www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 6 June 2017, Page No. 21557-21562
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.02

The Comparison of Various Decision Tree Algorithms for Data Analysis

Kiran Singh*, Raunak Sulekh**

* Assistant professor,Dronacharya group of Institution,Gr.Noida

** Assistant professor,GNIOT,Gr.Noida

*kiran13.singh@gmail.com , **meet.raunak123@gmail.com

ABSTRACT
The Main objective of this paper is to compare the classification algorithms for decision trees for data analysis. Classification problem is important task in data mining. Because today's databases are rich with hidden information that can be used for making intelligent business decisions. To comprehend that information, classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Several classification techniques have been proposed over the years e.g., neural networks, genetic algorithms, Naive Bayesian approach, decision trees, nearest-neighbor method etc. In this paper, our attention is restricted to decision tree technique after considering all its advantages compared to other techniques.
There exist a large number of algorithms for inducing decision trees like CHAID, FACT, C4.5, CART etc. But in this paper, these five decision tree classification algorithms are considered – ID3, SLIQ, SPRINT, PUBLIC and RAINFOREST.

Keywords *Decision Tress, ID3, SLIQ*

1. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining known by different names as – knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, data dredging, knowledge discovery in databases (KDD).

Classification is an important problem in data mining. Given a database D= {t1,t2,…, tn} and a set of classes C= {C1,…, Cm}, the ***Classification Problem*** is to define a mapping f: D C where each ti is assigned to one class. It means that given a database of records, each with a class label, a classifier generates a concise and meaningful description for each class that can be used to classify subsequent records. Actually classifier divides the database into equivalence classes that is each class contains same type of records.

In other words, classification is the process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of companies into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "Good" and "Bad."

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. For example, you may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation, vulnerable to switching over to a competing long distance phone service, or a good candidate for a surgical procedure.

2. RELATED WORK

Many works related in this area have been going on. "In A New Approach for Evaluation of Data
Mining Techniques" by Moawia Elfaki Yahia[1],the authors tried to put a new direction for the evaluation of some techniques for solving data mining tasks such as: Statistics, Visualization, Clustering, Decision Trees, Association Rules and Neural Networks. The article on " A study on effective mining of association rules from huge data base " by V.Umarani, [2] It aims at finding interesting patterns among the databases. This paper also provides an overview of techniques that are used to improvise the efficiency of Association Rule Mining (ARM) from huge databases. In another
article " K-means v/s K-medoids: A Comparative Study" Shalini S Singh explained that portioned based clustering methods are suitable for spherical shaped clusters in medium sized datasets and also proved that K-means are not sensitive to noisy or outliers.[3]. In an article "Predicting School Failure Using Data Mining C". MÁRQUEZ-VERA explained the prediction methods and the application of classification rule in decision tree for predicting the school failures.[4].There are many research works carrying out related with data mining technology in prediction such as financial stock
market forecast, rainfall forecasting, application of data mining technique in health care, base oils
biodegradability predicting with data mining technique etc,[5].

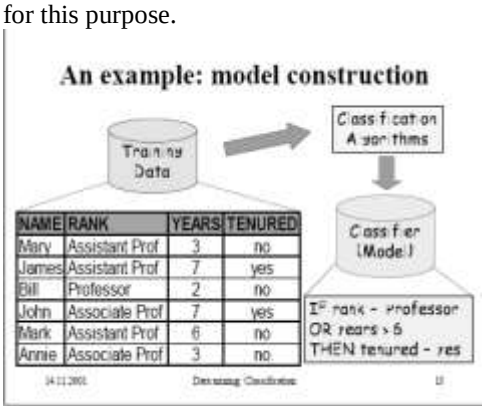
3. CLASSIFICATION- A TWO-STEP PROCESS

DOI: 10.18535/ijecs/v6i6.03

Data classification is a two-step process. In the first step, a model is build describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples, or objects. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Typically, the learned model is represented in the form of classification rules, decision trees, or mathematical formulae. The rules can be used to categorize future data samples, as well as provide a better understanding of the database contents.

In the second step, the model is used for classification. First, the predictive accuracy of the model (or classifier) is estimated. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned model's class prediction for that sample. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known. Classification and prediction have numerous applications including credit approval, medical diagnosis, performance prediction, and selective marketing.

For example, suppose that new customers are added to the database and that you would like to notify these customers of an upcoming computer sale. To send out promotional literature to every new customer in the database can be quite costly. A more cost-effective method would be to target only those new customers who are likely to purchase a new computer. A classification model can be constructed and used for this purpose.



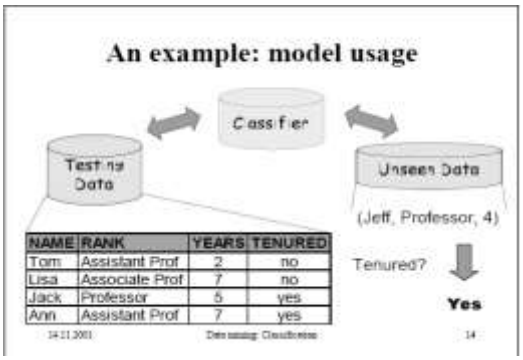


Fig 1: The data classification process

4. THE ROLE OF DATA MINING IN HIGH DIMENSIONAL ANALYSISII.

Due to the advancement in algorithm and changing scenario, new techniques have emerged in data analysis, which are used to predict and generate data patterns and to classify entities having multivariate attributes. These techniques are used to identify the pre-existing relationship in the data that are not readily available. Predictive Data mining deals with impact patterns of data.[6]

4.1 Models used in predictive Data Mining

The models mainly used in predictive data mining includes Regression, Time series, neural networks, statistical mining tools, pattern matching, association rules, clustering, classification trees etc[5] Regression model is used to express relationship between dependent and independent variables using an expression. It is used when the relationship is linear in nature. If there is a non linear relationship, then it cannot be expressed using any expression, but the relationship can be built using neural networks. In time series models, historic data is used to generate trends for the future. Statistical mining models are used to determine the statistical validity of test parameters and can be utilized to test hypothesis undertake correlation studies and transform and prepare data for further analysis.

Pattern matching are used to find hidden characteristics within data and the methods used to find patterns with the data includes association rules. [5] Association rules allows the analysts to identify the behavior pattern with respect to a particular event where as frequent items are used to find how a group are segmented for a specific set. Clustering is used to find the similarity between entities having multiple attributes and grouping similar entities and classification rules are used to categorize data using multiple attributes. So in this paper, I am going to analyze decision tree algorithms.

4. 2 Decision trees

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as

Follows:If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets *S*1, *S*2, . . . according to the outcome for each case, and apply the same procedure recursively to each

subset. Use either information gain or gain ratio to rank the possible tests.

Advantages	Limitations
Error rate is less	Decision trees typically require certain knowledge of quantitative or statistical experience to complete the process accurately. Failing to accurately understand decision trees can lead to a garbled outcome of business opportunities or decision possibilities.
Decomposition is easier as compared with other techniques	It can also be difficult to include variables on the decision tree, exclude duplicate information or express information in a logical, consistent manner. The inability to complete the decision tree using only one set of information can be somewhat difficult.
Represent the knowledge in the form of IF-THEN rules. Rules are easier for humans to understand.	While incomplete information can create difficulties in the decision-tree process, too much information can also be an issue.

Table 1: Decision Tree-Advantages and Disadvantages

5. DECISION TREE INDUCTION ALGORITHMS

5.1 ID3

J. Ross Quinlan originally developed ID3, which is a decision tree induction algorithm, at the University of Sydney. ID3 stands for "Iterative Dichotomizer (version) 3" and its later versions include C4.5 and C5. It is an early technique that influenced a large part of the research on decision trees is useful to look at in order to understand basic decision tree construction.

ID3 is based on the Concept Learning System (CLS) algorithm that is in fact a recursive top-down divide-and-conquer algorithm. The ID3 family of decision tree induction algorithms uses information theory to decide which attribute shared by a collection of instances to split the data on next. Attributes are chosen repeatedly in this way until a complete decision tree that classifies every input is obtained. If the data is noisy, some of the original instances may be misclassified. It may be possible to prune the decision tree in order to reduce classification errors in the presence of noisy data. The speed of this learning algorithm is reasonably high, as is the speed of the resulting decision tree classification system.

5.2 SLIQ

SLIQ is a decision tree classifier that can handle both numeric and categorical attributes. SLIQ uses a pre-sorting technique in the tree growth phase and this sorting procedure is integrated with a breadth-first tree growing strategy to enable SLIQ to classify disk-resident datasets. For the tree-pruning phase, SLIQ uses an algorithm, which is based on the Minimum Description Length Principle. This algorithm is inexpensive, and results in compact and accurate trees.

5.3 Sprint

SPRINT, is a new decision-tree based classification algorithm that removes all of the memory restrictions, and is fast and scalable. The algorithm has also been designed to be easily parallelized, allowing many processors to work together to build a single consistent model. SPRINT has excellent scale up, speedup and size up properties. The combination of these characteristics makes SPRINT an ideal tool for data mining. SPRINT has no restriction on the size of input and yet is a fast algorithm. It shares with SLIQ the advantage of a one-time sort, but uses different data structures. In particular, there is no structure like the class list that grows with the size of input and needs to be memory-resident.

5.4 Public

PUBLIC stands for Pruning and Building Integrated in Classification. It is an improved decision tree classifier that integrates the second "pruning" phase with the initial "building" phase. Generating the decision tree in two distinct phases could result in a substantial amount of wasted effort since an entire sub tree constructed in the first phase may later be pruned in the next phase. It is observed that pruning phase prunes large portions of the original tree-in some cases; this can be as high as 90% of the nodes in the tree. So, during the building phase, before splitting a node, if it can be concluded that the node will be pruned from the tree during the subsequent pruning phase, then we could avoid building the sub tree rooted at the node. Consequently, since building a sub tree usually requires repeated scans to be performed over the data, significant reductions in I/O and improvements in performance can be realized.

PUBLIC, a decision tree classifier that during the growing phase, first determines if a node will be pruned during the following pruning phase, and subsequently stops expanding such nodes. In order to make this determination for a node, before it is expanded, PUBLIC computes a lower bound on the minimum cost sub tree rooted at the node. This estimate is then used by PUBLIC to identify the nodes that are certain to be pruned, and for such nodes, not expend effort on splitting them. Thus, PUBLIC integrates the pruning phase into the building phase instead of performing them one after the other.

5.5 Rainforest

Rainforest is a unifying framework for decision tree classifiers that separates the scalability aspects of algorithms for constructing a decision tree from the central features that determine the quality of the tree. Also, this general algorithm is easy to instantiated with other specific algorithms like C4.5, CART, CHAID, FACT, ID3 and extensions, SLIQ, Sprint and QUEST.

This general framework, called RainForest, closes the gap between the limitations to main memory datasets of algorithms in the machine learning and statistics literature and the scalability requirements of a data mining environment.

6.RESULTS: PERFORMANCE EVALUATION AND SCALABILITY

The primary metric for evaluating classifier performance is **classication accuracy** i.e. the percentage of test samples that are correctly classified. The secondary matrices may be:

1. Classification time

2. Size of Decision Tree

The ideal goal for a classifier is to procude compact, accurate trees in a short time. The performance evaluation of the various decision tree classification algorithms are divided into two parts. First part compares the algorithm with other classification algorithms and the second part of performance evaluation examines classification accuracy and classification time of the particular classifier on some dataset. Then we examine the **scalibility** of the various classifiers along two dimensions:

1. Number of training examples, and

2. Number of training attributes in the data.

Since real-life data sets are generally small, synthetic data sets are used to study the performance of the classifiers on larger data sets. The purpose of the synthetic data sets is primarily to examine the classifier's sensitivity to parameters such as noise, number of classesand number of attributes. The performance evaluation of various classifiers is given below.

SPRINT and SLIQ

The data structures and how a tree is grown are very different in SPRINT and SLIQ. They consider the same types of splits at every node and use identical splitting index(gini index). Therefore they produce identical trees for a given dataset (provided SLIQ can handle the data set). The final trees obtained using the two algorithms are also identical since SPRINT uses the same pruning method as SLIQ uses. An often used benchmark in classification is STATLOG, however its largest dataset contains only 57000 training examples. Each record in this synthetic database consisits of nine attributes frou of which are shown in table.

Attribute	*Value*
Salary	Uniformaly distributed from 20k to 150k
Commission	Salary>=75k ⇒ commission=0 else uniformaly distributed from 10k to 75k Uniformaly ditributed from 20 to 80 uniformaly distributed from 0 to 500k.
Age	Uniformly distributed from 20 to 80
Loan	Uniformly distributed from 0 to 5000000

Table2 : Description of Attributes for Synthetic data

Some classification functions were also proposed to produce databases with distributions with varying complexities. Function 1 results in fairly small decision trees, while Function2 results in very large trees. Both these functions divide the database into two groups: Group A and Group B. Figure below shows the predicates for Group A for each function.

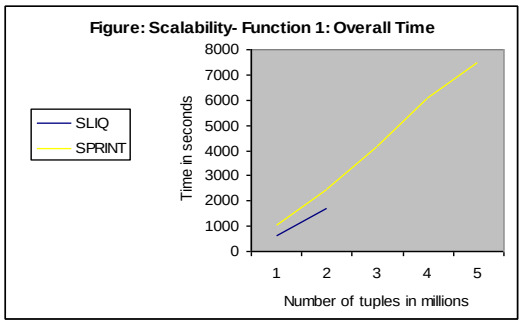
Function 1 – Group A:

((age<40) □ (50k<=salary<=100k)) □ ((40<=age<60) □ (75k<=salary>=125k)) □ ((age>=60) □ (25k<=salary<=75k))

Function 2 – Group A:

Disposable > 0 Where disposable = (0.67 * (salary + commission)) – (0.2 * loan – 20k)

For the performance analysis , we compare the response times fo SPRINT and SLIQ on training sets of various sizes. SPRINT is compared with SLIQ only becuase in most cases SLIQ outperforms other pouplar decision tree classifiers. For the desk-resident datasets, SLIQ is the only other viable algorithm. The training sets taken for the performance evaluation taken are ranging in size from 10000 records to 2.5 million records. This range is selected to examine how well SPRINT performs in operating regions where SLIQ can and can not run. The results are shown in graph(given below) on databases generated using Function 1.

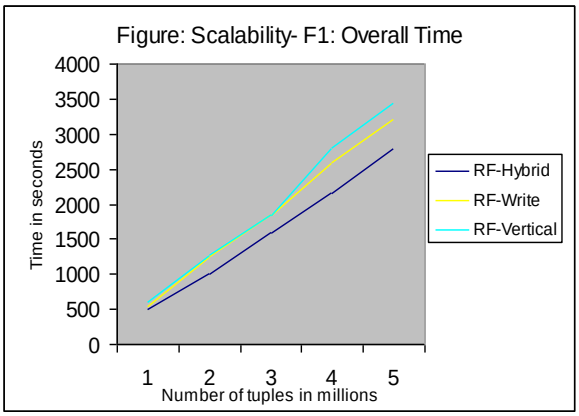


The graph shows that SPRINT is somewhat slower than SLIQ for the data sizes for which thte class list could fit in memory. However, as soon as we cross an input size threshold (about 1.5 million records, however it depends on the system configuration also), SLIQ starts thrashing, whereas SPRINT continues to exhibit a nearly linear scaleup. This shows that SPRINT removes all momory restrictions that limit existing decision-tree classification algorithms, and yet exhibits the same excellent behaviour as SLIQ.

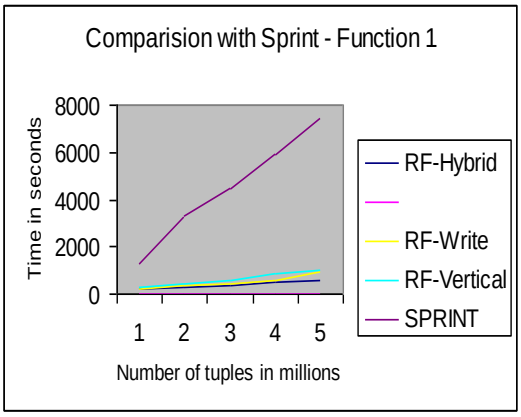
Rainforest

Rainforest is a unifying framework of family of algorithms for classifying the data using decision trees. The algorithms in the Rainforest framework are: RF-Write, RF-Read, RF-Hybrid and RF-Vertical. The performance evaluation of these algorithms is examined as the size of input database increases.For algorithms RF-Write and RF-Hybrid, the size of AVC-group buffer is fixed to 2.5 million entries; for algorithms RF_Vertical the AVC-group buffer size is fixed to 1.8 million entries. Figure (given below) shows the overall running time of algorithms as the number of tuples in the input database increases from 1000000 to 5000000.

DOI: 10.18535/ijecs/v6i6.03



The running time for all algorithms grows linearly with the number of tuples. Algorithm RF-Hybrid outperfoms both algorithms RF-Write andRF-Vertical in terms of running time. For small AVC-group sizes(40% and below), the times for RF-Vertical and RF-Hybrid are identical. The larger buffer-size only shows its effect for larger AVC-group sizes. The running time of RF-Write is not affected through a change in AVC-group size, since RF-Write writes partitions regardless of the amount of memory available. Figure shows the performance comparision of RAINFOREST with SPRINT algorithms for the function 1 whose predicate is given in SPRINT.



The figure shows that for function 1, RF-Hybrid and RF-Vertical outperform SPRINT by a factor of about 5.

PUBLIC

The integrated Public algorithms are implemented using the same code base as SPRINT except that they perform pruning while the tree is being built.

For performance evaluation, PUBLIC and SPRINT, both algoritms are tested on real-life datasets as well as synthetic datasets. The attributes of the synthetic dataset are shown in table 1 and the execution times for the datasets generated by functions 1 and 2 (whose predicates are given in SPRINT) for the both algorithms are shown in table 2.

Attribute	Description	Value
Salary	Salary	Uniformly distributed from 20000 to 150000
Commission	Commission	If salary>=75000 then commission=0 else uniformly distributed from 10000 to 75000
Age	Age	Uniformly distributed from 20 to 80
Elevel	Education level	Uniformly choosen from 0 to 4
Hvlaue	Home value	Uniformly distributed from 0.9k100000 to 1.5k10000 where k {0,...9} depends on zipcodes.
Years	Years house owned	Uniformly distributed from 1 to 30
Loan	Total loan amount	Uniformly choosen from 0 to 5000000
Car	Make of the car	Uniformly choosen from 1 to 20
Zipcode	Zipcode of the town	Uniformly choosen from available zipcodes

Table 3: Description of attributes is synthetic data sets

Function	1	2
SPRINT	1471	1277
PUBLIC(1)	720	615
PUBLIC(S)	604	559
PUBLIC(V)	593	510

Table 4: Synthetic data sets : Execution time (secs)

For each dataset, the noise factor was set to 10%. From the tables, we can easily see that PUBLIC outperforms SPRINT by a significant amount. Execution time increases as the noise is increased. This is because as the noise is increased, the size of the tree and thus the number of nodes generated increases. The execution time for SPRINT increases at a faster rate than those for PUBLIC, as noise factore os increased. Thus, PUBLIC results in better performance improvements at higher noise values.

7. CONCLUSION

Classification of large databases is an important data mining problem. Many classification algorithms have been propose, but so far there is no algorithms which uniformaly outperforms all other algorithms in terms of quality. SLIQ achieves comparable classification accuracy but produces small decision trees and has small execution times. However as soon as we cross an input size threshold, SLIQ starts thrashing, whereas another decision tree classifier which is the fastest classifiers, SPRINT contiues to exhibit a nearly linear scaleup. SPRINT removes all the momory restrictions that limit the decision tree classifiers, so the other algorithms, PUBLIC and RAINFOREST are compared to SPRINT for the performance issues as they are as good as SPRINT and under particular circumstances they perform better. Public integrates the pruning phase into the tree building phase, as a result fewer nodes are expanded during the building phase and the

amount of work (example I/O) to construct the tree is reduced. Rainforest which is a family of various algorithms uses AVC-group for the splitting criteria at a tree node, which is relatively compact. Depending on the available memory, these algorithms offer significant performance improvements over SPRINT. If there is enough memory to hold all AVC-groups for a node, the speed up is even better.

8. FUTURE ENHANCEMENT

We will be able to create a new hybrid algorithm by comparing the advantages and disadvantages of the existing ones. We can also take other techniques which are not included in this survey for comparison purpose and can find the best one by evaluating the advantages and limitations of the existing one.

REFERENCES

[1]. "A New Approach for Evaluation of Data Mining Techniques", Moawia Elfaki Yahia1, Murtada Elmukashfi El-taher2, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

[2]. "A study on effective mining of association rules from huge database" V.Umarani et. al. / IJCSR International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.

[3]. "K-means v/s K-medoids: A Comparative Study" Shalini S Singh, National Conference on Recent Trends in Engineering & Technology, May 2011.

[4]. Predicting School Failure Using Data Mining"C. MÁRQUEZ-VERA

[5]. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" K.Srinivas et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-

255.

[6]. en.wikipedia.org/wiki/Data_mining