International Journal of Pure and Applied Mathematics
Volume 118 No. 8 2018, 531-537
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Review on Prediction Algorithms in Educational Data Mining

A.Dinesh Kumar[1] , R.Pandi Selvam[2], K.Sathesh Kumar[3]
[1, 2] PG Department of Computer Science , Ananda College, Devakottai
[3]School of Computing, Kalasalingam University, Krishnankoil, – 626126, Tamil Nadu, India.
[1]dineshasoka661@gamil.com , [2]pandiselvamraman@gmail.com , [3]sathesh.drl@gmail.com

***Abstract*—In present day educational system, a student's performance is influenced by many factors. Students should be properly motivated to learn. Motivation leads to interest, interest leads to success. Proper assessment of abilities helps the students to perform better in their education. Data mining techniques have been applied to improve the performance of the student. Predicting student performance is one of the prominent research fields in EDM. Applying data mining methods in educational data is an interesting research area nowadays. Many studies of EDM have focused on the data mining algorithms related with the prediction. This paper reviews the prediction algorithms and data mining tools used in educational data mining and future insights of better prediction algorithm to be identified and new data mining tools to be used to predict the students' performance, which helps the instructor and institution to increase their study level.**

Keywords*—Data Mining, Educational Data Mining and Prediction.*

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data warehouse. KDD is the controlled process of identifying valid, new, useful, and understandable patterns from huge and complex data sets. Data Mining (DM) is the heart of the KDD process, concerning infer of algorithm that explore the data, develop the model and find out previously unfamiliar patterns. The model is used for understanding phenomena from the data, analysis and prediction [1,36].

Educational data mining is an independent research field in data mining. Applying data mining techniques in an educational data is a fascinating research field also known as Educational Data Mining (EDM). EDM examines data that are generated by the educational institution such as prediction of the student performance, learning analytics of the student, grouping the student according to their performance and recommending to the student.

EDM analyze data created by any type of information system supporting learning or education in schools, colleges, universities and other academic or professional learning institutions providing conventional model of teaching, as well as easy learning. Predicting students' results and student modeling have been the primary goals of educational data mining. These two issues are intensely connected with educational environment [2].

This paper is organized as follows. Section II introduces the related work of the prediction techniques. Section III provides prediction techniques used in educational data mining. Section IV provides a tabulated format of research works that have been carried out in prediction techniques that are used in educational data mining. Section V discusses the conclusion of this paper.

II. BACKGROUND STUDY

Mustafa Agaolglu conducted the research on predicting instructor performance in higher education. They have used the artificial neural networks, classification algorithms, decision trees, support vector machines and discriminant analysis algorithms for predicting the instructor performance. Their results shows C5.0 classifier is the best algorithm to predict the performance of instructor [3].

Sushil and Thakur applied fuzzy association rule mining to predict the student's performance in end semester. They used the

data like attendance, midsem marks, previous semester marks and previous academic records that collected from the previous data base. Based on these data they analysed some hidden patterns of student poor performance [4].

Mukesh Kumar and A.J. Singh explained their study based on the performance of 412 students. They used Naive Bayes, Random Forest, PART and Bayes Network for predicting the student performance. They were found random forest algorithm gave the best result comparing to other algorithms for predicting the student performance [5].

Kamal et al described a performance study on B.A first year students from Vikram University, Ujjain, India. They were used ID3, C4.5 and CART algorithm to predict the first year student performance. Based on these algorithms they found a final grade of student in a course [6].

Bevinda Alisha et al conducted a study on predicting drop out students' performance. They have used the attributes like Previous Semester Marks, Internal Grades to predict the student's final semester marks and also used different classification algorithms like ID3, C4.5, CART, CHAID. Their result shows that CHAID algorithm predicted the performance of drop out students with highest accuracy [7].

Shaleena and Shaiju Paul conducted a study on predicting the student performance by using decision trees, class imbalance and cost sensitive classification method. They found the relevant factors and relationships that lead to a student to pass or fail. They concluded some factors are related with the student failure [8].

Dinesh and R.V.Radhika invsetigated the performance of the students using feature selection method. They focused about the factors that are related with the performance of the student then they compared the environmental factor and educational institute factor which most affect the performance of the student. Finally they concluded environmental factors are affecting the performance of the student [2].

Dr.N. Tajuniza et al conducted a study on the factors that affect the academic achievement of the student. They used classification techniques and mapreduce technique to predict the performance of the student [9].

AS Galathiya et al analyzed a research on classification with an improved decision tree algorithm using feature selection technique. They used genetic search algorithm to improve the classification accuracy. The classification accuracy was improved by implemented the diversities of algorithm using RGUI with weka packages [10].

III. Prediction Techniques

In educational data mining, prediction techniques are used to predict the performance of the student. In order predict the performance of the student several tasks are used which are classification, regression and density estimation.

A. Classifications

Classification is a supervised learning technique whose aim is to create a model, in this specific case, called classifier, which can classify the class label of unknown data. In other words, a classifier is created from a training set and it is then used to classify unknown data, into one of the existing classes. Classification is a two step process: learning phase and the classification phase. Using the mapping function one can classify any attribute vector in the classification phase. To evaluate the classifier, an already classified input is considered and its accuracy is calculated as the percentage of correct classification obtained [11]. Several algorithms are used under classification tasks that have been applied to predict the performance of the student. They are Decision tree, Bayesian Classifier, Artificial Neural Network, Support Vector Machine and K-Nearest Neighbor algorithms.

B. Decision Trees

Decision trees are the best known classification model. DT is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision- making. Decision tree starts with a root node on which it is for users to take actions. From this node, users will split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome [17].

C. Bayesian Classifier

Bayesian classifiers are statistical classifier that is represented visually as a graph structure. This classifier predicts the class membership by probabilities, such as the probability that a given sample belongs to a particular class. Several Bayes algorithm have been developed, among which Bayesian and naïve Bayes are the two essential techniques. Naïve Bayes algorithm assumes that the effect that an attribute plays on a given class is independent of the values of other attributes [34].

D. Artificial Neural Network

Artificial neural networks are much admired in pattern recognition. It is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to

predict the correct class labels of the input tuples. Neural networks are well studied for continuous valued inputs and outputs. These are best at identifying patterns or trends in data and well studied for prediction of student performance [17].

E. Support Vector Machine

Support Vector Machines are best method when the class boundaries are nonlinear but here is too little data to learn composite nonlinear models. The fundamental idea is that when the data is mapped to a higher dimension, the classes become linearly divisible. In practice, the mapping is done only absolutely, using kernel functions. Support Vector Machine (SVM) focus on only the class boundaries; points that are any way easily classified are skipped. The objective is to find the "thickest hyperplane", which splits the classes [35].

F. K-Nearest Neighbor

K- Nearest neighbor classifier represents a totally dissimilar approach to classification. They do not build any clear universal model, but estimated it only locally and implicitly. The main idea is to classify a new object by examining the class values of the *K* most alike data points. The selected class can be either the most frequent class among the neighbors or class distribution in neighborhood. The only learning task in *K*-nearest neighbor classifiers is to select two important parameters; the number of neighbors *k* and distance metric *d* [35].

G. Regression

Regression analysis is a statistical methodology that is most often used in numeric prediction [12] .The objective of this task is to achieve a function of the independent variables that allows computing the conditional expectation of a dependent variable for prediction and forecasting exercises based on the minimization of a certain type of error via an iterative procedure. Nearly, Classification and Regression Trees (CART) summarizes these tasks [13].

H. Density Estimation

Density estimation is concerned with the estimation of probability masses, univariate densities, joint densities, and conditional densities. Most of the existing estimators assume that all the data instances are available at once. With the ever increasing amounts of data and a tendency towards online settings, however, there is an increasing demand for density estimation on data streams (LINK). Some density estimation techniques are inference, pattern mining, or outlier detection [14].

IV. Prediction Methods Used In EDM

For the purpose of this study, 20 papers were chosen from Educational Data Mining. We have analyzed the performance of the student is predicted by many classification algorithms. The research has been done to improve the performance of the student based on predicted results.x This prediction helped the instructor and institution to know about the weak student status and take proper assessment take on the student to improve their study level. In this paper we have reviewed the papers that have been published in the year from 2011 to 2017. Form their research work the paper title, algorithms, tools and results are listed in the below table. Most of the research has been done under decision tree classification algorithms naïve bayes classification algorithm to predict the performance of the student.

TABLE I

Prediction Algorithm As Applied In EDM

No	Title of Paper	Algorithm	Tools Used	Results
1	Data Mining: A Prediction for Performance Improvement Using Classification [15]	Bayesian Classification	MatLab	Predicted the High Potential Variable that effect student's performance
2	Data Mining: A Prediction of performer or under performer using classification [16]	Bayesian Classification	MatLab	Predicted Student's Final Mark
3	Mining Educational Data to Analyze Students' Performance [17]	ID3	Weka	Predicted The student End Semester Performance
4	Data Mining for Engineering schools Predicting Student Performance and Enrollment in Master Programs[18]	CART	Matlab	Predicted overall performance of students

5	Efficiency of Decision Trees in Predicting Student's Academic Performance [19]	C4.5 and ID3	Weka	Predicted student performance based on Internal and External Marks
6	Prediction of student performance by an application of data mining techniques [20]	Support Vector Machine, K-Means Clustering	Rapid miner	Predicted the relationships between student's behavioral and their success
7	Data Mining: A Prediction for performance Improvement of Engineering Student Using Classification [21]	C4.5, ID3 and CART	Weka	Predicted first year Engineering students performance
8	Data Mining techniques in EDM for predicting the performance of students[22]	OneR, C4.5, Multilayer Percepton, Nearest Neighbour algorithm	Weka	Predicted the factors that affect the student's performance
9	Mining Student Academic Performance [23]	Naïve Bayes	Weka	Predicted Students Academic Performance
10	Data Mining approach for Predicting Student Performance [24]	Naïve Bayes, Multilayer Perceptron and C4.5	Weka	Predicted Student success in a course and the performance of learning methods
11	Data Mining: A Prediction for Student's Performance Using Classification Method[25]	ID3	Weka	Predicted the final grade of the student
12	Mining Educational Data to Predicting Higher Secondary Students Performance	ID3, C4.5	We ka	Predicted Educational and Environmental Factors of student
	[2]			
13	Mining Educational Data to Predict Student's Performance Using Ensemble Methods[26]	Artificial Neural Network , C4.5	Weka	Predicted students behavioral features
14	Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms [27]	Naïve Bayes	Weka	Predicted student behavior of academic performance
15	Prediction of Students Performance using Educational Data Mining[28]	Naïve Bayes	Weka	Predicted student performance at top of the semester.
16	Educational Data Mining & Students' Performance Prediction[29]	C4.5 ID3, CART and CHAID	Weka and Rapid Miner	Predicted the students' performance based on related personal and social factors.
17	Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron[30]	Linear Regression, Multilayer Perceptron	Weka	Predicted student performance in final examination
18	Predicting Student Academic Performance using Data Mining Methods[31]	Decision Tree, ANN, Random Forest Tree	Rapid Miner	Predicted students graduation performance in final year

19	Is Alcohol Affect Higher Education Students Performance: Searching and Predicting pattern using Data Mining Algorithms [32]	BFTree, C4.5, RepTree and Simple CART	Weka	Predicted student performance based on alcohol consumption student data set
20	A Comparative Analysis of Decision Tree Algorithms for Predicting Student's Performance[33]	ID3, C4.5, CART, CHAID	Weka	Predicted Student Performance in University Results

Based on table we have listed the prediction algorithm as a figure that is most often used by the researchers in educational data mining for prediction. The Fig.1 Shows the prediction algorithm mostly used in EDM.

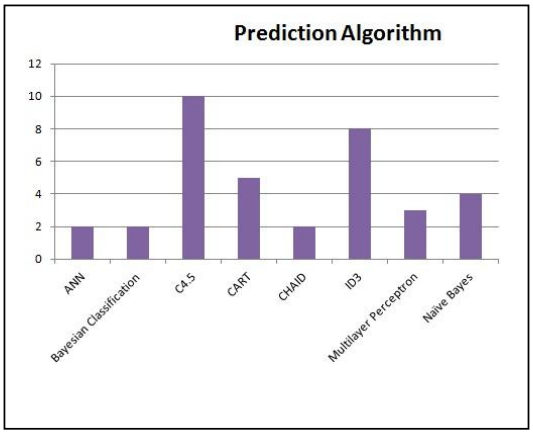

Fig.1 Prediction Algorithm of EDM

The Fig.1 says that the C4.5 (J48) is often used by most researchers to predict the performance of the student then ID3, CART and Naïve Bayes algorithms are frequently used by the researchers. The other algorithms are rarely used by the researchers. Some researchers used combination of these algorithms to predict the performance of the student in educational data mining.

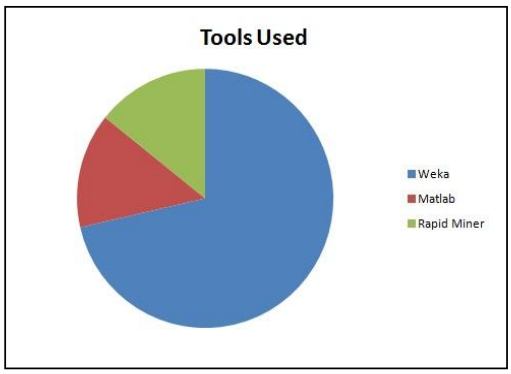

Fig. 2 Tools used in EDM

The Fig.2 illustrate most of the research has been applied in weka tool for prediction. Matlab and Rapid miner tools are rarely used to predict the performance of the student by the researchers.

V. Conclusion And Future Work

In this paper different prediction techniques and prediction tools are discussed in educational scenario to predict student performance. Prediction of student academic performance helps both the teacher and parents to predict about their success and failure in examinations. From this paper we concluded that most of the research has been done on same prediction algorithms, student variables and data mining tools. Their work was only different from to improve the classification accuracy and different factors related to the student performance. Symbiotic structure learning algorithm and feed forward neural-network-aided grey model is not yet applied for predicting student performance. In future this algorithm and model can be applied for predicting student performance or new algorithm; new student variable and new data mining tools can be also identified for better prediction based on this study.

References

[1] Oded Maimon Lior Rokach, "Data Mining and Knowledge Discovery Handbook",Second Edition, Springer.

[2] A.Dinesh Kumar ,Dr.V.Radhika, "Mining Educational Data to Predicting Higher Secondary Students Performance", International Journal of Computational Intelligence and Informatics(IJCII), Vol 6 ,September 2016.

[3] Mustafa Agaoglu," Predicting Instructor Performance Using Data Mining Techniques in Higher Education", IEEE, Vol 4, 2016.

[4] Sushil Kumar Verma, R.S. Thakur, "Fuzzy Association Rule Mining based Model to Predict Students' Performance", International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, No. 4, August 2017

[5] Mukesh Kumar Prof. A.J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance", Modern Education and Computer Science, 2017, 8, 25-31.

[6] Kamal Bunkar Rajesh Bunkar," Data Mining: Prediction for Performance Improvement of Graduate Students using Classification", IEEE, 2012.

[7] Bevinda Alisha Pereira, Anusha Pai," A Comparative Analysis of Decision Tree Algorithms for Predicting Student's Performance", International Journal of Engineering Science and Computing, Vol 7, 2017.

[8] Shaleena K.P Shaiju Paul," Data Mining Techniques for Predicting Student", International Conference on Engineering and Technology (ICETECH), IEEE 2015.

[9] Dr.N.Tajunisha and M.Anjali," Predicting Student Performance Using Mapreduce", International Journal of Engineering and Computer Science (IJECS) ,Vol.4, Issue1-2015.

[10] As.Galathiya and AP.Ganatra, "Classification with an improved Decision Tree Algorithm", International Journal of Computer Application (IJECS), Vol 46, 2012.

[11] Ricardo Mendes And Joao P.Vilela, " Privacy- Preserving Data Mining: Methods, Metrics, and Applications", IEEE, 2017.

[12] Jiawei Han and Micheline Kamber ," Data mining concepts and techniques", Second Edition.

[13] Chady EI Moucary "Data mining for Engineering Schools", International Journal of Advanced Computer Science and Applications(IJACSA) Vol.2, 2011.

[14] www.datamining.informatik.uni-mainz.de

[15] Birijesh Kumar Bharadwaj, Saurabh Pal," Data Mining: A Prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol 9, 2011.

[16] Umesh Kumar Pandey S.Pal, "Data Mining:A Prediction of performer or underperformer using classification", International Journal of Computer Science and Information Technologies (IJCSIT) Vol 2(2),2011.

[17] Birijesh Kumar, Saurabh Pal, " Mining Educational Data to Analyze Students' Performance", International Journal of Advanced Computer Science and Applications Vol 2, 2011.

[18] Chady EI Moucary, "Data Mining for Engineering schools Predicting Student Performance and Enrollment in Master Programs", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, 2011.

[19] S.Anupama Kumar, Dr.Vijayalskhmi, "Efficiency of Decision Tress in Predicting Student's Academic Performance", CCSEA,2011.

[20] Sajadin Sembiring and Zarlis ,"Prediction of student performance by an application of data mining techniques",International Conference on Management and Artificial Intelligence IPEDR vol.6 2011.

[21] Surjeet Kumar Yadav, Saurabh Pal, " Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), Vol 2, 2012.

[22] Ajay Kumar Pal, Saurabh Pal, " Data Mining Techniques in EDM for Predicting the Performance of Students", International Journal of Computer Science and Information Technology, Volume 2, 2013.

[23] Azwa Abdul, Nur Hafieza, Fadhilah Ahmad, " Mining Students' Academic Performance", Journal of Theoretical and Applied Information Technology", Vol 53, 2013.

[24] Edin Osmanbegovic and Mirza Suljic, "Data Mining approach for Predicting Student Performance", Journal of Economics and Business, May 2012.

[25] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby, "Data Mining: A Prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, 2014.

[26] Elaf Abu, Amrieh, Thair Hamtini, Ibrahim Aljarah, "Mining Educational Data to Predict Student's Performance Using Ensemble Methods", International Journal of Database and Theory and Application,Vol 9, 2016.

[27] Patinon Galvan," Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms", Future Technologies conference IEEE,2016.

[28] Ms.Tismy Devasia,Ms.Vinushree , "Prediction of Students Performance using Educational Data Mining ", Data Mining and Advanced Computing (SAPIENCE) IEEE-2016.

[29] Amjad Abu Saa , "Educational Data Mining & Students' Performance Prediction,International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, 2016

[30] Febrianti Widyahastuti , Viany Utami Tjhin , "Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron", IEEE, 2017.

[31] Raheela Asif , Saman Hina and Saba, "Predicting Student Academic Performance using Data Mining Methods", International Journal of Computer Science and Network Security (IJCSNS), VOL.17, 2017.

[32] Saurabh Pal**,** Vikas Chaurasia, "Is Alcohol Affect Higher Education Students Performance: Searching and Predicting pattern using Data Mining Algorithms", International Journal of Innovations & Advancement in Computer Science IJIACS, Volume 6, April 2017.

[33] Bevinda Alisha, Anusha Pai, "A Comparative Analysis of Decision Tree Algorithms for Predicting Student's Performance", International Journal of Engineering Science and Computing (IJESC) ,Vol 7,2017."

[34] Dorina Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernatics and Information Technologies, Vol 13, 2013.

[35] C Romero, Ventura, Pechenizkiy, and Baker Rsjd," Handbook of Educational Data mining", 2010.

[36] Shankar, K. "Prediction of Most Risk Factors in Hepatitis Disease using Apriori Algorithm." RESEARCH JOURNAL OF PHARMACEUTICAL BIOLOGICAL AND CHEMICAL SCIENCES 8.5 (2017): 477-484.