IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

Study and Analysis of Decision Tree Based Classification Algorithms

Harsh H. Patel[1*], Purvi Prajapati[2]

[1,2]Dept. of Information Technology, CSPIT, Charotar University of Science and Technology, Changa, Gujarat, India.

[*]*Corresponding Author: harshpatel4598@gmail.com, Tel.: +91 7043046700*

Available online at: www.ijcseonline.org

Accepted: 13/Oct/2018, Published: 31/Oct/2018

***Abstract*—** Machine learning is to learn machine on the basis of various training and testing data and determines the results in every condition without explicit programmed. One of the techniques of machine learning is Decision Tree. Different fields used Decision Tree algorithms and used it in their respective application. These algorithms can be used as to find data in replacement statistical procedures, to extract text, medical certified fields and also in search engines. Different Decision tree algorithms have been built according to their accuracy and cost of effectiveness. To use the best algorithm in every situations of decision making is very important for us to know. This paper includes three different algorithms of Decision Tree which are ID3, C4.5 and CART.

***Keywords*—** Machine Learning, Decision Tree (DT), WEKA tool.

I. INTRODUCTION TO DECISION TREE

Classification is the task of giving objects to categories which have many diverse applications.

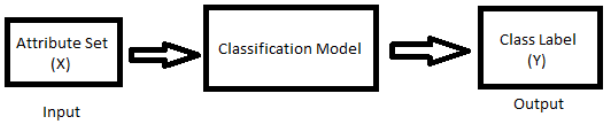

Fig. 1: Classification of mapping attribute set (X) to its class label (Y)

Decision Tree

A normal tree includes root, branches and leaves. The same structure is followed in Decision Tree. It contains root node, branches, and leaf nodes. Testing an attribute is on every internal node, the outcome of the test is on branch and class label as a result is on leaf node [3, 4]. A root node is parent of all nodes and as the name suggests it is the topmost node in Tree. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continues value) [4]. As decision trees mimic the human level thinking so it's so simple to grab the data and make some good interpretations. The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf.

II. RELATED WORK ON DECISION TREE

Decision Tree is similar to the human decision-making process and so that it is easy to understand. It can solve in both situations whether one has discrete or continuous data as input. The example of Decision Tree is as follow [15].

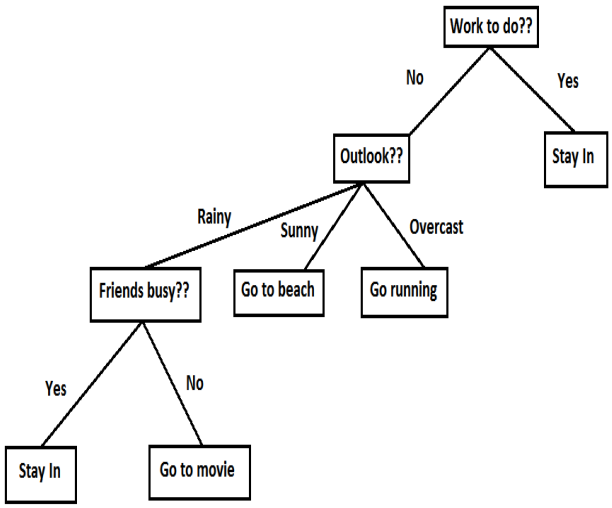

Fig. 2: Example of Decision Tree on what to do when different situations occur in weather.

When data does not offer benefits while splitting, it directly stops the execution. Try to find one test at a time rather than optimize the whole tree together.

Talking about the characteristics of Decision Tree, the ID3 algorithm is simulated only on WEKA tool and the data type of data set is only categorical. ID3 can not take continuous data set for simulation. Similarly, CART and C4.5 have same

© 2018, IJCSE All Rights Reserved

characteristics as ID3 has. The only difference is that C4.5 and CART both can take continuous data set as input for simulation purpose [11].

Table-1: Characteristics of DT.

Decision Tree Algorithm	Data Types	Numerical Data Splitting Method	Possible Tool
CHAID	Categorical	N/A	SPSS answer tree
ID3	Categorical	No Restriction	WEKA
C4.5	Categorical, Numerical	No Restriction	WEKA
CART	Categorical, Numerical	Binary Splits	CART 5.0

The decision tree makes explicit all possible alternatives and traces each alternative to its conclusion in a single view, to make easy comparison among the various alternatives [12]. Transparent in nature is one of the best advantages of Decision Tree.

Another main advantage is the ability to selecting the most biased feature and comprehensibility nature. It is also easy to classify and Interpretable easily. Also used for both continuous and discrete data sets.
Variable screening and feature section are good enough in decision tree [19]. By talking on its performance, non-linear does not affect any of the parameters of the decision tree.

III. DECISION TREE ALGORITHMS

Decision tree algorithms are used to split the attributes to test at any node to determine whether splitting is "Best" in individual classes. The resulting partitioned at each branch is PURE as possible, for that splitting criteria must be identical.

Table- 2: Decision tree algorithms

Algorithm name	Classification	Description
CART (Classification and Regression Trees)	Uses Gini Index as a metric.	By applying numeric splitting, we can construct the tree based on CART [4].
ID3 (Iterative Dichotomiser 3)	Uses Entropy function and Information gain as metrics.	The only concern with the discrete values. Therefore, continuous dataset must be classified within the discrete data set [5].
C4.5	The improved version on ID 3	Deals with both discrete as well as a continuous dataset. Also, it can handle the incomplete datasets. The technique called "PRUNNING", solves the problem of over-filtering [9].
C5.0	Improved version of the C4.5	C5.0 allows to whether estimate missing values as a function of other attributes or apportions the case statistically among the results [13].
CHAID (CHi-square Automatic Interaction Detector) [6]	Predates the original ID3 implementation.	For a nominal scaled variable, this type of decision tree is used. The technique detects the dependent variable from the categorized variables of a dataset [3, 11].
MARS (multi-adaptive regression splines)	Used to find the best split.	In order to achieve the best split, we can use the regression tree based on MARS [2, 10].

IV. METRICS

According to the values of the splitting attribute, the training data are partitioned into several subsets. Until all instances in a subset belong to the same class in any Decision Tree the algorithm proceeds recursively [6].

Table- 3: Splitting Criteria

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	$Gini\ Index, G = \left(\frac{1}{2n^2\mu}\right)\sum_{j=1}^{m}\sum_{k=1}^{m} n_j n_k \lvert y_j - y_k \rvert$

© 2018, IJCSE All Rights Reserved

Information Gain is biased towards multivariate attributes which are the main drawback of Information Gain [6]. The unbalanced split of data where one of the child nodes has more number of entries compared to the others Gain Ratio generally prefers that [7, 12]. Gini Index gives unfavorable results as with more than two categories in the data set. These are the drawbacks of splitting criteria [15].

V. EVALUATION MECHANISM

If the values are close to each other, the set can be said to be precise. If their average is close to the true value of the quantity being measured, the set can be said to be accurate. Only if given a set of data points from repeated measurements of the same quantity then one can measure above two terms [13].

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

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

TP = True positive, TN = True Negative
FP = False Positive, FN = False Negative

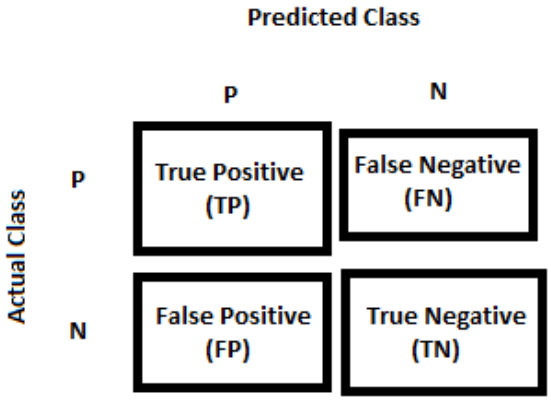

Fig. 3: Confusion Matrix sample in Decision Tree.

VI. DATASET DESCRIPTION

The dataset which is used in the experiment is car dataset. By applying this dataset on three algorithms of the decision tree which are ID3, C4.5, and CART. Dataset description is as follow.

The car dataset is in two parts. One is Car Acceptability and other is Technical Characteristic. Overall price (buying) and Price of Maintenance (maint) are two attributes of Car Acceptability. The Number of doors (doors), Capacity in terms of persons carries (persons), the size of luggage boot (lug_boot) and estimated the safety of the car (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%

VII. EXPERIMENT

The experiment is simulating on WEKA tool. For data mining tasks, WEKA is a collection of machine learning algorithms. For data pre-processing, classification, regression, clustering, association rules, and visualization Weka contains tools. Weka is open source software issued under the GNU General Public License. It is also well-suited for developing new machine learning schemes. The algorithms can either be applied directly to a dataset or called from your own Java code [18].

Table- 4: Theoretical results

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

© 2018, IJCSE All Rights Reserved

For the experiment, this paper distributes the same data sets on three different decision tree algorithms like ID3, C4.5, and CART. The results of all three algorithms in the terms time and accuracy with the help of the outcome from the below table [17]. The splitting Criteria column gives information about how the algorithm split in order to get a better result. The attribute type column gives information about what type of values the algorithm can handle. Whether the algorithm finds the missing value or not, the result defines from the Missing Value column and thus the algorithm is accurate or not we can find.

Table- 5: Practical results

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

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.

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.

Confusion Matrix:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  35 363  27   3 |   a = vhigh
 361   4  60   6 |   b = high
 267  54  11 100 |   c = med
 237  41 107  47 |   d = low
```

Fig. 2 – Confusion matrix for ID3

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 341  64  27   0 |   a = vhigh
 348  24  46  14 |   b = high
 261  37  48  86 |   c = med
 231  23  84  94 |   d = low
```

Fig. 3 – Confusion matrix for C4.5

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 360  61  11   0 |   a = vhigh
 341  44  39   8 |   b = high
 268  41  57  66 |   c = med
 246  20  73  93 |   d = low
```

Fig. 4 – Confusion matrix for CART

VIII. CONCLUSION

The Decision Tree algorithms ID3 C4.5 and CART were applied on the dataset. Decision tree outperforms others in terms of accuracy, time and precision. It quite relies on the algorithm used for recommendation to find interesting resources. At last, the comprehensive study is done about decision tree algorithms and this paper concludes that CART is the algorithm for this dataset is very precise and most accurate among the others.

IX. FUTURE WORK

In the future, this will be installed in the Apache server thus published on the internet. [17] Datasets are updated continuously and it will take online rating for the prediction. The prediction approaches can also be tried in different datasets to check the performance of the system.

REFERENCE

Research Papers

[1]. Sorower MS. A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis. 2010 Dec;18.

[2]. Utku A, Hacer (Uke) Karacan, Yildiz O, Akcayol MA. Implementation of a New Recommendation System Based on Decision Tree Using Implicit Relevance Feedback. JSW. 2015 Dec 1;10(12):1367-74.

[3]. Gershman A, Meisels A, Lüke KH, Rokach L, Schclar A, Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).

[4]. Jadhav SD, Channe HP. Efficient recommendation system using decision tree classifier and collaborative filtering. Int. Res. J. Eng. Technol. 2016;3:2113-8.

[5]. Beel J, Langer S, Genzmehr M, Nürnberger A. Introducing Docear's research paper recommender system. InProceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries 2013 Jul 22 (pp. 459-460). ACM.

[6]. Zhang X, Jiang S. A Splitting Criteria Based on Similarity in Decision Tree Learning. JSW. 2012 Aug;7(8):1775-82.

[7]. Bhargava N, Sharma G, Bhargava R, Mathuria M. Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Jun;3(6).

[8]. Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security. 2009 Jun;3(3):230-40.

© 2018, IJCSE All Rights Reserved

[9]. Freund Y, Mason L. The alternating decision tree learning algorithm. Inicml 1999 Jun 27 (Vol. 99, pp. 124-133).
[10]. Pandey M, Sharma VK. A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications. 2013 Jan 1;61(13).
[11]. Priyama A, Abhijeeta RG, Ratheeb A, Srivastavab S. Comparative analysis of decision tree classification algorithms. International Journal of Current Engineering and Technology. 2013 Jun;3(2):334-7.
[12]. Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security. 2009 Jun;3(3):230-40.
[13]. Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar 1;1(1):81-106.
[14]. Drazin S, Montag M. Decision tree analysis using weka. Machine Learning-Project II, University of Miami. 2012:1-3.
[15]. 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.
[16]. Jayakameswaraiah M, Ramakrishna S. Implementation of an Improved ID3 Decision Tree Algorithm in Data Mining System. International Journal of Computer Science and EngineeringVolume-2, Issue-3 E-ISSN. 2014.

Books

[17]. Larose D.T. (2005), Discovering Knowledge in Data: An Introduction to Data Mining, Wiley.
[18]. DATA MINING WITH DECISION TREES: Theory and Applications (2nd Edition) by Lior Rokach and Oded Maimon.
[19]. Lior R. Data mining with decision trees: theory and applications. World Scientific; 2014 Sep 3.

Authors Profile

Harsh H. Patel pursuing Bachelor in Information Technology from Chandubhai S. Patel Institute of Technology, CHARUSAT. His area of interest is Data Mining.

Purvi Prajapati is working as an assistant professor at Department of Information Technology in Chandubhai S. Patel Institute of Technology, CHARUSAT since June 2006. She had received the degree of Master of Technology in Computer Engineering from Chandubhai S Patel Institute of Technology, CHARUSAT in 2012 and Bachelor of Engineering in information technology from A D Patel Institute of Technology in 2004. She has 11 years of teaching experience including subject proficiency in Data Structure, Computer Network, Data Mining, Language Processor and Programming Languages (C, C++, Java, Python). Her research interest includes data mining in machine learning. She has published 09 papers in international journal and conferences and also a member of ACM.

© 2018, IJCSE All Rights Reserved