International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Comparative Analysis of Decision Tree Classification Algorithms

Anuja Priyam[a*], Abhijeet[a], Rahul Gupta[a], Anju Rathee[b], and Saurabh Srivastava[b]

[a]Computer science & Engineering, Kanpur institute of technology, Kanpur
[b]Lovely Professional university, Jalandhar

Accepted 24 March 2013, Available online 1June 2013, **Vol.3, No.2 (June 2013)**

Abstract

At the present time, the amount of data stored in educational database is increasing swiftly. These databases contain hidden information for improvement of student's performance. Classification of data objects is a data mining and knowledge management technique used in grouping similar data objects together. There are many classification algorithms available in literature but decision tree is the most commonly used because of its ease of execution and easier to understand compared to other classification algorithms. The ID3, C4.5 and CART decision tree algorithms former applied on the data of students to predict their performance. But all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory. This limits their suitability for mining over large databases. This problem is solved by SPRINT and SLIQ decision tree algorithm. In serial implementation of SPRINT and SLIQ, the training data set is recursively partitioned using breadth-first technique. In this paper, all the algorithms are explained one by one. Performance and results are compared of all algorithms and evaluation is done by already existing datasets. All the algorithms have a satisfactory performance but accuracy is more witnessed in case of SPRINT algorithm.

Keywords: *Data Mining, Educational Data Mining, Classification Algorithm, Decision trees, ID3, C4.5, CART, SLIQ, SPRINT*

1. Introduction

Education is a crucial element for the betterment and progress of a country. It makes the people of a country enlightened and well affected. Mining in educational environment is called educational data mining(M. Sukanya et al,2012). Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining and data mining etc. As we know, large amount of data is stored in educational database; data mining is the process of discovering interesting knowledge from these large amounts of data stored in database, data warehouse or other information repositories.

A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Classification is one of the most frequently studied problems by data mining and machine learning researchers(Brijesh et al,2011).

It consists of predicting the value of a categorical attribute based on the value of other attributes. Classification methods like decision trees, rule mining, Bayesian network etc. can be applied on the educational data for predicting the students behavior, performance in examination etc.

Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. It the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms (Surjeet kumar et al,2012). The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year.

Decision tree can be constructed relatively fast compared to other methods of classification. Trees can be easily converted into SQL statements that can be used to access databases efficiently. Decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm. The C4.5, ID3, CART decision tree algorithms

*Corresponding author: Anuja Priyam

are applied on the data of students to predict their performance. These algorithms are explained below:-

C4.5 Algorithm

It is an improvement of ID3 algorithm developed by Quilan Ross in 1993. It is based on Hunt's algorithm and also like ID3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. It accepts both continuous and categorical attributes in building the decision tree(Anju Rathee)

It has an enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute.

The algorithm C4.5 has following advantages:

- Handling attributes with different costs.
- Handling training data with missing attribute values- C4.5 allows attribute values to be marked as '?' for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling both continuous and discrete attributes- in order to handle continuous attributes,C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Pruning trees after creation- C4.5 goes back through the tree once it has been created and attempts to remove branches that do not help by replacing them with leaf nodes(Devi Prasad et al,2010).

ID3 Algorithm

Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. It is serially implemented and based on Hunt's algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node (Tarun Verma et al).

In order to select the attribute that is most useful for classifying a given sets, we introduce a metric – information gain. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked. Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise and it is serially implemented. Thus an intensive pre-processing of data is carried out before building a decision tree model with ID3.

SPRINT Algorithm

It stands for scalable parallelizable induction of decision tree algorithm. It was introduced by Shafer et al in 1996. It is fast, scalable decision tree classifier. It is not based on Hunt's algorithm in constructing the decision tree, rather it partitions the training data set recursively using breadth-first greedy technique until each partition belong to the same leaf node or class. It can be implemented in both serial and parallel pattern for good data placement and load balancing.

It uses two data structure: attribute list and histogram which is not memory resident making sprint suitable for large data sets, thus it removes all the data memory restrictions on data. It handles both continuous and categorical attributes (Sunita et ,2011).

CART Algorithm

It stands for classification and regression trees and was introduced by Breiman in 1984.It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt's algorithm and can be implemented serially. It uses gini index splitting measure in selecting the splitting attribute.

CART is unique from other Hunt's based algorithm as it is also use for regression analysis with the help of the regression trees (S.Anupama et al,2011). The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time.

It uses many single-variable splitting criteria like gini index, symgini etc and one multi-variable in determining the best split point and data is stored at every node to determine the best splitting point. The linear combination splitting criteria is used during regression analysis. SALFORD SYSTEMS implemented a version of CART called CART using the original code of Breiman(1984). CART has enhanced features and capabilities that address the short-comings of CART giving rise to a modern decision tree classifier with high classification and prediction accuracy.

SLIQ Algorithm

It stands for supervised learning in ques. It was introduced by Mehta et al (1996). It is fast scalable decision tree algorithm that can be implemented in serial and parallel pattern. It is not based on HUNT'S Algorithm for decision tree classification. It partitions a training data set recursively using breadth-first greedy strategy that is integrated with pre-sorting technique during the tree building phase. In building a decision tree model SLIQ handles both numeric and categorical attributes (Tarun Verma et al).

One of the disadvantages of SLIQ is that it uses a class list data structure that is memory resident thereby imposing memory restrictions on the data. It uses minimum description length principle(MDL) in pruning the tree after constructing it MDL is an expensive technique in tree pruning that uses the least amount of coding in producing tree that are small in size using bottom-up technique[12].

Table 1 Frequency usage of decision tree algorithms
Algorithm Usage frequency (%)

CLS	9
ID3	68
IDE3+	4.5
C4.5	54.55
C5.0	9
CART	40.9
Random Tree	4.5
Random Forest	9
SLIQ	27.27
Public '	13.6
OCI	4.5
Clouds	4.5

Comparison

The following table 1 shows the comparison between the working of existing algorithms. These algorithms are among the most influential data mining algorithms in the research community [4].

Table 2 Parameter Comparison of Decision tree algorithm

ALGORITHMS	ID3	CART	C4.5	SLIQ	SPRINT
Measure	Entropy info-gain	Gini diversity index	Entropy info-gain	Gini index	Gini index
Procedure	Top-down decision tree constructi on	Construct s binary decision tree	Top-down decision tree constructi on	Decision tree constructi on in a breadth first manner	Decision tree constructi on in a breadth first manner
Pruning	Pre-pruning using a single pass algorithm	Post pruning based on cost-complexi ty measure	Pre-pruning using a single pass algorithm	Post-pruning based on MDL principle	Post-pruning based on MDL principle

Table 3a Classifiers Accuracy

Algorithm	Correctly classified Instances	Incorrectly Classified Instances
ID3	52.0833%	35.4167%
C4.5	45.8333%	54.1667%
CART	56.2500%	43.7500%

Table 2 shows the accuracy of ID3, C4.5 and CART algorithms for classification applied on some data sets using 10-fold cross validation is observed. It shows that a C4.5 technique has highest accuracy of 67.7778% compared to other methods. ID3 and CART algorithms also showed an acceptable level of accuracy. The table also shows the time complexity in seconds of various classifiers to build the model for training data.

Table 3b Classifiers Accuracy

Algorithm	Class	TP Rate	FP Rate
ID3	Pass	0.714	0.184
	Promoted	0.625	0.232
	Fail	0.786	0.061
C4.5	Pass	0.745	0.209
	Promoted	0.517	0.213
	Fail	0.786	0.092
CART	pass	0.809	0.349
	Promoted	0.31	0.18
	Fail	0.643	0.105

Table 3 above shows the three machine learning algorithms that produce predictive models with the best class wise accuracy. From the classification accuracy it is clear that the true positive rate of the model for the FAIL class is 0.786 for ID3 and C4.5 decision trees that means model is successfully identifying the students who are likely to fail. These students can be considered for proper counseling so as to improve their result.

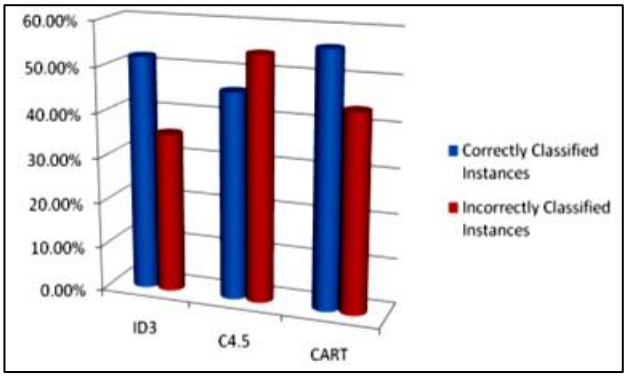

Fig.1 The classifiers accuracy on various data sets is represented in the form of a graph.

Conclusion

In this paper, three existing decision tree algorithms (ID3, C4.5, and CART) have been applied on the educational data for predicting the student's performance in examination. All the algorithms are applied on student's internal assessment data to predict their performance in the final exam. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the tutor to identify the weak students and improve their performance. C4.5 is the best algorithm for small datasets among all the three because it provides better accuracy and efficiency than the other algorithms. The main disadvantages of serial decision tree algorithm (ID3, C4.5 and CART) are low classification accuracy when the training data is large. But all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory. This limits their suitability for mining over large databases. This problem is solved by SPRINT and SLIQ decision tree algorithm. Still effective algorithms for decision tree should be developed.

References

Anju Rathee "survey on decision tree classification algorithms for the evaluation of the student performance*" ijct* Vol. 4 no. 2

Surjeet kumar yadav and Saurabh Pal(2012)"Data mining: a prediction for performance improvement of engineering students using classification", *World of science and information technology journal (WCSIT)* ISSN: 2221-0741, Vol 2, no. 2

S.Anupama Kumar and Dr. Vijayalakshmi M.N. (2011) "Efficiency of decision trees in predicting student's academic performance", D.C. Wyld, et al. (Eds): *CCSEA* 2011, CS & IT 02, pp. 335-343, 2011.

Matthew N. Anyanwu and Sajjan G.shiva "Comparative analysis of serial decision tree classification algorithms", *International journal of computer science and security, (IJCSS)* Volume 3: Issue (3).

Devi Prasad bhukya and S. Ramachandram (Aug 2010)" Decision tree induction- An Approach for data classification using AVL –Tree", *International journal of computer and electrical engineering,* Vol. 2, no. 4

Tarun Verma, Sweety raj,Mohammad Asif khan, Palak modi (2012) "Literacy Rate Analysis", *International journal of science & engineering research* volume 3, issue 7, ISSN 2229-5518.

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.

M. Sukanya, S. Biruntha, Dr. S. Karthik and T.Kalaikumaran "Data mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm", International conference on computing and control engineering (*ICCCE 2012*) 12 & 13 April, 2012.

Shaeela Ayesha, Tasleem Mustafa, M.Inayat Khan and Ahsan Raza Sattar(2010) "Data mining model for higher education system", *Europeon journal of scientific research,* ISSN 1450-216X Vol. 43 no. pp.24-29.EuroJournalsPublishing,inc.http://www.eurojournals.com/ejsr.htm

C.Romero and S.Ventra "Educational data mining: A survey from 1995 to 2005", 2006 *Elsevier ltd*. All rights reserved. www.elsevier.com/locate/eswa

Sunita B aher, Mr. LOBO L.M.R.J (2011) "Data mining in educational system using weka tool", *International conference on emerging technology trends (ICETT)*

John shafer, Rakesh agrawal, Manish Mehta "SPRINT: A scalable parallel classifier for data mining" *IBM Almaden Research Center,* 650 Harry road, San Jose, CA 95120.

Jorma Rissanen, Rakesh agrawal, Manish Mehta "SLIQ: A scalable parallel classifier for data mining" *IBM Almaden Research Center,* 650 Harry road, San Jose, CA 95120.

Brijsh Kumar bhardwaj and Saurabh Pal (2011)"Data mining: a prediction for performance improvement using classification", *International journal of computer science and information security*, vol. 9, no. 4.