2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)
GLA University, Mathura, Oct 26-28, 2017

Analyzing Performance of Students by Using Data Mining Techniques

A Literature Survey

Sagardeep Roy
Department of Computer Science and Engineering
ASET, Amity University
Noida, India
roy.sagard@gmail.com

Anchal Garg
Department of Computer Science and Engineering
ASET, Amity University
Noida, India
agarg@amity.edu

***Abstract*—With the birth of new technologies which can harness data associated with education, the field of Educational Data Mining (EDM) has bloomed. EDM is a research area which uses data mining techniques, machine learning algorithms and statistical techniques to understand how students learn, predict students' academic performance and how a student's learning can be improved. This paper conducts extensive review of the literature on the use of EDM for analyzing performance of student.**

Keywords—Educational Data Mining, Learning Analytics, Predicton, Classification, Regression, Decision Tree

I. INTRODUCTION

The capacity to predict performance of a student could be valuable in an extraordinary number of ways. Education system all over the world has changed rapidly since vast research in the field of Educational Data Mining (EDM) and Learning Analytics (LA). The use of Data Mining (DM) techniques, machine learning methods and different statistical techniques in education is EDM. EDM uses the above mentioned techniques to explore data from educational settings to find out different patterns of student's behaviors and predict performance of Student [1]. LA is similar to EDM and is a collection and analysis of usage data associated with student learning. This paper presents a literature review available on predicting student performance. It also explains various data mining techniques by which student's performance can be predicted. Predicting student performance now can be considered a sub domain of EDM and LA with the emphasis more on predicting student's success and failure and help student to achieve success. In this paper, will be discussing various data mining methods, machine learning algorithms and statistical techniques used in predicting student performance. The paper is divided into five sections. Various related literature review is presented in the second section. Third section describes data mining techniques and popular classification techniques used and tools which are generally used in research. Fourth section compares different classifiers; tools which supports techniques and authors cited by references who used techniques, algorithms and tools in their research, all represented in tables. Fifth section concludes the paper.

II. LITERATURE REVIEW

There are many papers published in the application areas of EDM & LA. Much research has been done in predicting academic performance of a student. In [2], a research was done to predict drop out students. The results proved that decision tress gives good accuracy. In [3], ANN techniques were applied to predict academic performance of students. In [4], three classification methods namely Naïve Bayes, neural networks and decision trees were used to predict student academic achievement. Naïve Bayes produced better results. In [5], Kabakcheiva used CRISP-DM model to predict student academic performance. He applied a Decision Tree (J48) classifier, Bayesian classifiers such as Naïve Bayes, BayesNet, Nearest Neighbor method (IBk) and rule learners namely OneR and JRip. Decision Tree had better accuracy than the rule learner JRip and IBk. Bayesian Classifiers had least accuracy. In [6], regression analysis was applied to predict student's marks in a distance learning system. In [7], Genetic Programming was used with DM techniques to predict student failure. In [8], different DM techniques such as Decision Trees, Random Forest, Neural Networks, SVMs and regression techniques were used to predict secondary school student performance. They concluded that it is possible to predict final grades if previous grades are known. CRISP-DM is more popular with researchers as other DM process models such as SEMMA which is developed by one of the popular and biggest producers of analytics software, SAS Institute. [9]. It concentrates only on modeling tasks forgetting the business angles unlike CRISP-DM which has "Business Understanding Phase". SEMMA intends to help clients of SAS Enterprise Miner. Applying it outside can be of no use. [10]. In [11], C. Romero introduced and applied classification via clustering while not using customary classification methods to predict pass or fail class. In [12], Jia et al predicted Students retention by combining SVM and neural network to improve classification accuracy. Predicting Student Performance has become popular after many detailed research in EDM and LA. Both research area has gave birth to many research areas. The trend in EDM methods was started by Romero and Ventura. There survey of EDM research from 1995 to 2015 consisting of more than 50 papers almost spanning a decade stated that relationship mining methods became the most distinguished

978-1-5386-3004-4/17/$31.00 ©2017 IEEE

EDM research in that decade as 43% of those papers involved Relationship Mining Methods. In that decade from 1995 to 2005, Prediction became the most popular and well-known research area after relationship mining as 27% of those articles used prediction methods. Clustering and Human judgment followed with 15% and 17% of those papers. [13]. But with the introduction of Big Data technologies, more concentration shifted towards prediction of Student Performance as it is the single most important factor in both EDM and LA. After 2005, prediction slowly became the most popular research area as it replaced Relationship Mining Methods. The first two years of EDM conference namely Baker's research in 2008 and Barnes work in 2009, saw a different pattern when compared with Romero and Ventura survey. "Relationship Mining" which was leading the trends in EDM from 1995 to 2005, got replaced by "Predicting Performance" and shifted to fifth place in 2008-2009 with only 10% of papers used the topic. Prediction which was second most researched area in 1995- 2005 became the top as many as 45% of papers published in those two years. Clustering and Human judgment research area did not lose much importance and relevance in 2008-2009 with 15% and 12% papers respectively. [14]. Both EDM and LA is associated with improving the quality of Education and finding ways how to improve student's learning ability with the help of DM techniques. Both topics have one common sub-domain which is Predicting Student Performance. This was found out by searching keywords "Educational Data Mining" and "Learning Analytics" on Google Scholar. The first 20 search results of both have some topics in common. The most common topic was Student or Leaner's performance prediction.

III. Data Mining Techniques

Some popular Data mining techniques explained here are Regression, Nearest Neighbor, Clustering and Classification.

A. Data Mining Methods

- Classification and Regression are main and most important DM techniques. Regression is mainly used when there is a need to predict values of variables which are dependent by estimating the relationship among variables. It helps to determine the degree of relationship between the variables and can extrapolate the values of dependent variable when the values of independent variables are known.
- Nearest Neighbor is a technique where the values are predicted based on the predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.
- Clustering- Data is divided into groups of similar objects. Data simplification is given priority and as a result some details are ignored. [15]
- Classification is the identification of the category or class to which a value belongs to, on the basis of previously categorized values. Here we are mainly concerned about Classification and Regression techniques.

B. Classification

The most used and useful data mining technique is classification as described by Ahmed and Ibrahim. They used decision tree classifier, a type of classification technique to predict student's final marks. [16]. Classification technique is a standout technique used by most researchers. Classification is a type of data analysis that extracts models describing data classes. A classifier predicts categorical labels (classes). [17]. In [18], one classification and one clustering method were used. SVM and K-means methods were applied to analyze the relationship between behavior of students and their academic success. It was concluded that psychological factors of student's affects their final grades. Classification techniques in DM are proficient to process a huge amount of data. There are different classification techniques. Some of them are explained below with features and limitations-

- C4.5 Algorithm: It is a Decision Tree Algorithm. The main features of C4.5 algorithm are as follows- C4.5 algorithm can be easily interpreted. C4.5 algorithm is very easy to implement. C4.5 algorithm uses both discrete and continuous values. Some limitations of C4.5 algorithm are as follows- Little deviation in data can lead to different decision tree. It works well with large datasets. If one has a small dataset one should not apply C4.5 algorithm on it.
- ID3 Algorithm: ID3 Decision Tree algorithm produces better accurate results than C4.5 Decision Tree algorithm. ID3 algorithm's detection rate is very high and space consumption is very small. ID3 algorithm takes a lot of searching time. ID3 algorithm produces very long rules which are very hard to prune. To store tree, ID3 algorithm needs a large memory.
- K-Nearest Neighbor Algorithm: K-NN algorithm has one distinct advantage as classes do not have to be separated linearly. There is no cost in the learning process. It is strong enough to handle noisy data. K-NN has many limitations. If the dataset is large, it can be very time- consuming. K-NN is sensitive to noise. The performance of K-NN algorithm depends on the dimensions used.
- Naïve Bayes Algorithm: Naïve Bayes Algorithm are easy to implement. The classification rate of Naïve Bayes is very high. It predicts accurate results for most of the classification and predication problems. Naïve Bayes algorithm needs large training dataset. Naïve Bayes algorithm's precision rate decreases if dataset is small. Results can be good only if data is large.
- Support vector-machine algorithm: SVM algorithm has high accuracy if dataset is small. It does not work well with large datasets. If there is missing value in data it's very difficult to deal with it.
- Artificial Neural Network Algorithm: It is very easy to use, it works with incomplete data. It learns quickly and implementation is rather easy. It is applicable to

most problems in real life. There are limitations and some of them are, it requires high processing time if neural network is large. Table 1 in Section IV provides the comparison of different classifiers.

C. *Tools*

Some open source tools which are used in applications of EDM and LA are as follows

- R: It is an open source language used for statistical and data analysis. It can run in multiple platforms (e.g. Windows, MacOS or Linux). [8]
- WEKA: It stands for Waikato Environment for knowledge analysis. It is non-propriety, freely available, and application-neutral standard for data mining projects. It is widely adopted in academic and business and has an active community (Hall et al., 2009) [9]
- Orange: It is a free data mining tool that also supports Data visualization. [20]
- Tanagra: It is an open source environment for teaching and research and is the successor to the SPINA software. [20]

IV. Comparison of Classifiers, Tools Which Supports Techniques And Authors Who Used Techniques, Algorithms, Tools In Their Research

TABLE I. Comparison Of Classifiers

CLASSIFIERS	PROS	CONS
Decision Tree	Easy to interpret and has capability of reasoning	Needs lot of data, difficult to deal with missing data
Bayesian Networks	Easy to interpret and has the capability of updating and reasoning	Needs a large training data
Support Vector Machines	Nonlinear application and has accuracy in small data	Cannot deal with missing data and does not support mixed variable
Neural Network	Can work with incomplete data and has the capability of both updating and reasoning	Difficult to deal with missing data and does not support mixed variable. Needs a lot of data to train
K-Nearest Neighbor	Can work with incomplete data and has the capability of updating	Needs a lot of data and it can't deal with missing data

TABLE II. Tools Which Support Techniques

	R	WEKA	Orange	Tanagra
K-Means Clustering	Supports	Supports	Supports	Supports
Regression	Supports	Supports	Supports	Supports
Naïve Bayesian Classifiers	Supports	Supports	Supports	Supports
Decision Tree	Supports	Supports	Supports	Supports
Time Series Analysis	Supports	Supports	Does not support	Does not support
Big Data Processing	Supports	Supports	Does not support	Does not support
Text Analytics	Supports	Supports	Supports	Does not support

TABLE III. Authors Who Used Techniques, Algorithms And Tools In Their Research

Reference	Techniques	Algorithms	Tools
2	Classification	SimpleCart, J48, BayesNet, Logistic Regression, JRip, RandomForest	WEKA
3	Classification	Artificial neural networks (ANN)	SPSS v.19 – Neural Network Module
4	Classification	Naïve Bayes, J48, MLP	WEKA
5	Classification	Naïve Bayes, BayesNet, J48, IBk, OneR, JRip	WEKA
6	Regression	Model Trees M5, Neural Networks, Linear regression, Locally weighted linear regression, SVM	WEKA
7	1. Classification. 2. Grammar-based genetic programming (G3P)	1. Rule based- JRip, NNge, OneR, Prism, Rindor. Tree based- J48, SimpleCart, ADTree, RandomTree, REPTree. 2. Interpretable Classification Rule Mining (ICRM)	WEKA
8	1. Classification (Binary, 5-level). 2. Regression.	1. Decision Trees, Random Forest, Neural Networks, SVM. 2. Logistic Regression.	R

11	1. Classification. 2. Classification via Clustering.	1. Rule based-DTNB, JRip, Nnge, Ridor. Tree based- ADTree, J48, LADTree, RandomForest. Function based- Logistics, MLP, RBFNetwork, SMO. Bayes based- BayesNet, NaiveBayes. 2. EM, Hierarchical Cluster, sIB.	WEKA
12	Classification	They used all classification algorithms	WEKA

V. Conclusion

Due to in depth research in education, birth of new advanced technologies which can harness data associated with education to improve the quality of Education with emphasis more on predicting student's success and failure and help student to achieve success, predicting student performance has become a very popular research area. With thorough research in the field of EDM and LA the new trend is predicting student's performance with the help of Data Mining Techniques, Machine Learning Algorithms and Statistical techniques and approaches. Much research has been done in "Predicting Student Performance". Most research has been done by collecting data through questionnaires. Most of the work has been done through different Classification Techniques, using CRIPS-DM model. Most Common tools used are WEKA and R.

References

[1] B. Guo, R. Zhang, G. Xu, C. Shi and L. Yang. "Predicting students performance in educational data mining.", 2015 International Symposium on Educational Technology (ISET), Wuhan, 2015, pp. 125-128, doi: 10.1109/ISET.2015.33

[2] G. Dekker, M. Pechenizkiy and J. Vleeshouwers. "Predicting students drop out: A case study." In Educational Data Mining 2009.

[3] M.F. Musso, E. Kyndt, E. C. Cascallar and F. Dochy, "Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks", Frontline Learning Research, 1(1):42–71, 2013

[4] E. Osmanbegović , M. Suljić, "Data Mining Approach for Predicting Student Performance", Journal of Economics and Business, Vol. X, Issue 1, May 2012

[5] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetics and Information Technologies- The Journal of Institute of Information and Communication Technologies of Bulgarian Academy of Sciences, DOI: 10.2478/cait-2013-0006

[6] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades",Artificial Intelligence Review, 37(4):331–344, 2012.

[7] C. Vera, A.Cano, C. Romero,S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data", DOI 10.1007/s10489-012-0374-8

[8] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance", In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[9] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185. Archived January 9, 2013, at the Wayback Machine.

[10] S. S. Rohanizadeh, M.B. Moghadam, "A Proposed Data Mining Methodology and its Application to Industrial Procedures", Journal of Industrial Engineering 4 (2009) pp 37-50

[11] C. Romero, M. López, J. Luna, S. Ventura, "Predicting students' final performance from participation in on-line discussion forums", ELSEVIER, Computers & Education 68 (2013) 458–472, 2013

[12] Jia, Ji-Wu. "Machine learning algorithms and predictive models for undergraduate student retention at an HBCU." PhD diss., Bowie State University,2013.

[13] C. Romero and S. Ventura (2007), "Educational data mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146.

[14] R. Baker, K. Yacef, "The state of educational datamining in 2009: A review and future visions", 2009, JEDM-Journal of Educational Data Mining 1.1 (2009), pp. 3-17.

[15] P. Berkhin, "Survey of Clustering Data Mining Techniques" , 2006, pp 25-71, 10.1007/3-540-28349-8_2

[16] D. Ahmed, I. Sayed, "Data Mining: A Prediction for student's performance using classification method", World journal of Computer Application and Technology-2014

[17] C. Romero, S. Ventura, "Data Mining algorithm to Classify Students", IEEE Trans Sys Man Cybern.

[18] S. Sembering, M.Zarlis, " Prediction of student academic performance by an application of data mining techniques", International conference on management and Artificial Intelligence-2011.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutmann & I. Witten (2009), "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter, 11(1), 10-18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.3671

[20] Wimmer, Hayden, and Loreen Marie Powell. "A comparison of open source tools for data science." Journal of Information Systems Applied Research 9, no. 2 (2016): 4.