Strategies for Vibration Testing of Decision Tree-Based Classification Systems.

Kandimalla Gopi , Assistant Professor, Guru Nanak Institutions Technical Campus, Hyderabad, kandimalla.gopi@gmail.com, 9182653788
Author Name2:Veera Swamy PittalaAssistant Professor,Lakireddy Bali Reddy College of Engineering,Mylavaram,veeraswamypittala@gmail.com,9493843237

***Abstract*—** *" Without any intervention from a person, computers are capable of "learning" new things by analyzing data in various ways (training and testing) and making conclusions. Machine learning includes decision trees. Decision Tree techniques are extensively employed in many different industries. These algorithms have a wide variety of potential applications, including search engines, text extraction, and companies that provide medical certifications. Decision tree algorithms that are both accurate and affordable are now at our fingertips. Whenever a choice is necessary, it is critical to know what the best choice is. With the use of tools like WEKA, ML, and DT, we present three decision tree algorithms in this study: ID3, C4.5, and CART.*

I. EXPLORING THE DECISION TREE

Classification is the process of assigning things to categories, and it has many different uses.

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

Decision Diagram

The main parts of a typical tree are the trunk, the branches, and the leaves. Decision Tree follows the same pattern. With its trunk, branches, and foliage, this structure is reminiscent of a tree. As stated in references 3, and 4, attribute checking occurs at each leaf node. At the leaf node, you can see the class name, and the results are passed down the branch. Biological parents of all subsequent nodes in a tree are located at the root node, the initial node in the tree. According to [4], a "node" represents a quality or attribute, a "branch" a choice or rule, and a "leaf" a result, whether continuous or categorical. Decision trees allow for the fine-tuning of data collecting and analysis since its architecture is grounded on human thought processes. We want to create such a tree for all the data, with the goal of processing a single result at each leaf.

CONNECTED RESEARCH ON THE DECISION TREE

Decision Tree is simple because it mimics the way humans make decisions. Problems inregardless of whether one is dealing with discrete or continuous data. Here's a sample [15] of a Decision Tree.

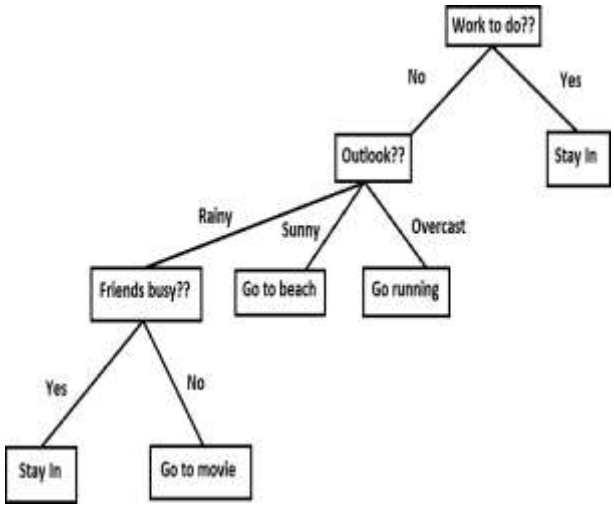

Fig. 2: Example of Decision Tree on what to do whendifferent situations occur in weather.

Splitting is instantly terminated if any data is deemed useless. Finding specific tests is more effective than trying to optimize the tree overall.

Bear in mind that the data set only gives categorical information, and that the ID3 technique can only be simulated using the WEKA tool, when you analyze the properties of Decision Tree. When running simulations, ID3 does not enable continuous data collecting. Several parallels exist between C4.5 and CART.

features identical to those of ID3. C4.5 and CART are similar in that both may use continuous data sets as input for simulation purposes [11], but there is one key distinction.

Table-1: Characteristics of DT.

Decision Tree Algorithm	**Data Types**	**Numerical Data Splitting Method**	**PossibleTool**
CHAID	Categorical	N/A	SPSS answ ertree
ID3	Categorical	No Restriction	WEKA
C4.5	Categorical, Numerical	No Restriction	WEKA
CART	Categorical, Numerical	Binary Splits	CART 5.0

One way to visually display all the potential outcomes and the processes needed to get to each one is via a decision tree [12]. Decision Tree is powerful in its transparency and honesty.You also get to choose the most biased and understandable nature, which is a huge plus. I can also understand it and put it into categories. Able to effortlessly process info that is either continuous or discrete. The decision tree only needs to be able to segment features and filter variables [19]. Performance is affected by non-linear, although the decision tree's parameters are unaffected.

I. ALGORITHMS BASED ON DECISION Determining the "Best" way to split an attribute between two categories is possible using a decision tree approach. It is essential to have a consistent criteria for producing the splits so that the partitions at each branch are as pure as possible.

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 ofthe 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].

I. METRICS

A number of subsets of the training data are created based on the values of the splitting property. The procedure iterates recursively [6] until all instances in a subset belong to the same class in any Decision Tree.

Table- 3: Splitting Criteria

Metrics	Equation
InformationGain	$Information\ Gain = I(p,n) = \left(\frac{-p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{n+p}\right)\log_2\left(\frac{n}{p+n}\right)$
Gain Ratio	Gain Ratio=I(p,n)-E(A) I(p,n)= Information beforesplitting E(A)= Information after splitting
Gini Index	$Gini\ Index, G = \left(\frac{1}{2n^2\mu}\right)\sum_{j=1}^{m}\sum_{k=1}^{m} n_j\, n_k \lvert y_j - y_k \rvert$

Information Gain's primary flaw is its favoritism of characteristics with many variables [6]. Unfair data partitioning occurs when one of the child nodes has an excessively large amount of records in comparison to the rest. Gain Ratios that are higher are better [7, 12]. When data is represented by more than two groups, the Gini Index loses some of its credibility. The problems with splitting criteria are listed here [15].

Information Gain's primary flaw is its favoritism of characteristics with many variables [6]. Unfair data partitioning occurs when one of the child nodes has an excessively large amount of records in comparison to the rest. Gain Ratios that are higher are better [7, 12]. When data is represented by more than two groups, the Gini Index loses some of its credibility. The problems with splitting criteria are listed here [15].

Information Gain's bias toward multivariate characteristics over univariate ones is a major drawback of the model [6]. In an unfair data partition, one of the child nodes contains a disproportionately high number of records compared to the others. More favorable are gain ratios that are larger [7, 12]. When more categories are included in the data, the Gini Index loses its validity. The following issues often arise while trying to divide criteria: [15].

A set is considered exact if and only if its elements are closely packed. Accurate measurements are those in which the mean value of the variables comes near to matching the true value. Only with a collection of data points obtained from many measurements of the same quantity is it feasible to measure quantities with more than 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

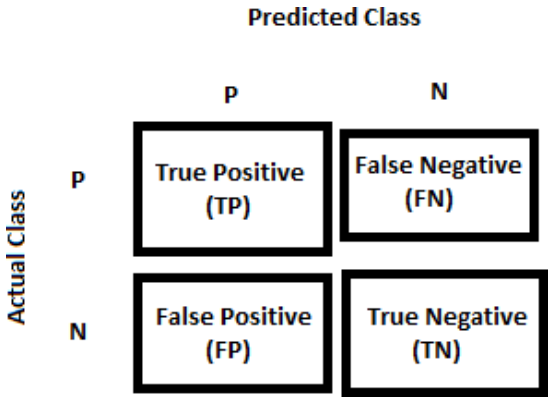

Fig. 3: Confusion Matrix sample in Decision Tree.**II. DESCRIPTION OF THE DATASET**

The vehicle dataset is used for this experiment. Running this dataset through the CART, ID3, and C4.5 decision tree algorithms. Here is how this dataset is defined.The vehicle dataset consists of two sections. Automotive Technology and Popularity. The vehicle's acceptability is affected by a number of elements, including its purchase price and the cost to operate it. Crashworthiness is affected by a variety of factors, including as the quantity of trunk space, safety features, the number of doors, and the size of the passageway (measured in terms of predicted passenger capacity).It appears 1728 times. There are six characteristics. In the absence of an attribute, it serves no purpose.Excellence in Personality Traits:

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%

II. EXPERIMENT We are using WEKA to try to reproduce the experiment. One such tool for data mining projects is the WEKA suite of machine learning techniques. To get data ready for analysis, Weka provides tools for cleaning, sorting, regressing, grouping, discovering associations, and presenting. Weka is free, open-source software that anybody may use according to the terms of the GNU Public License. Additionally, it may be used to create novel methods for machine learning. The algorithms may be

executed on a dataset or call them directly from your Java code [18].

Table- 4: Theoretical results

Algorithm	**AttributeType**	**Missing Value**	**Pruning Strategy**	**Outlier Detection**
ID3	Only categorical values	No	No	Susceptible to outlier
CART	Categorical and Numerical both	Yes	Cost complexity pruning isused	Can handle
C4.5	Categorical and Numerical both	Yes	Error based pruning isused	Susceptible to outlier

This study compares the effectiveness of three decision tree algorithms—ID3, C4.5, and CART—on the same datasets. Results for the three approaches are summarized in the following table [17] according to runtime and accuracy. The splitting Criteria column details how the algorithm was divided to enhance performance. The attribute type column specifies the possible values that the algorithm can handle. You can see whether the algorithm is effective by seeing if it fills in the Missing Value box.

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

The following table displays the output of three algorithms, ID3, C4.5, and CART, in a realistic setting. CART's execution time is 0.5 seconds, ID3's is 0.02 seconds, and C4.5's is 0.06 seconds. CART's execution time is the slowest, while ID3's is the quickest.

Although CART is the slowest of the three algorithms, it provides the most accurate results and is hence the most preferable choice. Based on the data shown in the table above, it seems that CART is the superior algorithm among the three under consideration

.

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

III. CONCLUSION

To process the data, we used the decision tree algorithms ID3, C4.5, and CART. With respect to accuracy, speed, and reliability, decision trees are superior than alternative methods. A lot of people rely on the recommendation system to help them locate relevant material. After much discussion, the article's writers have settled on the conclusion that, when tested on this dataset, CART achieves the highest levels of accuracy and precision among decision tree methods.

REFERENCE

Research Papers

[1]: Yes. Sorower, my dear. Review of the literature on algorithms for multi-label learning. December 18, 2010; Oregon State University, Corvallis, Oregon.

[2]. Yuldiz O. Akcayol, Utku A., and Karacan Hacer (Uke) Utku. Decision-Tree-Based Recommendation System with Implicit Relevance Feedback Implementation. Journal of Safety and Health, 2015, 10(12):1367-74. Published on December 1, 2015.

Written by Gershman, Meisels, Lüke, Rokach, Schclar, and Sturm [3]. A Decision Tree-Based Recommender System. Volume 17, Issue 3, June 3, 2010, International Institute for Cognitive Science.

"Jadhav SD" and "Channe HP" served as references. An efficient recommendation system is a decision tree classifier that uses collaborative filtering. Journal of Engineering Research and Technology: An International Journal, 2016; 3: 2113–2118.

The authors of the article are Nürnberger, Genzmehr, Langer, and Beel. We introduce Docear's method for academic article recommendation here. Delivered as part of the proceedings from the ACM/IEEE-CS

2013 Digital Libraries Conference on July 22, 2013. Modern Language Association (ACM). Zhang X. and S. Jiang. Using a similarity split-off criteria for decision tree learning. JSW published the article in 2012 Aug;7(8):1775-82.
Sharma, G., Mathuria, M., Bhargava, N., Bhargava, R., and N. A decision tree research on data mining utilizing the j48 algorithm. Volume 3, Issue 6, of the International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSEE 2013) was published in June 2013.
strategies for serial decision tree classification and comparing them. [8] Myanwu MN, Shiva SG. Volume 3, Issue 3, June 2009, Pages 230-140, Journal of International Computer Science and Security
.