ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

A Survey On Educational Data Mining Algorithms And Its' Accuracy

K. SELVI
Research Scholar,
Dept, of Computer Science,
Jairams Arts and Science College,
Karur 639003.

Dr. B. MUKUNTHAN
Research Advisor,
Dept, of Computer Science,
Jairams Arts and Science College,
Karur 639003.

Abstract: The aim of this research paper is to study the data mining algorithms and its accuracy in prediction. The main objective of this research work is to compare the accuracy of various existing algorithms.. By analyzing and comparing the existing algorithms, this research work tries to suggest a new framework for better prediction. Computer science undergraduate student's data are collected and analyzed to study the family background, and individual difference factors and their links with their exam performance. Based on the data collected from an arts and Science College, the student's performance is predicted with various algorithms. The required developments are suggested by introducing the cognitive load factors, personality differences and individual's prior experiences, etc. From the research it has been observed that the prior experience of the learner's has positive impact on exam performance. Those who have studied their school education in a particular type of management and practice have more advantage in their exam performance. This prediction of exam performance is due to the prior experience and their similarity in the environment in the same set-ups.

Keywords: Data mining algorithms, prediction of accuracy, new framework, cognitive load, prior experiences, exam performance, etc.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

1. Introduction

Due to the digitization of large educational data, the possibilities to extract the new knowledge have many potential applications. Even though predicting the performance based on the background of the learners is an old concept, still there are wide gaps exists to predict the outcome with maximum accuracy. This research paper addresses the issues related to various data mining algorithms and its accuracy. Data mining algorithms are broadly categorized into two types. They are supervised and unsupervised algorithms. Data mining has three main components. They are classification, association and sequence analysis. Classification is a method of categorizing the discrete data. Classification algorithms are used to find the useful hidden information from the large data. The major supervised classification algorithms are Naive Bayse, J48, Random Forest, Random Tree, Neural network algorithms, genetic algorithms, CART-Classification and Regression Tree, MLP-Multi Layer Perceptron, etc. In this research paper, section 1 describes the data mining concepts, section 2 explains the various algorithms and its accuracy, section 3 discuss the Research methodology applied, section 4 helps to discuss the finding and suggestions and the section 6 concludes the findings from the research work.

Objectives

- To study the various classification algorithms.
- To compare their accuracy in prediction.
- To suggest the ways to improve the algorithm.
- To propose a framework to improve the existing algorithm.
- To outline the various individual factors for prediction.

2. Related Works

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

Educational Data mining:

Educational data mining is an interdisciplinary approach to study the educational data using data mining techniques. Each educational problem requires a specific approach and techniques to apply the data mining algorithms. Traditional methods and algorithms are suited to solve the problems [1].

Romero et al. conducted a survey to find out the developments in educational data mining during the period 1995 to 2005. Also they studied the application of data mining techniques to traditional learning system, web learning system and web content management system. The researchers found out the needs to apply the various techniques such as clustering, association, visualization, classification, pattern mining, etc., [2]

Alex Freitas conducted research in educational data mining algorithms and its improvements. Also they found out the application of evolutionary genetic based algorithms to study the performance. They also studied extensively the application of genetic operator, fitness level and individual representations, etc. Furthermore, they studied the prediction of accuracy and interpretation of the new knowledge [3].

Mitra and Pal conducted research on soft computing techniques, data mining functions and preference criterions on data mining. Furthermore they studied the hybridization of algorithms neural networks and new models to discover the new knowledge. Throguh survey method data were tabulated based on the soft computing method, learner's preferences, and functions.[4].

Data set, algorithm and their performances were studied by the researchers Shiqug wang et al. Their research showed the requirement of algorithms and its accuracy to predict the missing data Regression analysis used to identify the hidden data. It was effective to detect the unknown

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

hidden information. But the evolutionary neural network was more effective than the regression analysis. This research work helped to identify the need to find the accurate algorithm [5].

Ajay pal conducted research to study the relationships between the job recruitment and students' performance using Navie Bayse. They found out the accuracy level by using the algorithm by prediction rate up to 86 percentages. Error rate was very low in their research. There used four basics steps involved in conversion of raw data into a new knowledge. The stages were data selection, preprocessing the data, transferring the data and finding patterns. [6]

Mohamed Abu Tahir used various classifications, association, and clustering techniques to identify the student's performance based on the background data of fifteen years duration. Naive Bayse and rule induction classifier method used by them to identify the performance to improve it further [7].

Yang Zhang et al. studied the concept of student's retention level which was the main indicator to find the exam performance. The student's retention level was calculated by the application of data mining techniques. By identifying the retention problem they tried to for a new strategy and intervention program to improve the overall performance [8].

Sourabi pal tried to identify the dropout rates using data mining techniques among the engineering students By identifying the potential engineering students, the researchers tried to conduct intervene program to reduce the dropout rate. The prediction was done by using data mining techniques from the incoming new students. They found out the use of data mining algorithm to predict the outgoing students from the first year [9].

Bharathwaj and Pal studied the student's background, social status, exam performance and college performance among the selected college student's of bachelor of computer applications.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

They selected seventeen criteria using Bayseian classification and tried to find the relationships between the background of the student, past educational performance, mother's qualifications, etc., [10].

Decision Tree Induction Model:

Naïve Bayse

Naïve Bayse is based on decision tree induction model for prediction. Mohammed M. Abu Tair et al. applied the educational data mining classification algorithms Naïve Bayse to predict the performance of the learners. The graduate learners' background data gathered from a college and they investigated the performance. They elaborately studied the methods such as data cleaning, data processing, data extraction and applying various methods related to classification, association, and clustering to predict the learners' performance. The data mining tool called Rapid Miner used by the researchers to find out the new knowledge. The researchers' used various to predict the low performed students and tried to improve their exam marks and performance based on the recommendations. Their research data were collected from the college data base. They applied various case study methods to find out the improvement areas. Rule Induction method and Naïve Bayes classifier were applied to study the performance of the students. Two types of classification methods used to study the performance. They were distance-based method and density-Based method.

J48 Classification algorithm

J48 algorithms' function is based on decision tree induction classification method. In this method of classification, one attribute of the node is developed for splitting the data. A tree structure is formed from the root information which provides the developed information or new knowledge.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

The advantage of J48 algorithm is better for prediction because it reduces the computing time. he branches of the nodes helps to form a new knowledge. Also it is efficient to form a class label with maximum accuracy. Pruning is a process in this algorithm which is more effective to improve the effectiveness of the accuracy and performance.

Random Tree

Random tree is a decision tree induction algorithm which functions based on randomly selected attributes. It helps to form a uniform of trees from the samples. It is more accurate because there is no pruning process in this algorithm. Also it helps to form probability of trees based on back fitting.

The stages of Random tree are as follows:

(i) New data set is developed from the training dataset. This process is developed from the process of replacement and sampling of new data.

(ii) Randomly selected nodes and leaves develop into a new tree.

(iii) Pruning is not used in random process for the development.

(iv) New information is tabulated from each classified tree of data.

Genetic Algorithm

Genetic algorithm is based on the selection and evolution model of chromosomes in genes. Genetic algorithm has as fitness based selection and recombination for solution. The parent genes produce the new offspring which has the set of new solution with the parent gene's chromosomes. The methods like selection, crossover, mutation, recombination, etc. are applied

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

in the GA. The fuzzy If-then logic is used in the GA to develop the new knowledge. In this method each chromosome represents the solution to the problem which has a fitness value.

Artificial Neural Network

It is also a biologically inspired method of algorithm which is more efficient and effective to find solution to data mining problems. Multi Layer Perceptron (MLP) is an efficient approximate model for prediction and classification of data mining. The MLP has a more complex feed forward non linear neurons with multi layer structure. The MLP structure can be illustrated by the following diagram.

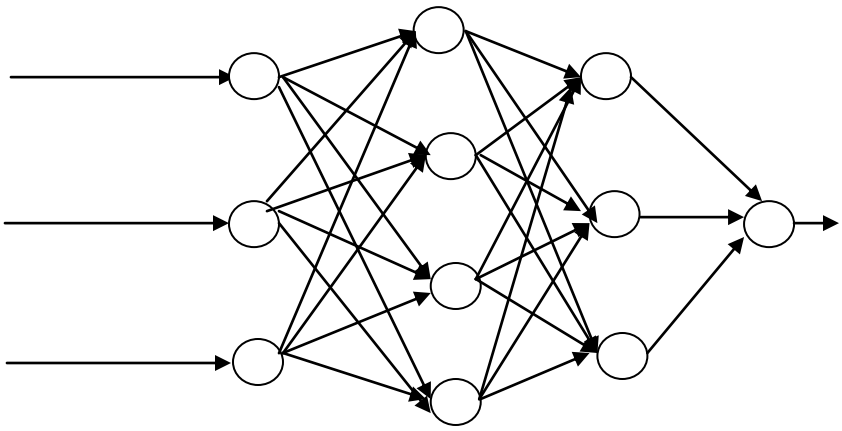


Figure 1: Artificial Neural Network Algorithm

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

3. Research Methodology

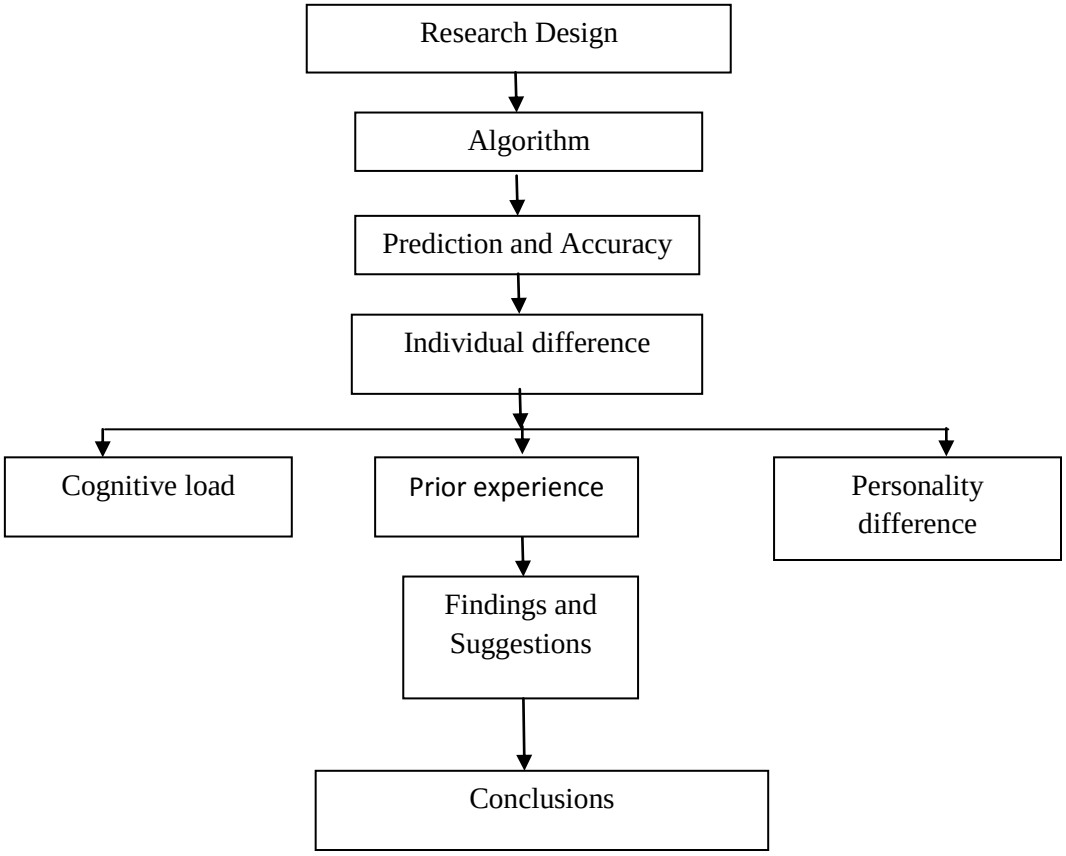


Fig: 1 Design Methodology

Profile of the Respondents:

- All the students studied B.Sc computer science from 2010 to 2015
- 65 % of the students preferred the same background in schooling and college education.
- Majority of the learners suffered due to their language differences between school and colleges.
- All the respondents are studying in autonomous colleges under a government university.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

4. Findings of the Research

The prediction of the outcome various depends upon the data set and the algorithm applied. The results shown here are gathered from the research outcome of the research work conducted by the researchers.

Algorithm	Accuracy in prediction cognition
Naïve Bayes	49%
J48	67%
Random Forest	75%
Random Tree	69%
MLP	72%
PMLP	76%
Genetically optimized PMLP	85 %

Those who have prior experience in the same type of educational environment, management, and similar factors have more possibilities to score well in their exam. Due to the comparable learning environments in the schooling and college education they have more possibilities to score well. Medium of instruction in the schooling impacts the performance in college education. Even though many other factors involves in the process of attaining exam performance, medium of instruction, prior experience about the institutional management and methods, individual differences have significant impacts on the exam performance. Same type of institutional transformation has many advantages which lead to better performance.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

5. Conclusions

This research work concludes that student's cognitive load, learning style, individual difference has major role in algorithm improvements. Improving the current classification algorithms is possible through the incorporation of individual differences, learning style and cognitive load. This research work helps to find the ways to improve the existing algorithms by the considerations of other factors such as learners' differences, learning style, etc. This research is also shows the family background has more impact on the adaptable nature of the student with the new environment. Exam performance difference happens due to the differences in their learning environment. From this study Genetically Optimized Multilayer Perceptron classification algorithms are more accurate in the prediction of performance than other algorithms. There are also possibilities exists to improve the existing algorithms by incorporation of individual differences, languages, prior learning experiences, etc.

Reference:

1. Christobal Romero, *"Data mining in education",* Wiley online library. DOI: 10.1002/widm.1075, 2012
2. Christobal Romero and ventura, *"A survey from 1995 to 2005",* ELSEVIER Educational data mining, DOI.org/10.1016/j.eswa.2006.04.005Get rights and content, 2006
3. Alex Freitas, *"A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery",* Advances in Evolutionary Computing. Natural Computing Series. Springer, Berlin, Heidelberg, pp 819-845 Springer 2003.

ISSN: 0971-1260
Vol-22-Issue-3-July-September-2019

4. Mitra and Pal*, "Interestingness measures for data mining: A survey",* IEEE transactions, Volume 13, Issue 3, 2006.
5. Shiqup wang*, "A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research"* Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on IEEE *Xplore*, 2010.
6. Ajay pal, *"Overview of the algorithms educational data mining for higher education institutions",* IJERT, vol 3. Iss 2 2014.
7. Mohamed Abu Tahir *"Mining educational data to improve student's performance",* IJICT, A case study 2017.
8. Ying Zhang*"A case study-use of data mining to improve students retention in higher education",* Conference: ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1, DISI, Funchal, Madeira, Portugal, June 8 - 12, 2010
9. Sourabi pal *"Mining educational data to reduce the dropout rates of engineering",* I J I Engineering and Electronic Business, 2012, 2, 1-7 published Online April 2012.
10. B.K. Bharadwaj and S. Pal. *"Data Mining: A prediction for performance improvement using classification",* International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136- 140, 2011.