International Journal of Engineering Sciences Paradigms and Researches Vol. 08, Issue 01, October 2013
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Behavior of Classification & Clustering Algorithms on Diverse Data

Navneet[1], Dr. Nasib Singh Gill[2]

[1,2]Department of Computer Science & Applications, Maharishi Dayanand University, Rohtak, Haryana, India
[1]navneet.khatri@gmail.com

Abstract

Decision Tree is well known and widely used method of classification. Decision tree is simple and easy to interpret as classification rule. Attribute selection is considered as a challenging job in various data mining applications like student result prediction, medical diagnosis, fraud detection, credit card analysis etc. Various Decision Tree algorithms were developed to ease decision making process. Besides classification there are many clustering techniques like DB Scan, Hierarchical, K-means clustering technique. This paper presents Comparative evaluation among various decision trees algorithms and Clustering Techniques in terms of various criteria of classification accuracy, time taken to build model and size of tree.

General Terms: *Data mining, Decision Tree, Classification algorithms, attribute selection.*

Keywords: *Decision Tree, Classification Algorithms, Split criteria.*

1. Introduction

Generation of decision tree has been used as a method of machine learning for efficient acquisition of knowledge from massive amount of data, simplicity and comprehensibility. Decision Tree learning is typically a greedy, top down and recursive process starting with entire training set and empty tree. There are various Decision Tree algorithms like C4.8 called as J48 in WEKA, REP, Naïve Bayes Tree, Simple Cart. C4.5 is a typical Decision tree algorithm developed by Quinlan in 1993[2].Whole dataset is divided into train set and test set. Training set is used to train model and test set is used to deduce result and predict the desired class. Several decision Tree algorithms were developed based on various split criteria's[1].Various models of classification has been developed so far like neural network model, linear model of statistics, decision trees was developed[9][10]. Decision Trees are widely used in various data mining applications as they are fast, simple and easy[11]. In many areas Decision Tree classifier give sometimes better accuracy than others.

For Comparative evaluation among various decision trees algorithms we have taken data set with two hundred instances having various numbers of attributes.

Liangxio surveyed various measures of attribute selection for selecting best split and empirical study on best classification was done and ranking performance of decision trees were given[3].Matthew performed experimental analysis on sample records to evaluate the performance of commonly used serial decision tree algorithms This analysis shows SPRINT and C4.5 algorithms have good classification accuracy[4]. Wray buntine paper conclude that random splitting leads to increased error and performs significantly worse than others[5][6]. John Mingers conducted experiment various domains G-Statistics, chi-square, gain ratio, gini index, random and find out size of trees. Choice of measure significantly influences size of unpruned tree. Gain ratio measure generates smallest tree [7]. D. Sarvana Kumar proposed a system which deals with automated selection of Decision Tree Algorithms based on training data size which takes less time and avoids memory problem [8]. Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both).

2. Decision Tree Algorithms

Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes. The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures. The tree leaves is made up of the class labels which the data items have been group.

Decision tree classification technique is performed in two phases: tree building and tree pruning. Tree building is done in top-down manner. It is during this phase that the tree is recursively partitioned till all the data items belong to the same class label. It is very tasking and computationally intensive as the training data set is traversed repeatedly. Tree pruning is done is a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set). Over-fitting in decision tree algorithm results in misclassification error. Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once.

There are various approaches of decision tree are C4.5 [13], NB, REP, CART[14] are based on different node splitting criteria. Out of these CART and C4.5 are most famous decision tree algorithms for commercial use.

2.1 CART (Classification and Regression Trees)

CART (Classification and regression trees) was introduced by Breiman, (1984) [14]. It builds both classifications and regressions trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt's model of decision tree construction and can be implemented serially (Breiman, 1984)[14]. CART is based on Gini coefficient can only make a binary tree and backward pruning.

2.2 C4.5

C4.5 is a well-known classification algorithm that constructs decision trees of arbitrary depth in a top-down recursive divide-and-conquer strategy with splits maximizing the Gain Ratio [20]. It is biased, however, in favour of continuous attributes, a weakness partly addressed by later improvements [21]. C4.5 employs a pruning technique that replaces sub trees with leaves, thus reducing overfitting. In a number of datasets the accuracy achieved by C4.5 was comparatively high [20, 21]. It is improved form of ID3 and also based on entropy. Over the years several efforts were made to develop a method of decision tree induction with more accuracy. C4.5 is a classification algorithm that induces decision trees and rules from datasets that could contain categorical and numerical attributes. Categorical values of attributes from new records can be predicted by using rules. It is based on divide and conquer strategy.[17]

2.3 Naïve bayes DT

The NB-Tree provides a simple and compact means to indexing high–dimensional data points of variable dimension, using a light mapping function that is computationally inexpensive. The basic idea of the NB-Tree is to use the Euclidean norm value as the index key for high–dimensional points. Thus, values resulting from the dimension reduction can be ordered and later searched in the resulting one–dimensional structure.

It has been widely used in many data mining applications, and performs surprisingly well on many applications. However, due to the assumption that all features are equally important in naive Bayesian learning, the predictions estimated by naive Bayesian are sometimes poor [22]. For example, for the problem of predicting whether a patient has diabetes, his/her blood pressure is supposed to be much more important than his/her height. Therefore, it is widely known that the performance of naive Bayesian learning can be improved by mitigating this assumption that features are equally important given the class value. Many enhancements to the basic naive Bayesian algorithm have been proposed to resolve this problem.

2.4 REP DT

Reduced Error Pruning is an algorithm that has been used as a technique to remove the problems of decision tree learning suffers from the inadequate functioning of the pruning phase, like the size of the resulting tree grows linearly with the sample size, even though the accuracy of the tree does not improve.

In this paper for Experimental setup most popular only three algorithms of Decision Tree are taken- J48, Naive Bayes, REP, Simple CART Decision Tree Algorithm.

3. Clustering Techniques

Clustering is an important technique in the data-mining area [23]. If data mining is considered a knowledge discovery from amount of data, we can obtain useful information from the number of clusters into how many clusters the whole data can be divided, as well as the available maximum sample size and the computational complexity. Clustering is a division of

data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst them and dissimilar compared to objects of the other groups. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data object by few clusters and hence it models by its clusters [24].

There are various Clustering Techniques in Data Mining like DBScan, hierarchical Clustering, K-means clustering.

3.1 DBScan Clustering Technique

DBSCAN algorithm (density-based spatial clustering of applications with noise) discovers clusters of arbitrary shapes and is efficient for large spatial databases. The algorithm searches for clusters by searching the neighborhood of each object in the database and checks if it contains more than the minimum number of objects [15]

3.2 Hierarchical Clustering

Hierarchical clustering which is depicted by a tree or dendrogram. There are two approaches to hierarchical clustering: we can go from the bottom up", grouping small clusters into larger ones, or from the top down", splitting big clusters into small ones. These are called agglomerative and divisive clustering's respectively. [16]

3.3 K-means Clustering

K-means clustering is a method commonly used to automatically partition a data set into k groups. This proceeds by selecting k initial cluster centers and then iteratively refining them as follows:

The k-means algorithm [18] randomly selects k data points as initial means. K clusters are formed by assigning each data point to its closest cluster mean. The algorithm uses the Euclidean distance. Virtual means for each cluster are calculated by using all data points contained in a cluster. Second and third step is iterated until a predefined number of iteration is reached or the consistence of the clusters does not change anymore. The runtime for the algorithm is O(n).

4. Database Sources

All five datasets were taken from WEKA toolkit which represents a wide range of domains and data characteristics shown in Table1 below.

Table 1: Range of domains and data characteristics

Sr.No.	Dataset	No. of instances	No. of attributes
1	Ionosphere	351	35
2	Segment test	810	20
3	Soyabean	683	36

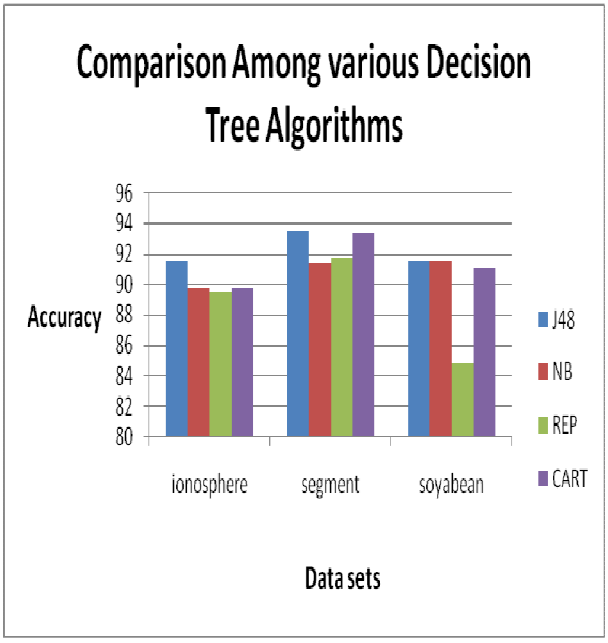
5. Experimental Setup

Experiments were conducted under the framework of Weka to study the various kinds of Decision Tree Classification Algorithms and clustering techniques on these three datasets. We conducted our experiments to compare decision trees results in terms of classification measured by percentage accuracy of no. of correctly classified instances .The accuracy of each tree on each dataset is obtained via 5 runs of 10-fold cross-validation. Runs with the various tree algorithms are carried out on the same training sets and evaluated on the same test sets. In particular, the cross-validation folds are the same for all the experiments on each dataset.

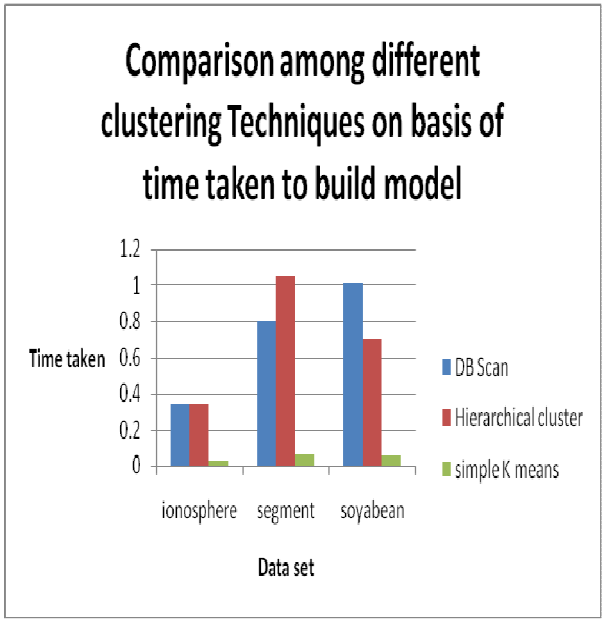
Table 2 shows comparison of various DT Algorithms on basis of Percentage of Correctly classified instances of different trees on each dataset.

Table2: Accuracy among different Decision Tree Algorithms on different datasets

	ionosphere	segment	soyabean
J48	91.45	93.45	91.5
NB	89.74	91.35	91.5
REP	89.45	91.7	84.77
CART	89.74	93.33	91.06



Graph1: Comparison between Various Decision Tree Algorithm



Graph2: Comparison between Various Clustering Techniques

Graph1 shows comparison of various DT Algorithms on basis of Percentage of Correctly classified instances of different trees on each dataset.

Table 3 respectively shows comparison of various Clustering Algorithms on basis of No. of cluster produced and Time taken in seconds by different models of Clustering Techniques on each dataset.

Table3: Comparison among different clustering Techniques on basis of time taken to build model

	ionosphere	segment	soyabean
DB Scan	0.34	0.8	1.01
Hierarchical cluster	0.34	1.05	0.7
simple K means	0.03	0.06	0.05

6. Conclusion

After Comparative analysis on a wide range of domains and data characteristics result found that classification accuracy to develop decision tree model in case of J48 is maximum in all cases of datasets. Time taken to build Model using k mean clustering technique is minimum. Now a days, Decision Tree classifier is the major thrust area in data mining research which ultimately leads to small decision making process. So some new model will be proposed in future by cascading j48 and k mean clustering technique to improve feature of accuracy.

References

[1] Ron Kohavi, Ross Quinlan, "Decision Tree Discovery",1999
[2] Quinlan, J.,"C4.5: Programs for machine learning", Morgan Kaufmann: San Fransisco, 1993.
[3] Liangxizo jiang,"An Empirical Study on Attribute Selection Measures in Decision Tree Learning", 2010, Journal of Computational Information Systems, pp 105-112.
[4] Matthew N.Anyanwu, "Comparative Analysis of Serial Decision Tree Classification Algorithm", International Journal of Computer Sc. And security IJCSS, vol3, issue3, 2009.
[5] Leo Brieman,"Technical note: Some Properties of Splitting Criteria", Machine Learning24, 41-47, 1996.

[6] Wray Buntline, "A Further Comparison of splitting rules for decision tree induction", Machine Learning, pp 75-85, 1992.
[7] John Mingers,"An empirical Comparison of Selection Measures for Decision Tree Induction", Machine Learning, 319-342, 1989.
[8] D. Saravana Kumar, "An Approach to automation Selection of Decision Tree based on Training Data Set", International Journal of Computer Applications(0975-8887) vol 64-No.21,Feb2013.
[9] L. Breiman J. H. Friedman,"Classification and Regression Trees, Wadsworth, Belmont", 1984.
[10] J. Bala, J. Hung, "Hybrid learning using Genetic Algorithms and Decision Trees for Pattern Classification",2003
[11] Carala E. Brodley Paul E. Utgoff,"Multivariate verses univariet Decision trees", Coiuns Technical Report92-8 Jan 1992.
[12] Quinlan J. R., "Induction of decision tree", Machine Learning, vol. 1, pp. 81–106, 1986 Kluwer Academic Publishers, Boston - Manufactured in The Netherlands,1986 .
[13] Quinlan J.R,"Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, Vol 4, pp 77-90., 1996.
[14] Breiman L., Friedman J. H."Classification and Regression Trees ", Wadsworth, Belmont, CA Chapman & Hall, New York. 1984.
[15] Ester M., Kriegel H.P., Sander S., and Xu X., A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. Fayyad, editors, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 226-231, Menlo Park, CA, 1996. AAAI, AAAI Press.
[16] Laure Berti-Equille L Bertossi - Handbook of Data Quality, 2013 – Springer
[17] X. Wu, V. Kumar et. al. "Top 10 algorithms in data mining", Knowledge Information System, 2008.
[18] MacQueen, J. B., "Some Methods For Classification And Analysis Of Multivariate Observations". Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, University of California Press, pp281- 297, 1967.
[19] Fonseca,M.J. and Jorge, J.A. (2003a) `Indexing high-dimensional data for content-based retrieval in large databases',Proceedings of the 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003), Kyoto, Japan: IEEE Computer Society Press, pp.267±274.
[20] Quinlan, J.R.: C4.5: Programs for Machine Learning, San Mateo, Morgan Kaufmann, 1993.
[21] Quinlan, J.R.: Improved Use of Continuous Attributes in C4.5, Journal of AI Research 4, Morgan Kaufmann 1996, pp. 77-90.
[22] Chang-Hwan Lee et. al., "Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure", 11th IEEE International Conference on Data Mining, 1550-4786/11$26.00,2011, IEEE.
[23] Tsunenori Ishioka et. al., "An Expansion Of Means For Automatically Determining The Optimal Number Of Clusters Progressive Iterations of Means And Merging Of The Clusters ", Proceedings of fourth IASTED international Conference Computational Intelligence, July 4-6, 2005, Calgary, Alberta, Canada.
[24] Osama Abu Abbas et. al, " Comparison Between Data Clustering Algorithm", The international arab journal of information technology, Vol. 5, no.3,July 2008.