INTERNATIONAL RESEARCH JOURNAL OF HUMANITIES AND INTERDISCIPLINARY STUDIES
(Peer-reviewed, Refereed, Indexed & Open Access Journal)
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 6.865 (SJIF 2023)

A Comparative Study of Recursive Partitioning Algorithms (ID3, CART, C5.0) for Classification

Abhijeet D. Mankar
Assistant Professor,
Department of Computer Science,
Tuljaram Chaturchand College of Arts, Science and Commerce (Autonomous), Baramati,
Dist. Pune (Maharashtra, India)
E-mail: abhijeetmankar@gmail.com

Dr. Sudhakar D. Bhoite
Associate Professor,
Chhatrapati Shahu Institute of Business Education and Research,
Kolhapur (Maharashtra, India)
E-mail: sdbhoite05@gmail.com

DOI No. 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/03.2023-16164445/IRJHISIC2302052

Abstract:

Classification is a technique used to predict categorical dependent variables. Classification is known as supervised learning since class labels are known in advance. It has applications in various areas. In this paper, we have performed a comparative study of ID3, CART, C5.0 decision tree algorithms. Decision tree algorithms are effective and easy to interpret as compared to other classification algorithms. We have used free datasets from UCI machine learning repository from different subject areas to illustrate how the recursive partitioning algorithms differ from one another. We have chosen datasets with small, medium and large number of observations to avoid any bias in the comparative analysis. We have reviewed ID3, CART, C5.0 algorithms and by experimental analysis we focused on the aspects like time taken to build the model, accuracy, depth, and breadth of the resulting classification tree, as well as size of the tree, that is, the total number of nodes.

Keywords: *Classification, decision tree, recursive partitioning, ID3, CART, C5.0*

1. Introduction:

The recursive partitioning approach reflects a tree representation. The prediction is located at the leaf nodes in a tree. Trees designed for classification predicts categorical dependent variable whereas trees designed for regression predicts continuous dependent variable. We discuss the approaches used in recursive partitioning. Development of statistical decision theory and use of network models to represent decision rules led to what is popularly known as decision trees. Researchers then argued that prediction is also some kind of decision and hence it should be possible to use tree structures to represent prediction models. Every domain of application has some data and want to use the available data for its purpose. Advances in data acquisition technologies have resulted in huge amounts of multidimensional data. The classical statistical prediction models are based on the assumption of the homogeneity of data. But in reality, data is of big volume and also it is in various formats. The recursive partitioning of data addresses these two issues. Some researchers have conducted comparative study to know the accuracy of these algorithms. Hetal Bhavsar and Amit Ganatra compared different classification techniques on different aspects like overfitting, noise, speed, accuracy, missing values. Through their comprehensive study of classification techniques, a researcher can gain insight into existing classification techniques and their advantages and disadvantages [2]. Venkatadri. M and Lokanatha C. Reddy in their study have focused on learning time, tree size and accuracy of the algorithms [5].Sayali Jadhav and H. P. Channe compared K-NN, Naïve Bayes and decision tree techniques and their advantages and disadvantages [3].In the study performed by Venkatadri. M, Lokanatha C. Reddy and Sayali Jadhav, H. P. Channe, Weka tool was used. In our study we have considered depth and breadth, and the tree size along with the accuracy of these algorithms. We have used RStudio tool for experimental analysis.

2. Recursive partitioning algorithms:

CART algorithm was developed by Leo Breiman et al. CART follows the greedy search approach. CART prefers pruning the tree instead of stopping rules. CART can handle both continuous and nominal attributes data. CART can also deal with missing values. The CART algorithm is used for classification and regression. CART uses GINI index for Classification and sum of squared residuals for Regression. CART algorithm for classification is implemented in R Studio through rpart package with split="gini".

Ross Quinlan developed the ID3 algorithm in 1986. The author developed an approach to synthesize decision trees and describes ID3 (Iterative Dichotomiser) in detail. ID3 is iterative in nature. ID3 allows upto two classes for any induction task. It is used for classification. It uses entropy and information gain. ID3 algorithm is implemented in RStudio through rpart package with split="information".

C5.0 was proposed by Ross Quinlan in 1994. It provides several improvements on C4.5. C5.0 is used for classification and entropy is used in this algorithm. In speed comparison, C5.0 is faster than C4.5. The C5.0 decision trees are comparatively smaller as compared to C4.5 with matching results. C5.0 algorithm is implemented in R Studio through C50 package.

3. Data and Methods:

3.1 Data:

The datasets used in this study are downloaded from UCI Machine Learning Repository. We have chosen datasets containing small, medium and large number of observations. This is because we want to avoid any bias in experimental analysis. The datasets used in this study are from different subject areas. The description is given in the Table 1.

Table 1: Datasets used in the study

Name of the dataset	No. of observations	No. of variables	Response variable	Type of response variable	Subject area(as mentioned in UCI ML repository)
Obesity estimation	2111	17	Nobeyesdad	categorical	Life
Bank marketing	41188	21	Y	binary	Business
Adult	32561	15	>50K, <=50K	binary	Social
Mushroom	8123	23	class	binary	Life
Glass identification	214	11	Type_of_glass	categorical	Physical
Statlog(german credit data)	1000	21	class	categorical	Financial
Nursery	12960	9	class	categorical	Social

3.2 Methods:

ID3, CART and C5.0 are applied on the datasets mentioned in the Table 1. For implementation purpose RStudio2022.12.0+353 version is used. The tree structure is obtained after applying these algorithms on the datasets.

4. Experimental Analysis:

We performed experiment to compare the decision tree algorithms. We chose seven datasets from different subject areas from UCI machine learning repository. We performed experiment on these datasets using RStudio 2022.12.0+353 version. The time taken to build the model on all seven datasets along with the accuracy are mentioned in the Table 2.

Table 2 : Time taken to build the model, accuracy and data size of the model

Dataset	Algorithm	Time taken to build model(using system. time function)	Accuracy	Data size (no. of rows* no. of columns)
Obesity estimation	ID3	0.14	0.9147	35887
	CART	0.04	0.8799	
	C5.0	0.10	0.9494	
Bank marketing	ID3	0.86	0.918	864948
	CART	0.97	0.9167	
	C5.0	1.66	0.9149	
Adult	ID3	0.97	0.8479	488415
	CART	0.75	0.8479	
	C5.0	1.45	0.8684	
Mushroom	ID3	0.08	0.9955	186829
	CART	0.09	0.9955	
	C5.0	0.28	1	
Glass identification	ID3	0.02	0.5938	2354
	CART	0.02	0.5469	
	C5.0	0.02	0.4844	
Statlog (german credit data)	ID3	0.05	0.7067	21000
	CART	0.03	0.6967	
	C5.0	0.05	0.7333	
Nursery	ID3	0.06	0.8786	116640
	CART	0.08	0.8786	
	C5.0	0.36	0.9933	

We have also compared the tree representation produced by the three recursive partitioning algorithms and summarized it in the Table 3.

Table 3: Depth, Breadth, Tree Size and Shape

Dataset	Algorithm	Depth	Breadth	Tree size	Shape=Breadth/Depth
Obesity	ID3	7	20	39	2.85

estimation	CART	8	18	35	2.25
	C5.0	12	69	135	5.75
Bank marketing	ID3	4	8	15	2.00
	CART	4	7	13	1.75
	C5.0	16	336	559	21.00
Adult	ID3	3	5	9	1.66
	CART	3	5	9	1.66
	C5.0	25	112	204	4.48
Mushroom	ID3	2	3	5	1.50
	CART	2	3	5	1.50
	C5.0 (removing 'veil-type' variable)	5	7	12	1.40
Glass identification	ID3	7	11	21	1.57
	CART	7	10	19	1.42
	C5.0	8	22	43	2.75
Statlog (german credit data)	ID3	5	7	13	1.40
	CART	8	12	23	1.50
	C5.0	13	72	123	5.53
Nursery	ID3	5	6	11	1.20
	CART	5	6	11	1.20
	C5.0	12	160	300	13.33

5. Conclusion:

In this study we compared the performance of recursive partitioning algorithms in terms of accuracy. We also compared the tree representation produced by these algorithms. From this study, we found that C5.0 algorithm takes more time to build the model as compared to ID3 and CART. On the other hand, C5.0 has more accuracy as compared to ID3 and CART. In addition to that, the depth, the breadth and the tree size is more in C5.0 tree. Some researchers have compared these algorithms from performance point of view (accuracy, time taken to build the model). We observe that these algorithms need to be compared from process point of view (comparing internal steps of these algorithms, tree structure).

Acknowledgments:

We acknowledge the suggestions provided by Dr. S. D. Gore. We thank our family members, administrative authorities of our institutions and our friends and colleagues who are constant source of inspiration for us.

6. References:

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone C.J.: Classification and Regression Trees. Chapman and Hall/CRC (1984)

[2] Bhavsar, Hetal & Ganatra, Amit: A Comparative Study of Training Algorithms for Supervised Machine Learning. International Journal of Soft Computing and Engineering (IJSCE). Vol.2, Issue 4 (2012).

[3] Sayali D. Jadhav, H. P. Channe: Comparative Study of K-NN, Naïve Bayes and decision tree classification techniques. International Journal of Science and Research (IJSR), Vol. 5, Issue 1, pp. 1842-1845 (2016).

[4] Quinlan, J.R. : Induction of Decision Trees. Machine Learning, 1, pp. 81-106 (1986).

[5] Venkatadri. M, Lokanatha C. Reddy : A comparative study on decision tree classification algorithms in data mining. International Journal of Computer Applications in Engineering, Technology and Sciences (IJ-CA-ETS), Vol.2 Issue 2, pp. 24-29 (2010).

[6] Estimation of obesity level based on eating habits and physical condition. (2019). UCI Machine Learning Repository.

[7] Moro, S., Rita, P. & Cortez, P.. (2012). Bank Marketing. UCI Machine Learning Repository.

[8] Adult. (1996). UCI Machine Learning Repository.

[9] Mushroom. (1987). UCI Machine Learning Repository.

[10] German, B.. (1987). Glass Identification. UCI Machine Learning Repository.

[11] Hofmann, Hans. (1994). Statlog (German Credit Data). UCI Machine Learning Repository.

[12] Rajkovic, Vladislav. (1997). Nursery. UCI Machine Learning Repository.