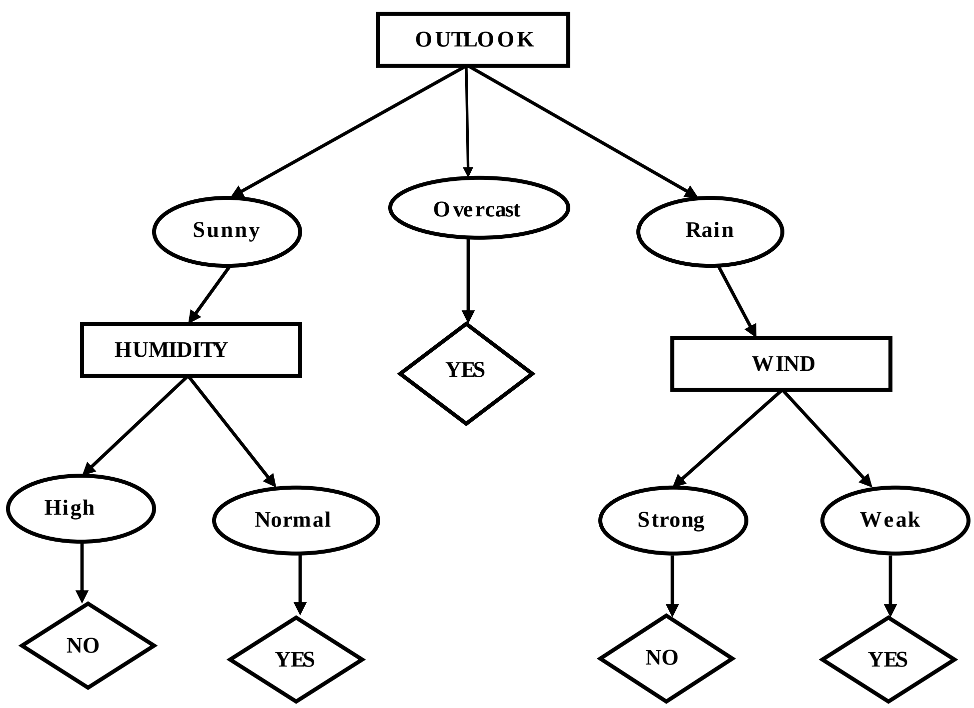


Fig: 1 Decision Tree Example

2. The target function has discrete output values. It can easily deal withinstance which is assigned to a Boolean decision, such as 'true' and 'false','p(positive)' and 'n(negative)'.

3. The training data may contain errors. This can be dealt with pruning techniques that we will not cover here.

We will cover the 3 widely used decision tree learning algorithms: ID3, CART and C4.5.

Decision Tree learning an attractive Inductive learningmethod because of the following reason [1]:

1. Decision tree is a good generalization for unobservedinstance, only if the instances are described in terms of
features that are correlated with the target class.

2. The methods are efficient in computation that is proportional to the number of observed training instances.
3. The resulting decision tree provides a representation of the concept that appeal to human because it renders the Classification process self-evident.

III SPLITTING CRITERIA

Basically all decision tree algorithms require splitting criteria for splitting a node to form a tree.In most of the cases, the discrete splitting functions are univariate. Univariate means that an internal node is split according to the value of a single attribute [2]. The algorithm used searches for the best attribute upon which to split. There are various splitting criteria based on impurity of a node. The main aim of splitting criteria is to reduce the impurity of a node. There are many measures of splitting that can be used to determine the best way to split the records. These splitting measures are defined in terms of the class distribution of the records before and after splitting. The example of impurity measures includes [3]:

Entropy: It is the measure of impurity of a node. It is defined (for a binary class with values) as:

$$\textbf{Entropy (t)} = -\sum p(i/t) \log_2 p(i/t)$$

Gini Index: It is another measure of impuritythat measures the divergences betweenthe probability distributions of the target attribute's values. The Gini Index has been used in various works such as (Breiman *et al*., 1984) and (Gleaned*et al*., 1991) and it is defined as: [3]:

$$\textbf{Gini Index} = 1-\sum [p(i/t)]^2$$

Classification Error: It is computed as:

$$\textbf{Classification error (t)} = 1-\max [p(i/t)]$$

Where, p (i/t) denote the fraction of records belonging to class i at a given node t.

Suppose an Attribute A has three values and Class label has two classifier then Entropy, Gini Index and Classification Error are computed as follow in Fig: 2
From Table 2 in Fig: 2 it is observed that all the three measures attain their maximum value when the class distribution is uniform (i.e., when p= 0.5, where p refers to the fraction of records that belong to one of the two class).The maximum values for the measures are attained when all the records belong to the same class (i.e., when p equals 0 or 1).

Information Gain: Information gain is an impurity-based criterion that uses the entropy measure as the impurity measure (Quinlan, 1987).It is the difference between the entropy of the node before splitting (parent node) and after splitting (child node).

$$\textbf{Info Gain } (\nabla) = \textbf{Entropy (parent node)} - \textbf{entropy (child node)}$$

Gain Ratio: The gain ratio "normalizes" the information gain as follows (Quinlan, 1993) [2].

$$\textbf{Gain Ratio} = \frac{\textbf{Information gain } (\nabla)}{\textbf{Entropy}}$$

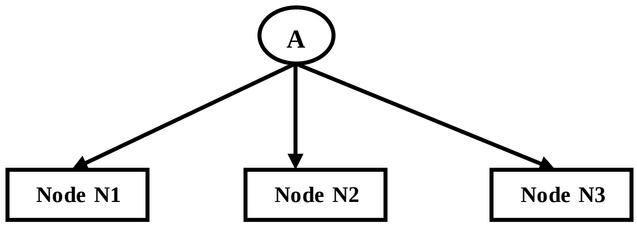
Impurity measures such as entropy and Gini Index tend to favor attributes that have large number of distinct values. Therefore Gain Ratio is computed which is used to determine the goodness of a split[3]. Every splitting criterion has their own significance and usage according to their characteristic and attributes type.

Twoing Criteria: The Gini Index may encounter problems when the domain of the target attribute is relatively wide (Breiman *et al*., 1984). In this case it is possible to employ binary criterion called twoing criteria. This criterion is defined as:

$$\textbf{Twoing Criteria (t)} = \frac{P_L P_R}{4}\left(\sum \left(|p(i/t_L) - p(i/t_R)|\right)\right)^2$$

Where, p (i/t) denote the fraction of records belonging to class i at a given node t.
Suppose an Attribute A has three values and Class label has two classifier then Entropy, Gini Index Index and Classification Error are computed as follow in Fig: 2

From Table 2 in Fig: 2 it is observed that all the three measures attain their maximum value when the class distribution is uniform (i.e., when p= 0.5, where p refers to the fraction of records that belong to one of the two class).The maximum values for the measures are attained when all the records belong to the same class (i.e., when p equals 0 or 1).



	N1	N2	N3
CLASS NO	**0**	**1**	**3**
CLASS YES	**6**	**5**	**3**

Table: 1 Showing Class Distribution

Impurity measure	Node 1	Node 2	Node 3
Entropy	$-(0/6)\log_2(0/6)-(6/6)\log_2(6/6)=0$	$-(1/6)\log_2(1/6)-(5/6)\log_2(5/6)=0.650$	$-(3/6)\log_2(3/6)-(3/6)\log_2(3/6)=1$
Gini Index Index	$1-(0/6)^2-(6/6)^2=0$	$1-(1/6)^2-(5/6)^2=0.278$	$1-(3/6)^2-(3/6)^2=0.5$
Classification error	**1-max[0/6,6/6]=0**	**1-max[1/6,5/6]=0.167**	**1-max[0/6,6/6]=0.5**

Table: 2

Figure: 2 Computing Impurity Measures

Where L and R refer to the left and right sides of a givensplit respectively, and p (i|t) is the relative frequency of class i at node t (Breiman, 1996). Twoing attempts to segregate data more evenly than the Gini Index rule, separating whole groupsof data andidentifying groups that make up 50 percent ofthe remaining data at each successive node

Stopping Criteria

The splitting (growing) phase continues until a stopping criterion is triggered. The following conditions are common stopping rules [1]:

1. All instances in the training set belong to a single value of *y*.
2. The maximum tree depth has been reached.
3. The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
4. If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.
5. The best splitting criteria is not greater than a certain threshold.

IV DECISION TREE ALGORITHMS

ID3

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 givendata in a *top-down* fashion, starting from a set of objects and a specification of properties.At each node of the tree, one property is tested based on maximizing information gain andminimizing entropy, and the results are used to split the object set. This process isrecursively done until the set in a given sub-tree is homogeneous (i.e. it contains objectsbelonging to the same category). This becomes a leaf node of the decision tree [4].The ID3 algorithm uses a greedy search. It selects a test using the information gaincriterion, and then never explores the possibility of alternate choices. This makes it a veryefficient algorithm, in terms of processing time. It has the following advantages and disadvantages [5]:

Advantages

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests.
- Whole dataset is searched to create tree.

Disadvantages

- Data may be over-fitted or over-classified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.
- Does not handle numeric attributes and missing values.

CART

CART stands for Classification and Regression Trees (Breiman *et al*., 1984).It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the twoing criteria and the obtained tree is pruned by cost–complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least–squared deviation). The prediction in each leaf is based on the weighted mean for node [5]. It has the following advantages and disadvantages[6]:

Advantages

- CART can easily handle both numerical and categorical variables.
- CART algorithm will itself identify the most significant variables and eliminate non-significant ones.
- CART can easily handle outliers.

Disadvantages

- CART may have unstable decision tree. Insignificant modification of learning sample such as eliminating several observations and cause changes in decision tree: increase or decrease of tree complexity, changes in splitting variables and values.
- CART splits only by one variable.

.

C4.5

C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993). The C4.5 algorithm generates a decision tree for the given data by recursively splitting that data. The decision tree grows using *Depth-first* strategy. The C4.5 algorithm considers all the possible tests that can split the data and selects a test that gives the best information gain (i.e. highest gain ratio). This test removes ID3's bias in favor of wide decision trees [8]. For each discrete attribute, one test is used to produce many outcomes as the number of distinct values of the attribute. For each continuous attribute, the data is sorted, and the entropy gain is calculated based on binary cuts on each distinct value in one scan of the sorted data. This process is repeated for all continuous attributes. The C4.5 algorithm allows pruning of the resulting decision trees. This increases the error rates on the training data, but importantly, decreases the error rates on the unseen testing data. The C4.5 algorithm can also deal with numeric attributes, missing values, and noisydata. It has the following advantages and disadvantages [7]:

Advantages

- C4.5 can handle both continuous and discrete attributes. In order to handle continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it [8].
- C4.5 allows attribute values to be marked as? For missing. Missing attribute values are simply not used in gain and entropy calculations.
- C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Disadvantage]

- C4.5 constructs empty branches; it is the most crucial step for rule generation in

C4.5.We have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for classification task. Rather it makes the tree bigger and more complex [7].

- Over fitting happens when algorithm model picks up data with uncommon characteristics. Generally C4.5 algorithm constructs trees and grows it branches 'just deep enough toperfectly classify the training examples'. This strategy performs well with noise free data. But most of the time this approach over fits the training examples with noisy data. Currently there are two approaches are widely using to bypass this over-fitting in decision tree learning [7].
- Susceptible to noise.

The basic Characteristic of the above three algorithms are explained in Table 3 below:

Characteristic(→) Algorithm(↓)	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible on outliers
CART	Towing Criteria	Handles both Categorical and Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical and Numeric value	Handle missing values.	Error Based pruning is used	Susceptible on outliers

Table 3: basic characteristic of decision tree algorithms

V CONCLUSION

In this paper we have studied the various basic properties of the decision tree algorithms which provides as a better understanding of these 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 best result on a specific type of data set.

REFERENCES

[1] Wei Peng, Juhua Chen and Haiping Zhou, of ID3,*' An Implementation Decision Tree Learning Algorithm'*, University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia .

[2] LiorRokach and OdedMaiman, *"Chapter 9 Decision Trees* "Department of InduatrialEngineering,"Tel-Aviv University"

[3] Vipin Kumar, Pang-Ning Tan and Michael Steinbach, "Introduction to Data Mining" Pearson.

[4]AjanthanRajalingam, DamandeepMatharu, KobiVinayagamoorthy, NarinderpalGhoman, "*Data Mining Course Project: Income Analysis*" SFWR ENG/COM SCI 4TF3, "*December 10, 2002*".

[5] Ahmed Bahgat El Seddawy, Prof. Dr Turkey Sultan, Dr. Ayman Khedr, "*Applying Classification Technique Using DID3 Algorithm to Improve Decision Support Under Uncertain Situations*". 'Department of Business Information System, Arab Academy for Science and Technology and Department of Information System, Helwan University, Egypt. "*International Journal of Modern Engineering Research ", Vol 3, Issue 4, July- Aug 2013 pp-2139-2146.*

[6]Roman Timofeev to Prof. Dr. Wolfgang Hardle"*Classification and Regression Trees (CART). Theory and Applications*," CASE- Center of Applied Statistics and Economics, Humboldt University, Berlin Dec 20, 2004.

[7] Mohammad M Mazid,A B M Shawkat Ali, Kevin Tickle, "*Improved C4.5 Algorithm for Rule Based Classification*" School of Computing Science, Central Queensland University, Australia.

[8] http://www.wikipedia.com/

Authors Profile

Sonia Singh received M.Sc. Degree from Delhi University in Computer Science in 2012. Received bachelor's degree from Delhi University in B.Sc.(H) Computer Science in 2010. .Currently working as Assistant Professor in Delhi University.

Priyanka Gupta received M.Sc. Degree from Delhi University in Computer Science in 2012. Received bachelor's degree from Delhi University in B.Sc. (H) Computer Science in 2009. .Currently working as Assistant Professor in Delhi University.