INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 28-30

COMPARATIVE EXAMINATION OF DECISION TREE CLASSIFICATION ALGORITHMS

M Nancy[1]

[1]Fatima College, India.

ABSTRACT

The volume of data in educational databases is growing rapidly, holding hidden insights for improving student performance. Data classification, a key technique in data mining and knowledge management, groups similar data objects together. Among classification algorithms, decision trees are popular due to their simplicity. However, traditional algorithms like ID3, C4.5, and CART are limited to small datasets stored entirely in memory. This issue is overcome by SPRINT and SLIQ algorithms, which efficiently handle large databases. In our study, we compare these algorithms' performance using existing datasets, with SPRINT showing the highest accuracy.

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

1. INTRODUCTION

Education plays a pivotal role in advancing and improving a nation. It fosters enlightenment and positive development among its populace. Within the educational realm, the process of extracting insights is termed educational data mining (M. Sukanya et al., 2012). This emerging field intersects with established areas of research such as e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, and data mining. Given the vast amount of data stored in educational databases, data mining becomes instrumental in uncovering valuable knowledge from these repositories.

Various data mining techniques have been applied to educational data to enhance student performance, including regression, genetic algorithms, Bayesian classification, k-means clustering, association rules, and prediction. These techniques aid in understanding the learning process by identifying, extracting, and evaluating relevant variables. Among the most studied problems in data mining and machine learning is classification, which involves predicting categorical attributes based on other attributes.

Classification methods such as decision trees, rule mining, and Bayesian networks are commonly utilized in educational data analysis to predict student behavior and examination performance. Decision trees, characterized by a flow-chart-like structure, are particularly popular due to their ease of implementation and comprehension. They efficiently predict outcomes such as the number of students likely to pass, fail, or progress to the next academic year.

Decision tree construction is relatively fast compared to other classification methods. Moreover, trees can be easily converted into SQL statements for efficient database access.

Decision tree classifiers often achieve comparable or superior accuracy compared to alternative methods. The implementation of decision tree algorithms can be tailored to suit different scenarios, whether in a serial or parallel fashion, depending on factors such as data volume, available memory, and algorithm scalability.

C4.5 ALGORITHM

The C4.5 algorithm, an advancement of the ID3 algorithm introduced by Quinlan Ross in 1993, builds upon Hunt's algorithm and shares its serial implementation approach. One of its key features is pruning, where internal nodes are replaced with leaf nodes to decrease error rates. Unlike ID3, C4.5 accommodates both continuous and categorical attributes during decision tree construction (Anju Rathee).

C4.5 employs an improved tree pruning technique to mitigate misclassification errors caused by noise and excessive detail in the training dataset. Similar to ID3, data sorting occurs at each tree node to identify the optimal splitting attribute. It utilizes the gain ratio impurity method to assess the splitting attribute.

SPRINT ALGORITHM

The SPRINT algorithm, abbreviated for Scalable Parallelizable Induction of Decision Tree, was introduced by Shafer et al. in 1996. It represents a rapid and scalable decision tree classifier. Unlike traditional methods based on Hunt's algorithm, SPRINT adopts a recursive partitioning approach using breadth-first greedy technique on the training dataset until each partition belongs to the same leaf node or class.

This algorithm can be implemented in both serial and parallel patterns to ensure optimal data placement and load balancing.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 28-30

SPRINT utilizes two key data structures: an attribute list and a histogram, which are not memory resident. This feature renders SPRINT well-suited for handling large datasets, eliminating memory constraints on data. Moreover, it effectively manages both continuous and categorical attributes.

CART ALGORITHM

The CART algorithm, which stands for Classification and Regression Trees, was introduced by Breiman in 1984. Unlike traditional decision tree algorithms, CART constructs both classification and regression trees. Its classification tree construction relies on binary splitting of attributes, following Hunt's algorithm, and can be implemented serially. CART employs the Gini index splitting measure to select the splitting attribute.

What sets CART apart from other Hunt's algorithm-based approaches is its ability to perform regression analysis through regression trees (S. Anupama et al., 2011). This feature enables forecasting of a dependent variable based on a set of predictor variables over a specified timeframe.

In CART, various single-variable splitting criteria such as the Gini index and symgini are utilized, alongside a multi-variable criterion, to determine the optimal split point. Data is stored at each node to facilitate this process. During regression analysis, a linear combination splitting criterion is employed.

A version of CART, developed by Salford Systems, implements the original code by Breiman (1984). This enhanced version of CART addresses its shortcomings, resulting in a modern decision tree classifier with improved classification and prediction accuracy.

ID3 ALGORITHM

The Iterative Dichotomiser 3 (ID3) is a straightforward decision tree learning algorithm devised by Quinlan Ross in 1986. It operates in a serial manner and is grounded in Hunt's algorithm. The fundamental concept behind the ID3 algorithm involves constructing a decision tree through a top-down, greedy search across provided datasets, testing each attribute at every tree node (Tarun Verma et al.).

To determine the most useful attribute for classifying a given dataset, the ID3 algorithm introduces a metric known as information gain. The aim is to minimize the number of questions asked in order to find an optimal classification approach. Information gain serves as a function to measure which questions yield the most balanced splits. ID3 employs the information gain metric to select the splitting attribute, exclusively accepting categorical attributes for constructing the tree model.

However, ID3 may yield less accurate results in the presence of noise, necessitating intensive pre-processing of data prior to building the decision tree model.

SLIQ ALGORITHM

The SLIQ algorithm, which stands for Supervised Learning In Quest, was introduced by Mehta et al. in 1996. It represents a fast and scalable decision tree algorithm that can be implemented in both serial and parallel patterns. Unlike traditional methods relying on Hunt's Algorithm, SLIQ adopts a recursive partitioning strategy using a breadth-first greedy approach, which is integrated with a pre-sorting technique during the tree construction phase. SLIQ effectively handles both numeric and categorical attributes when building a decision tree model (Tarun Verma et al.).

One drawback of SLIQ is its utilization of a class list data structure that resides in memory, thus imposing memory constraints on the data.

Additionally, SLIQ employs the Minimum Description Length (MDL) principle for tree pruning after construction. MDL is an expensive technique for pruning trees, aiming to produce compact trees with minimal coding using a bottom-up approach.

Table 1: Classifiers Accuracy

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

It shows that a C4.5 technique has highest accuracy of 67.7778% compared to other methods. ID3 and CART algorithmsalsoshowed an acceptable level of accuracy. The tablealso shows the time complexity in seconds of various classifiers to build the model for training data.

Table2: Parameter Comparison of Decision tree algorithm

ALGORITHMS	ID3	CART	C4.5	SLIQ	SPRINT
Measure	Entropyinfo-gain	Ginidiversityindex	Entropyinfo-gain	Giniindex	Giniindex
Procedure	Top-downdecisiontree construction	Constructs binarydecisiontree	Top-downdecisiontree construction	Decisiontree construction in abreadthfirst manner	Decisiontree construction in abreadthfirst manner
Pruning	Pre-pruningusing asinglepass algorithm	Postpruningbasedoncost-complexity measure	Pre-pruningusing asinglepass algorithm	Post-pruningbasedonMDL principle	Post-pruningbasedonMDL principle

2. CONCLUSION

In this study, three established decision tree algorithms (ID3, C4.5, and CART) were applied to educational data to predict students' performance in examinations. These algorithms were employed on internal assessment data to forecast students' outcomes in the final exam. The efficacy of these decision tree algorithms was assessed based on their accuracy and the time taken to generate the decision tree. The predictions generated by the system have aided tutors in identifying weaker students and enhancing their performance.

Among the three algorithms, C4.5 emerged as the most suitable for small datasets, offering superior accuracy and efficiency compared to the others. However, the serial implementation of decision tree algorithms (ID3, C4.5, and CART) exhibits shortcomings in terms of classification accuracy when dealing with large training datasets. Moreover, these algorithms require that either the entire dataset or a significant portion of it remain permanently in memory, limiting their applicability to mining large databases. This limitation is addressed by the SPRINT and SLIQ decision tree algorithms. Nonetheless, there remains a need to develop more effective algorithms for decision trees.

3. REFERENCES

[1] 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

[2] Tarun Verma, Sweety raj,Mohammad Asif khan, Palak modi (2012) “Literacy Rate Analysis”,International journal of science& engineering research volume 3,issue 7,ISSN2229- 5518.

[3] 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.

[4] 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.

[5] ShaeelaAyesha, Tasleem Mustafa, M.InayatKhanandAhsan Raza Sattar(2010)“Dataminin gmodelforhig hereducation system”, Europeonjournal of scientific research, ISSN 1450- 216X Vol. 43 no. pp.24-29. Euro Journals Publishing,inc.http://www.eurojournals.com/ejsr.htm