ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

COMPARATIVE ANALYSIS OF DECISION TREE ALGORITHMS FOR THE PREDICTION OF ELIGIBILITY OF A MAN FOR AVAILING BANK LOAN

M. Mohankumar[1], S. Amuthakkani[2] and G. Jeyamala[3]
[1]Assistant Professor, Dept of CSE
[2,3]UG Student, Dept of CSE
[1,2,3] Sri Vidya College of Engineering & Technology, Tamilnadu
[1]m.mohankumar18589@gmail.com
[2]amuthakkani@gmail.com
[3]jeyamalag95@gmail.com

Abstract: An evaluation of machine learning algorithm for prediction of person, whether who is eligible to loan or not. Decision tree and linear regression algorithms are useful for such predictions. To predict, the system needs information of a person which will be read by the system via dynamic web page which are analyzed. Decision tree learning algorithms are been successfully used in expert systems in capturing knowledge. The main work performed in this system is using inductive methods to the attributes of an unknown object to determine appropriate classification according to decision tree rules. There are many decision tree algorithms available namely ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE, and CTREE. In this paper, three most commonly used decision tree algorithms are discussed namely ID3, CART, C4.5 and the algorithms are compared with the parameters....loan amount required, age, asset value, income, family details, guarantor asset value, guarantor income, already avail loan on asset. These parameters are used to compare the three different kind of decision tree algorithms.

Keywords : Decision tree, ID3, C4.5, CART

I. INTRODUCTION

A decision tree is a tree like structure in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision. 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. A decision tree uses the traditional tree structure in which internal nodes represent the test on an attribute, each branch represents the classification Rules. It

starts with a single root node that splits into the multiple branches and leading to further nodes, each of which may further split or else terminate as a leaf node. Splitting depends upon the values of the attributes.

Associated with each non leaf node, there will be a test which may determine how to proceed and which branch to follow from that node. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.ID3,CART and C4.5 are the most common decision tree algorithms in data mining which uses different splitting criteria for splitting the node at each level to form a homogeneous (i.e. it contains objects belonging to the same category) node.

II. DATA SCIENCE

Data Science is an interdisciplinary field about processes and systems to extract knowledge or an insights from data in various forms, either structured or unstructured, which is a continuation of the analytics fields such as statistics, data mining, and predictive analysis & Knowledge Discovery in Databases (KDD). Data science employs techniques and theories drawn from many fields within the broad areas such as mathematics, statistics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modelling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing. Methods that scale to big data are of particular interest in data science, although the discipline is not generally considered to be restricted to such big data, and big data solutions are often focused on the organizing and pre processing the data instead of analysis. The development of machine learning has been enhanced the growth and importance of data science.

Data science affects the academic and applied research in many domains it includes machine translation, speech recognition, robotics, search engines, digital economy, biological sciences, medical informatics, health care, social sciences and the humanities. It heavily influences the economics, business and finance. From the business perspective, data science is said to be an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities, such as data mining and data analysis.

III DECISION TREE

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences that including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to help for identifying a strategy most likely to reach a goal, but are also a popular tool in machine learning.

A decision tree is simply like a flowchart structure in which each internal node represents a "test" on an attribute (for e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The path from root to leaf represents the classification rules. In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree consists of 3 types of nodes that includes:

1. Decision nodes - represented by squares
2. Chance nodes - represented by circles
3. End nodes - represented by triangles

IV ATTRIBUTE GENERATION

The attributes for the prediction of eligibility for a man to get loan from the bank is described here. Those information's provided by the person who expects loan and values of the attributes captured via Dynamic web pages where the user supposed to fill the authenticated information and there will be a validation of information given by the user by Bank Officials.

Attributes	Values
Loan amount required	<500000
Age	19-60
Assert Type	Dead or Fixed
Assert Value	20% of loan , 20%-50% of loan , 50 % of loan
Employment Type	Private , Government , Government undertaking
Income	>= 150% of loan
Marital Status	Single, Married
Having Children	Yes, no
Education details for children	
Parents living together	Yes , no
Guarantor's Assert Type	Dead or Fixed
Guarantor's Assert Value	20% of loan , 20%-50% of loan , 50 % of loan
Bank deposit	Yes or no
Bank deposit value	
Repayment period	
Already availing loan on assert	Yes or no
Insurance on assert	Yes or no

Table 1 Attribute generation

V COMPARISION OF ALGORITHM

ID3

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan, it is used to generate a decision tree from a dataset. ID3 is the precursor of C4.5 algorithm, and is typically used in the machine learning and natural language processing domains.

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 given data in a top-down approach, starting from a set of objects and a specification of properties. At each node of the tree, one property is tested based on maximization of information gain and minimization of entropy, and the results are used to split the object set. This process is recursively done until the set in a given sub-tree is homogeneous (i.e. it contains the objects that belonging to the same category). This becomes a leaf node of the decision tree .The ID3 algorithm uses a greedy approach. It selects a test using the information gain criterion, and then never explores the possibility of alternate choices. This makes that it is a very efficient algorithm, in terms of processing time. The advantages and disadvantages are:

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 the test data to be pruned, reducing number of tests. Whole dataset is used to create tree.

Disadvantages:

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

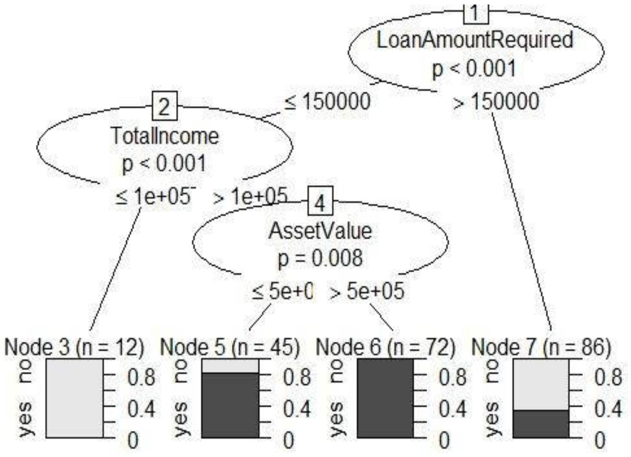


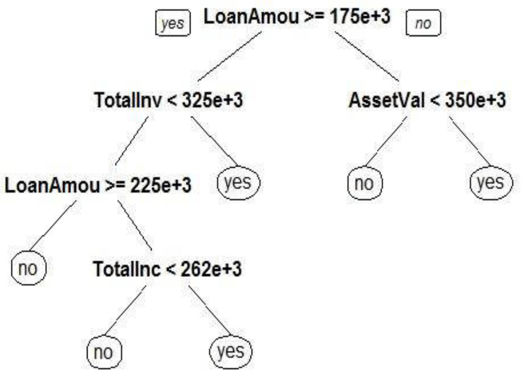
Figure 1 ID3 Algorithm

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 then the obtained tree is pruned by cost–complexity Pruning. When provided, CART can be consider misclassification costs in the tree induction. It also enables users to provide the prior probability distribution. An important feature of CART is that 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 prediction in each leaf is based on the weighted mean for node in a tree. Some of the advantages and disadvantages are:

Advantages:

- CART can handle both numerical and categorical variables.



- CART algorithm will itself identify the most significant variables and eliminate non significant variable.
- It 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.

Figure 2 CART Algorithm

C4.5

C4.5 is an evolution of ID3, presented by the same author. The C4.5 algorithm generates a decision tree for the given data by recursively splitting that data. The decision tree grows using the Depth-first strategy. The C4.5 algorithm considers all the possible tests that can be split the data and selects a test that gives the best information gain (i.e. highest gain ratio). This test removes that ID3's bias in favour of wide decision tree . For each discrete attribute, one test is used to produce many outcomes as the number of distinct values of that attribute. For each continuous attributes, the data is sorted, and then 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 the 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. It can also deal with numeric attributes, missing values, and noisy data. The advantages and disadvantages are:

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.
- C4.5 allows attribute values to be marked as for missing. The 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:

- It 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 more complex and bigger.

Characteristic(→)
Algorithm(↓)
	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection	Speed

ID3	Information Gain	Handles only Categorical value	Do not handle missing values	No pruning is done	Susceptible on outliers	Low
CART	Towing Criteria	Handles both Categorical and Numeric value	Handle missing values	Cost-Complexity pruning is used	Can handle Outliers	Average
C4.5	Gain Ratio	Handles both Categorical and Numeric value	Handle missing values	Error Based pruning	Susceptible on Outliers	Faster than ID3

Table 2 Comparison of Algorithm

Algorithm	Accuracy	Recall	Precision
ID3	0.54	0.604	0.763
C4.5	0.733	0.92	0.821
CART	0.701	0.608	0.736

Table 3 Performance Metrics Comparison

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

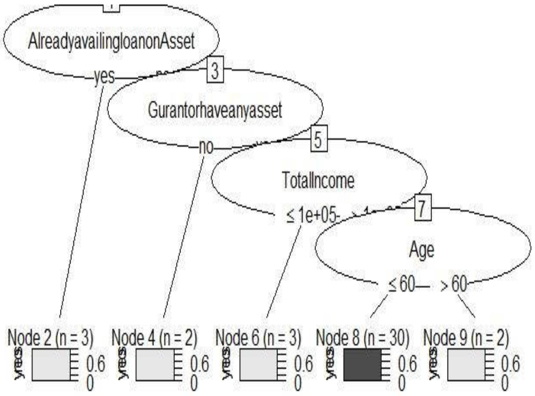


Figure 3 C4.5 Algorithm

Metrics:

Entropy:

Entropy *H(S)* is a measure of the amount of uncertainty in the (data) set *S* (i.e. entropy characterizes the (data) set *S*).

$$\mathbf{H(S)= -\sum_{x\in X} p(x) \log_2 p(x)}$$

where

S - The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)

X - Set of classes in **S**

p(x) - The proportion of the number of elements in class *x*

When *H(S)=0* the set is *S* perfectly classified (i.e. all elements in *S* are of the same class).

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set *S* on this iteration. The higher the entropy, the higher the potential to improve the classification here.

Information Gain:

Information gain *IG(A)* is the measure of the difference in entropy from before to after the set S is split on an attribute *A*. In other words, how much uncertainty in *S* was reduced after splitting set S on attribute *A*.

Where,

$$\mathbf{IG(A, S) = H(S) - \sum_{t\in T} p(t)H(t)}$$

H(S)- Entropy of set *S*

T- The subsets created from splitting set *S* by attribute *A*

p(t)- The proportion of the number of elements in *t* to the number of elements in set *S*

H(t)- Entropy of subset *t*

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set *S* on this iteration.

Gini Index:

It is another measure of impurity that measures the divergences between the probability distributions of the target attribute's values. The Gini Index has been used in various works such as (Breiman et al., 1984) and (Gleanedet al., 1991) and it is defined as :

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

Classification Error:

It can be computed as:

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

Where,

p (i/t)- The fraction of records belonging to class i at a given node t.

Gain Ratio:

The gain ratio "normalizes" the information gain as follows :

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

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

Towing 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:

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

Where,

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

VI VARIABLE COLLECTION FOR DECISION TREE

Nominor Details :

(i) Personal Details

- Name
- Aadhar ID No
- Husband's/ Father's Name
- Present Address
- Mobile No
- Permanent Address
- Gender
- Date of Birth
- Age(Completed Years)
- Educational Qualification

(ii) Job Details

- Working in
- Designation
- Are they permanent employee?
- Total Income

(iii) Family Details

- Are you Married?
- Spouse Name
- Spouse Occupation
- Spouse Salary
- How many kids you have?
- Kids Name
- Kids Age
- Kids studying institution
- Kids studying standard
- Are you live with your parents?
- Parent's age
- Do they have any physical or mental problems?

(iv) Asset Details

- Total investment amount
- Total liabilities amount
- Are you have any asset or not?
- Asset value

(v) **Proposed Loan Details**

- Loan amount required
- Purpose of loan
- Repayment period

(vi) **Previous Loan Details**

- Loan amount sanctioned
- Purpose of loan
- Repayment period

Guarantor Details :

- Guarantor Name
- Guarantor Husband's/ Father's Name
- Guarantor Address
- Guarantor Mobile No
- Guarantor Bank Account No
- Guarantor Occupation
- Guarantor Income
- Is there Guarantor having any asset?
- Guarantor Asset Value

VII CONCLUSION

In this paper we have studied the different basic properties of the decision tree algorithms which provides a better understanding of these various algorithms .We can apply them on various kinds of data sets having various kinds of values and properties can attain a best result by knowing that which algorithm will give the best result on a specific type of data set.

REFERENCES :

1. 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.

2. Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

3. J.R Quinlan, "Induction of Decision Trees Machine Learning". Vol.1, pp81-106, 1986

4. Rupali Bhardwaj , Sonia Vatta," Implementation of ID3 Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering

5. Harvinder Chauhan, 2Anu Chauhan "Implementation of decision tree algorithm c4.5", International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013

6. Rafik Khairul Amin, Indwiarti, Yuliant Sibaroni, Implementation of Decision Tree Using C4.5 Algorithm in Decision Making of Loan Application by Debtor (Case Study: Bank Pasar of Yogyakarta Special Region)