An Elusive Study of Decision Tree Classifiers in Machine Learning

Chetna Sharma[1] Aman Kumar Sharma[2]
[1,2]Department of Computer Science &Engineering
[1,2]Himachal Pradesh University, Shimla, India

Abstract—*In Machine Learning, Classification is included in significant techniques with applications in Artificial Intelligence, Fraud Detection, Medical Diagnosis and many other disciplines. Classifying objects contingent on their characteristics in the default categories is the broadly studied problem. Classification is the problem of automatically assigning an object to one of the default categories based on object attributes. Some of the commonly used classification algorithms are linear classifiers, boosted trees, logit regression, neural networks, k-nearest neighbors and decision trees. These decisions tree gives a demonstrating system that is straightforward for humans and simplifies the process of classification. Decision trees are exceptionally helpful for identification of disease as decision trees are widely used for the diagnosis of ovarian cancer, heart related issues and bosom cancer using ultrasonic images. This paper endeavors to give a similar examination of four generally used classifiers namely ID3, C4.5, CART and Random Forest with an experimental approach using Scikit-learn tool in Anaconda Environment on Breast Cancer dataset regarding classification and prediction accuracy.*

Keywords—Machine Learning, Classification Accuracy, Decision Tree, Iterative Ditchotomizer3, C4.5, CART, Random Forest, Scikit-Learn

I. INTRODUCTION

Machine learning methods are utilized for data analysis and pattern detection [1]. In machine learning computers are programmed to learn from data. They play a major part in the development of data mining applications [25]. Datasets in machine learning applications contain arrangement of components that have their own traits, so grouping procedures have been created to address many items that have several attributes [7]. In machine learning, classification is the essential errands and has been broadly considered in data mining, statistics, neural networks and systems experts for decades [9]. The entry for classification is an arrangement of training record instances, where each record has different attributes and a distinct attribute named class tag. The template is used to estimate class tags from unknown objects [16].

Classification is otherwise called "supervised learning", since the learning of model is "supervised", that is, each training example is labeled indicating its class [16]. Classification has been effectively applied to an extensive variety of application fields, for example, logical analyses, restorative diagnostics, climate gauges, credit endorsement, client division, target promoting and extortion recognition [17]. Decision tree classifiers are utilized widely for the finding of ovarian growth, solid breast or bosom tumor and heart analysis in ultrasonic pictures [21].

Machine Learning with Decision trees assumed to be imperative part in medicinal finding to analyze the symptoms of a patient [6]. In this paper, efficiency of different classifiers is validated using medical dataset. Further, decision tree techniques are selected as they are: easy to decipher, quick generation of trees and yields better precision [8].

The remaining paper is divided in three sections. In Section 2 the survey of decision tree classifiers are elaborated. Section 3 discusses test investigation and examinations of different decision tree classifiers. Section 4 gives conclusions and future scope.

II. DECISION TREE CLASSIFICATION

The section illustrates briefly the classification of decision trees, Iterative Ditchotomizer3, C4.5, Random Forest, CART decision tree classifiers briefly.

The decision tree performs classification in two phases [10]:

- Growing (or building)
- Pruning (cutting back)

1) Growing

The tree is built by dividing the training set according to optimal criteria until all or most of the records refer to most of the partitions that carry the same class tag.

2) Pruning

The pruning stage sums up the tree by expelling the commotion or noise and the anomalies. The classification seems to be more accurate.

The structure of decision tree is given in two stages as under:

```
BuildTREE(dataset S)
If all records in dataset S have a place with same class;
//Growing Phase
Return;
For each attribute Ai, assess splits on t attribute Ai;
//Pruning Phase
Utilize best fit found to section dataset S;
buildTREE(S1);
buildTREE(S2);
ENDBuildTREE;
```

Decision Trees are implemented in parallel and serial. The parallel form of implementation is desirable to ensure a fast formulation of results, considerably with the order/forecast of substantial or big datasets [12]. In any case, when small datasets are included, the serial implementation is used.

A. ID3

ID3 (Iterative Ditchotomizer3) is a basic classifier created by Ross Quinlan in 1983. The ID3 calculation depends on the system named Concept Learning System (CLS) algorithm. The CLS is the fundamental algorithm for decision tree learning. The development period of the CLS is done by selecting attribute to test each and every node by the trainer. ID3 enhances the CLS by adding a heuristic attribute selection [14]. In the first phase of the tree generation, ID3 classifier is utilized for information gain, entropy, to select the best division or split attribute. ID3 does not give a correct outcome when there is excessive noise in the training data, so extensive pre-handling of the

data is done before building a decision tree [18]. One of the primary downsides of ID3 is that the measure gain tends to support attributes with different values [19].

B. C4.5

C4.5 is also given by Ross Quinlan in 1993. C4.5 is enhanced classifier adaptation of ID3, this classifier uses Gain Ratio as a division criteria [13]. This classifier handles both ceaseless and discrete attributes. The primary favorable circumstances of C4.5 is when assembling a decision tree, it can manage the datasets that have guidelines with unknown values. This classifier handles training data with attribute values by permitting those values to be assigned as missing. Missing attributes are just not utilized as a part of gain or entropy calculations [14]. It has an improved strategy for tree pruningthat diminishes misclassification mistakes because of noise or a lot of detail in the dataset.

C. Cart

The CART classifier stands for Classification and Regression Trees. It is an algorithm for the exploration and estimation of data. In the early 80's, the CART was developed by Leo Breiman, Jerome Friedman and then later accompanied by Richard Olshen and Charles Stone, who began working with the decision tree in Southern California. The CART is described by the way it fabricates binary trees which implies that each inner node has precisely two active edges while both ID3 and C4.5 classifiers generate the decision trees with multiple branches per node [3]. CART is different from other classifiers, as it is can also generate regression trees. The CART uses the Gini index for the division procedure.The CART mediation includes automatic class rolling (optional), cost-sensitive learning, automatic value missing direction, dynamic property construction [20].

D. Random Forest

The general method for random forest was first suggested in 1995 by Ho, who confirmed that forests of trees breaking with slanting hyper-planes, if it is randomly confined to be susceptible to chosen feature dimensions, can obtain accurately and grow without overtraining [21]. Random Forest is ensemble of pruned binary decision tree, unlike other it generates numerous trees which creates forests like classification [22]. Ensemble learning method of the random forest is very promising technique in terms of accuracy. In tree development period of the standard trees above explained ID3, C4.5, CART all nodes utilizes the best split among all variables [23].

E. Comparison

Classifiers	Entropy	Information Gain	Gini Index
ID3	✓	✓	
C4.5		✓	
CART			✓
Random Forest			✓

Table 1: Comparison of Classifiers on the Basis of Study

Table1 demonstrate the results of four decision tree classifiers namely ID3, C4.5, CART and Random Forest based on the literature study. In which random forest gives better productionwhen contrasted with that of other classifiers. This method is computationally effective, does not over fit, is vigorous to noise and can also be enforced when the number of variable is significantly larger than number of samples [24].

III. EXPERIMENTAL RESULTS

In this section the four decision tree classifiers namely ID3, C4.5, CART and Random Forest are compared based on their Accuracy, Learning Time and Tree Size. The simulations were conducted using a Breast Cancer datasettaken from the UCI Repository:www.archive.ics.uci.edu/ml/datasets.html. This data set incorporates two classes with 201 and 85 instances.

Scikit-learn is a free and open source library software in Python used for machine learning. It is a simple and efficient tool for data mining and data analysis. It is based upon numpy, scipy and matplotlib. It is taken from the repository https://github.com/scikit-learn/scikit-learn.

This library incorporates a well ordered tools for classification, regression, clustering, and dimensionality reduction in machine learning. It is notified that scikit-learn is used to create models.

For this, comparison with the help of tool there are few parameters used like:

- Accuracy
- Learning Time
- Tree Size

A. Accuracy

It is the parameter used for testing the samples which are correctly classified. As accuracy is used for comparing different approaches, considering the experimental results given in Table 2 where Random Forest is better than other single tree decision tree algorithms.

B. Learning Time

It is the time taken for learning and generating decision trees. As given in Table 2 the time is given in mille-seconds. CART takes less learning time than C4.5, Random forest and ID3.

C. Tree Size

Size relies on the number of operations whichhas tobe done for classification. Therefore, more the number of operations, more the size of the tree for the classification process. Therefore it clarifies, the approach reduces the classification time.

Classifiers	Accuracy	Learning Time	Tree Size
ID3	98.23%	190ms	156
C4.5	97.34%	145ms	189
CART	99.23%	123ms	185
Random Forest	99.24%	152ms	130

Table 2: Comparison of Classifiers using Tool

It is observed from the Table 2 that Random forest has the highest classification accuracy (99.34%) lowest tree size and learning time. The second highest classification accuracy for CART classifier is 99.23% moreover ID3 have 97.34% and C4.5 classifier results in lowest accuracy followed with tree size and learning time which is 98.23% among four classifiers.

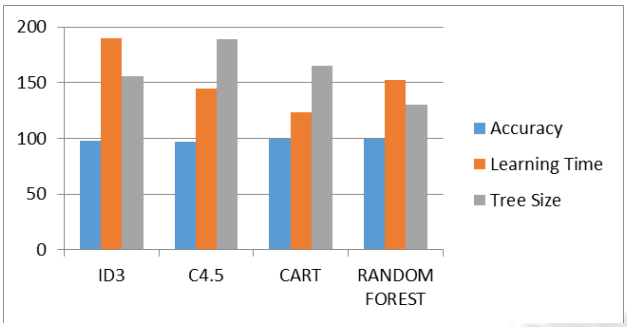

Fig.1: Comparison of Classifiers Accuracy

The Fig.1 shows graphical representation of experimental results. It can be found in the graph that the parameter values of Random Forest are much closer to the CART algorithm followed by C4.5 and ID3 which lack behind. But Random Forest is more efficient than CART algorithm as it ensemble whole forest taking less amount of time than CART algorithm which generates only single decision tree.

IV. Conclusions & Future Scope

In the study experiments are carried out to find the classification accuracy of four classifiers in terms of which classifier better determine whether a particular symptom leads to breast cancer or not with the help of attractive machine learning tool Scikit-learn. Four classifiers namely ID3, C4.5, Random Forest and CART were compared on parameters such as accuracy, learning time and tree size. All these four fall under classification methods of machine learning which makes relationship between dependent (output) variable and independent (input) variable.

It is clear from the simulation results that Random Forest classifier has the highest accuracy, less learning time and tree size. Future work may include the improvisation of Random forest in scikit-learn tool.

References

[1] Andreas C Holzinger, "Machine Learning and Knowledge Extraction", International Cross-Domain Conference (ICDC), Italy, Springer, 978-3-319-66808-6, 2017.

[2] Bevinda Alisha Pereira, Anusha Pai, Cassandra Fernandes, "A Comparative Analysis of Decision Tree Algorithms for Predicting Students Performance", International Journal of Engineering Science and Computing (IGESC), Vol. 7, pp. 10489-10492 April 2017.

[3] Breiman, "Classification Algorithms and Regression Trees", International Journal of Man-Machine Studies (IJMMS), Vol. 27, Issue 3, pp. 221-234, 1984.

[4] Charu C Aggarwal, "Data Classification Algorithms and Applications", Data Mining and Knowledge Discovery Series, CRC Press, IBM T. J. Watson Research Center, Yorktown Heights, New York, USA, 2009.

[5] D.Lavanya, "Performance Evaluation of Decision Tree Classifiers on Medical Data", International Journal of Computer Applications (IJCA), 0975 – 8887, Vol. 26, pp. 1-4, July 2011.

[6] Gang Zheng, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", Research Gate, IEMS Vol. 4, No. 1, pp. 102-108, June 2005.

[7] J. G. Carbonell, "Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification and Recognizing Textual Entailment", Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April11-13, 2005.

[8] Joaquín Abellán, "A comparative study on base classifiers in ensemble methods for credit scoring", Elsevier Journal, Expert Systems With Applications, 0957-4174, pp. 1-10, 2017.

[9] Joe el Quinque Ton, "Emergence in Problem Solving, Classification and Machine Learning", Proceedings of the 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC),0-7695-2740, Vol. X, 2006.

[10] J R Quinlan, "Induction of Decision Trees", Centre for Advance Computing Sciences, Kluwr Acedamic Publishers, Boston, pp. 81-106, 1986.

[11] J R Quinlan, "Learning Decision Tree Classifiers", ACM Computing Surveys, Vol. 28, No. 1, pp. 71-72, March 1996.

[12] Leonard A. Breslow, "Simplifying Decision Tree: A Survey", NCARAI Technical Report No. AIC-96-014, pp. 1-47 1997.

[13] Lior Rokach, "Top-Down Induction of Decision Trees Classifiers—A Survey", IEEE Transactions on Systems, MAN, And Cybernetics—Part C: Applications And Reviews, Vol. 35, NO. 4, pp. 476-487, November 2005.

[14] Lomax. S, "A survey of cost-sensitive decision tree induction algorithms", ACM Computing Surveys, pp. 34-44, 2013.

[15] Miroslav Kubat, "An Introduction to Machine Learning", 2nd edition, Springer International Publishing AG, ISBN 978-3-319-63912-3, 2017.

[16] Peter W. Eklund, "A Performance Survey of Public Domain Supervised Machine Learning Algorithms", Research Gate publications, School of Information Technology, Griffith University, Australia, 1998.

[17] Raj Kumar, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET), ISSN: 2319 – 1058, Vol. 1, pp. 1-8, 2 August 2012.

[18] Rodrigo C. Barros, "Automatic Design of Decision-Tree Induction Algorithms", Springer Briefs in Computer Science, 2015.

[19] Rodrigo C. Barros, "A Survey of Evolutionary Algorithms for Decision Tree Induction", IEEE Transactions on Systems, MAN and Cybernetics – Part C: Applications & Reviews, Research Gate, Vol X, and January 2012.

[20] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica, pp. 249-268, 2007.

[21] Songul Cinaroglu, "Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures", International Journal of Computer Applications, ISSN: 0975 – 8887, Volume 138 – No.1, pp. 37-41, March 2016.

[22] S. Rasoul Safavian, "A Survey of Decision TreeClassifier Methodology", School of Electrical EngineeringPurdue University, West Lafayette, TR-EE 9054, September 1990.
[23] Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No.6, pp. 1890-1895, 5 May 2011.
[24] Tjen-Sien Lim, "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms", Machine Learning, Kluwer Academic Publishers, Netherlands, pp. 203–228, 2000.
[25] Venkatadri .M, "A Comparative Study on Decision Tree Classification Algorithms in Data Mining" International Journal of Computer Applications in Engineering, Technology and Sciences (IJ-CA-ETS), ISSN: 0974-3596, Volume 2, pp. 24-29, April 2010.