CART, J-48graft, J48, ID3, Decision Stump and Random Forest: A comparative study

*Taram Nayab Shah, Muhammad Zakir Khan, Mumtaz Ali, Bilal Khan, Nauman Idress
City University of Science and Information Technology, Peshawar.
*Taramshah82@gmail.com, zakir@cusit.edu.pk, mumtaz@cusit.edu.pk, bilalsoft63@gmail.com, khanouman1@gmail.com

Abstract*: Data mining techniques are very useful in the discovery of the hidden knowledge amid a huge amount of data, as will for merging similar data objects. Different algorithms and techniques are available for data mining like clustering, classification, association rule mining and neural networks to solve problems of data discovery and arrangement. The discussed algorithms are supervised learning whose labels are defined, in which classification is the most well-known method. Classification is a widely used method for a number of useful applications like artificial intelligence, credit card rating and fraud detection, etc. A number of Weka classifiers families are available such as Bayes, Lazy, functions, Meta, misc, rules and tree with their own pros and cons. Amid these algorithms, decision trees are the simplest and easiest algorithms for understanding, decision making and to compare with others due to hierarchal structure in nature. There is a number of Decision Tree algorithms used and employed by researchers that are available in the literature. However, this study focuses on the comparison of six decision tree algorithms that are CART, J-48graft, J48, ID3, Decision Stump and Random Forest. The objective of this study is to compare various decision tree algorithms to conclude the best algorithm for a particular problem using Python and Weka tool.*

Keyword: Decision Trees, Classification, Weka, Python

1. Introduction

Data mining is the process of sorting large datasets into meaningful forms to improve understanding and establishes a relationship to solve a problem of data cleansing, clustering, and classification, etc. using data analysis techniques. The basic idea of data mining techniques is summarizing and analyzing the data from different directions and dimensions. Recently data is increasing exponentially and to deal with such a huge amount of data needs data mining techniques to search, analysis and order to find the meaningful patterns within that data [1]. A large number of techniques are using in research either supervised or unsupervised. Classification is described as supervised learning in which the class labels are known in advance while clustering is described as unsupervised learning due to unknown labels for classes. Classification is the most prominent problem in machine learning and data mining problem [4], it is preferred over clustering technique due to class labels are known and can easily select features which help in accurate prediction. In classification, data is of two types first is training and the second is testing. Training data is used to build the model while testing data is to evaluate that model. The trained model is then used for a new set of data which is different from both test and train data [3, 4]. Hence, this paper works only the decision tree classification algorithms due to their hierarchal nature, diversity, and simplicity, these are Simple CART [7], J48graft [6], J48 [2], ID3 [8], Decision Stump [13] and Random Forest [11]. These algorithms are employed on different datasets e.g. Weather, Mushroom, Nursery, Vehicles, and Zoo, taken from UCI and KEEL data repositories. Decision Trees are also known as a statistical classifier that frequently used in Classification problems. The decision tree divides the features into partitions which help to reduce the recursion at every stage of the same stage. The root node is considered to examine the tree and to predict the data of the new label [16]. The experimental results show that J48 performs well on all those sets as compared to other algorithms. The experiments are taken using the WEKA tool and Python.

2. Related Work

2.1 Decision Tree

A decision tree is an effective approach to classification for supervised learning. This is an induction technique which works on continuously partitioning of the data set into a breadth-first search or depth-first search approach [5, 11] until it reaches to the particular class. It is a tree-type structure in which the inner node is represented by an oval and the leaf is represented by a rectangle.

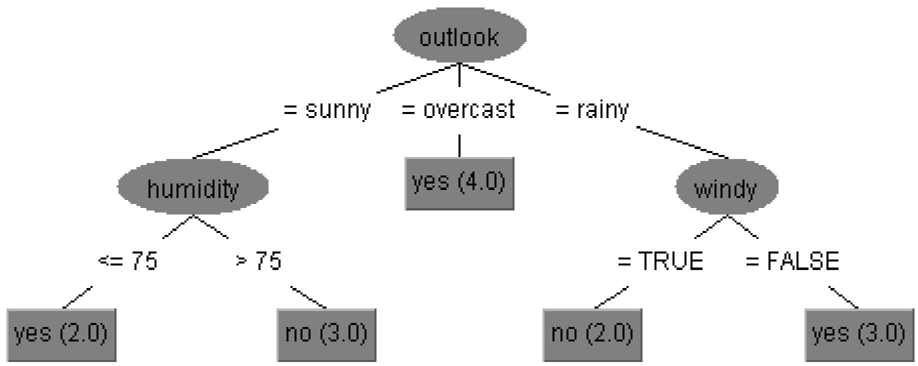

Figure 1. Example of Decision Tree Illustration

The Decision tree consists of three portions. The top from where splitting starts is a root node. The leaf is the ending node and the intermediate node that is used for further splitting. The edge represents the path between the two nodes shows the outcome of the test while the intermediate node represents the test on the attributes and it this node again the decision is taken until impurities measures [6] and the leaf node represent the decision of the class or holding the class label. The decision tree algorithm consists of modeling techniques that simplify the classification process and comprehended for human understanding [3].

2.1.1 Splitting Criteria

All Decision tree algorithm need splitting of a node to build a tree but most of the time the splitting function is univariate (mean the internal node is splitting according to the single attribute detail in [12]). Each algorithm tries to split the tree with the best attribute that is dominated by applying various criteria like the Gini index, information gain, gain ratio, etc. The aim of splitting is to reduce node impurity.

2.1.1.1 Entropy

Entropy is the measurement of uncertainty associated with a random variable for the impurity of the node.

$$\textbf{Entropy (t)} = -\sum P(I/t)\log_2 p(I/t) \qquad 1$$

2.1.1.2 Gini Index

Gini Index is used to determine attributes that are generating the branch. Gini index measures the impurity between the target attribute value and the probability of distribution. It has been widely used by many researchers in their work and it is explained in [13].

$$\textbf{Gini Index} = 1 - \sum [P(I/t)^2] \qquad 2$$

2.1.1.3 Classification Error

$$\textbf{Classification Error} = 1 - \max [P(I/t)] \qquad 3$$

Where p, (i/t) mean the division of records belonging to class I in node t.

2.2 Decision Tree Algorithms

Number of algorithms for classification has been used and using so far like **Bayes** (Bayes Net, Naïve Bayes, Naïve Bayes Multinomial, Bayes Multinomial Updatable, Naïve Bayes and Naïve Bayes Simple) , functions (Gaussian Processes, Linear Regression, Logistic, Multilayer Perceptron, SGD, SGD Text, Simple Linear Regression, Simple Logistic, SMO, SMO reg and Voted Perceptron), Lazy (IB1, IBK, K Star and LWL), meta, Misc, rules (Decision Table, JRIP, M5Rule, oneR and Part) and Trees (Decision Stump, Hoeffding tree, ID3, J48, J48 graft, LMT, M5P, Random Forest, Random Tree and Simple CART) to solve the particular problems like finding student Performance etc.

We have considered SimpleCart, ID3, J48graft, J48, RandomForest, Decision Stump. SimpleCart which has the property to handle the missing attribute and can handle both categorially and numerical data [7]. The creation of regression trees generates capability where leavers predict the real number instead of the class label. SimpleCART helps to lower the prediction of square error. ID3 is a very simple algorithm for testing the value of the object, it can identify the classification of the objects. This splitting is continuous until it reaches to make a subtree of homogenous objects. J48Graft uses the splitting criteria to split the data by providing the information gain. J48graft removes the biasness of ID3 with the wide decision tree. For continuous values, it performs a binary cut based on entropy gain in one scan of sorted data [9]. J48 is a predictive model of machine learning that decides the dependent variable (target value) based on available data [10]. The different attributes are denoted by the internal node of the decision tree. Random Forest is combining the method of learning for both regression and classification that works on a multitude of decision trees during training time and outputting the class that is the mode of the classification [11] or regression of the decision tree. Decision Stump is the machine learning classification model having a one-level decision tree [13]. Decision Stump is depending upon the type of feature passing as input, if it is a nominal feature, it may contain a leaf for all possible feature values or two leaves.

3. Research Methodology

There is a number of algorithms in the decision tree family but we have chosen six algorithms that are rarely used on these datasets and having diversity in their nature. All the algorithm belongs to the same hieratical in nature and the simplest to use. It is also called a statistical classifier and frequently used in Classification problems. The nature of the algorithms is very simple by making a tree which is decision rules. The decision tree divides the features into the division which helps to reduce the recursion at every stage of the same stage. The root node is considered to examine the tree and to predict the data of the new label [16].

To compare the various classification algorithms, two different tools are used i-e Weka and Python. The implementation contains two parts. One is through Weka and the other one is through Python. In the Weka tool, the algorithm is being imported but if it is not present in Weka than select K-fold cross-validation, selects the data and run. As Python is concerned, first the libraries are imported and the algorithm is being implemented in IDE, granted the path where datasets are stored and executed the classifier. The dataset is from KEEL and UCI repository download the dataset.

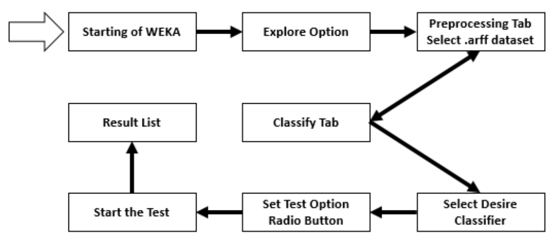

Figure 4. Weka Data Flow Diagram

Figure 6. Python environment preparation

4. Experiments and Result

For comparison of various classification algorithms, five datasets have been selected that come from two different repositories i-e KEEL [14] and UCI [15]. The summary of the datasets is given in Table 2

Table 2. Dataset Information

Dataset Name	No. of Instances	No of Attributes	No. of Classes	Test Method
Weather	14	5	2	2-CV
Mushroom	8124	23	6	10-CV
Nursery	12960	9	5	10-CV
Vehicles	846	19	4	10-CV
Zoo	101	18	7	2-CV

In the above datasets, 2-CV (Cross-validation) and 10-CV (Cross-validation) test methods are used. 2-CV for those which contain less than 500 instances and 10-CV for those which contain more than 500 instances. The performance for the given datasets is shown.

Table 3. Comparison of various Decision tree algorithms on Weather Dataset

Algorithm	Accuracy	MAE	Precision	Recall	F-Measure	ROC	Time Taken
ID3	42.857%	0.5714	0.469	0.429	0.440	0.422	0.02
CART	50%	0.5347	0.375	0.500	0.429	0.400	0.02
J48graft	**50%**	0.5442	0.375	0.500	0.429	0.311	0.02
J48	50%	0.544	0.375	0.500	0.429	0.311	0.01
Decision Stump	50%	0.5143	0.375	0.500	0.429	0.478	0.0009
Random Forest	35.7143%	0.5165	0.321	0.357	0.338	0.422	0.03

Table 4. Comparison of various Decision tree algorithms on Mushroom Dataset

Algorithm	Accuracy	MAE	Precision	Recall	F-Measure	ROC	Time Taken
ID3	100%	0	1	1	1	1	0.8
CART	99.9385%	0.0011	0.999	0.999	0.999	0.999	4.65
J48graft	100%	0	1	1	1	1	0.29
J48	**100%**	0	1	1	1	1	**0.02**
Decision Stump	88.6755%	0.1912	0.898	0.887	0.886	0.883	0.02
Random Forest	100%	0.0004	1	1	1	1	0.38

Table 5. Comparison of various Decision tree algorithms on Nursery Dataset

Algorithm	Accuracy	MAE	Precision	Recall	F-Measure	ROC	Time Taken
ID3	98.1867%	0.0018	0.995	0.996	0.995	0.991	0.04
CART	99.5756%	0.0019	0.996	0.996	0.996	0.999	2.92
J48graft	97.0756%	0.0153	0.970	0.971	0.970	0.995	0.21
J48	97.0525%	0.0153	0.970	0.971	0.970	0.995	0.02
Decision Stump	66.25%	0.1429	0.496	0.663	0.551	0.828	0.02
Random Forest	99.0664%	0.0295	0.990	0.991	0.991	1.000	0.56

Table 6. Comparison of various Decision tree algorithms on Vehicles Dataset

Algorithm	Accuracy	MAE	Precision	Recall	F-Measure	ROC	Time Taken
ID3	89.35%	0.02	0.964	0.962	0.9630	0.946	0.06
CART	69.1489%	0.1667	0.683	0.691	0.686	0.850	0.26
J48graft	72.4586%	0.142	0.715	0.725	0.718	0.861	0.06
J48	72.5768%	0.1415	0.721	0.726	0.723	0.862	0.03
Decision Stump	40.1891%	0.3374	0.282	0.402	0.297	0.638	0.04
Random Forest	76.0047%	0.1567	0.752	0.760	0.755	0.936	0.24

Table 7. Comparison of various Decision tree algorithms on Zoo Dataset

Algorithm	Accuracy	MAE	Precision	Recall	F-Measure	ROC	Time Taken
ID3							
CART	40.5941%	0.217	0.165	0.406	0.234	0.491	0.14
J48graft	92.0792%	0.0256	0.929	0.921	0.920	0.955	0.02
J48	**92.0792%**	0.0256	**0.929**	0.921	0.920	0.955	**0.01**
Decision Stump	59.4059%	0.1368	0.454	0.594	0.492	0.821	0.01
Random Forest	92.0792%	0.12	0.928	0.921	0.908	0.998	0.01

The final Accuracy Table of all the algorithms showed in table 2 is clarified in Table 8:

Table 8. Comparison of various algorithms Accuracy on the given datasets

Dataset Size	Algorithms Name and Accuracy					
	ID3	**CART**	**J48graft**	**J48**	**Decision Stump**	**Random Forest**
14	42.857%	50%	50%	50%	50%	35.7143%
8124	100%	99.9385%	100%	100%	88.6755%	100%
101	98.1867%	99.5756%	97.0756%	97.0525%	66.25%	99.0664%
12960	89.35%	69.1489%	72.4586%	72.5768%	40.1891%	76.0047%
846	0	40.5941%	92.0792%	92.0792%	59.4059%	92.0792%

These tables having different records are simplified by the algorithms resulting in different values that show the best algorithm regarding the accuracy. The graph illustrated below shows the accurate algorithm and clarifies the weak and strong algorithm.

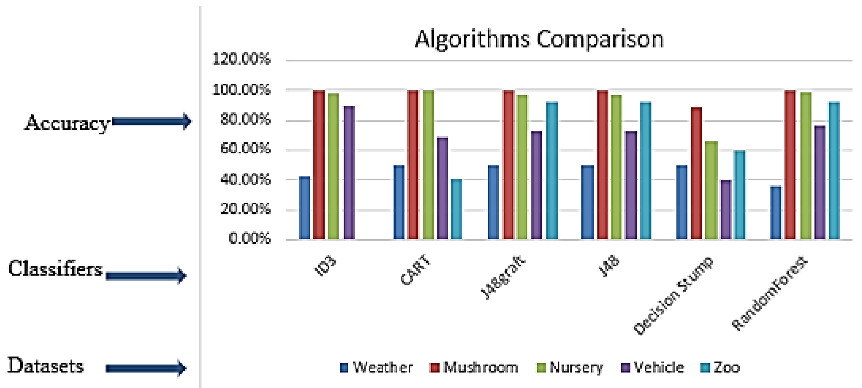

Figure 6. Classification result for Decision Tree Algorithm

5. Conclusion

Decision tree algorithms are one of the most frequently used algorithms in classification for different classification problems. We have considered only the decision tree algorithms. We have done various classification algorithms comparative analysis to know which one is better. For that purpose, we have applied on different datasets taken from UCI and KEEL repository. All algorithms showed better results but J48 win from all with respect to Precision and Time took to make a decision tree. There is always trade-off some algorithms showed better results. Out of five datasets of different sizes, J48 shows better than ID3 with respect to Accuracy and Precision but ID3 showed better Performance, Precision and Less time is taken in one Dataset. Random forest showed relevantly better than remaining. Out of six algorithms of the decision tree, J48 showed better performance. Similarly, we have compared Decision tree algorithms with other classification algorithms still decision tree algorithms showed better performance and Accuracy with the minimum time taken.

6. Acknowledgment

This research was supported by Mr Muhammad Zakir Khan. We thank our colleagues from City University of Science and Information Technology, who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001

[2] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, vol (1), 81-106

[3] Utgoff, P. E., & Brodley, C. E. (1990). An incremental method for finding multivariate splits for decision trees. Proceedings of the Seventh International Conference on Machine Learning , 58-65.

[4] Brijesh Kumar baradwaj and Saurabh pal (2011) “Mining educational data to analyze students’ performance”, (IJACSA) International Journal of Advanced computer science and applications. Vol. 2 no. 6.

[5] Hunt, E.B., Marin. and Stone, P.J. (1966). Experiments in induction, Academic Press, New York.

[6] Quinlan, J. R. (1993). C45: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

[7] Lewis, R.J. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. 2000 Annual Meeting of the Society for Academic Emergency Medicine, Francisco, California

[8] Ahmed Bahgat El Seddawy, Prof. Dr. Turkey Sultan, Dr. Ayman Khedr (2013), “Applying Classification Technique Using DID3 Algorithm to Improve Decision Support Under Uncertain Situations”. “International Journal of Modern Engineering Research “, Vol 3, Issue 4, July- Aug 2013 pp-2139-2146.

[9] Ajanthan Rajalingam, Damandeep Matharu, Kobi Vinayagamoorthy, Narinderpal Ghoman (2002), “Data Mining Course Project: Income Analysis” SFWR ENG/COM SCI 4TF3.

[10] Wikipedia contributors, ―C4.5_algorithm, ‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 28 Jan-2019.

[11] L. Breiman (2001), “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32.

[12] LiorRokach and OdedMaiman, “Chapter 9 Decision Trees “Department of Industrial Engineering,” Tel-Aviv University ”.

[13] Wikipedia contributors (2012), “Decision_stump,” Wikipedia, The Free Encyclopedia. Wikimedia Foundation.

[14] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L.Sánchez, and F. Herrera (2011), “KEEL Data - Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework.,” J. Mult.-Valued Log. Soft Comput., vol. 17.

[15] A. Asuncion and D. Newman (2007), UCI machine learning repository Irvine.

[16] D. Oreski, S. Oreski, B. Klicek, “Effects of dataset characteristics on the performance of feature selection techniques”, Applied Soft Computing, Vol. 52, pp. 109-119, 2017