ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93885, May 2016* ISSN (Online) : 0974-5645

Relevance and Resonance of Data Science in Performance Prediction and Visualization

A. Princy Christy* **and N. Rama**

Post Graduate and Research Department of Computer Science, Presidency College, Chennai, India;
princyphilana@gmail.com, nramabalu@gmail.com

Abstract

Objectives: This paper aims at comparing and contrasting relevant researches on predicting the performance of students in Data Science perspective that encompasses machine learning and data mining. **Analysis:** An architectural framework has been devised for educational data mining. The ultimate motive is to extensively investigate the techniques like classification, regression and recommender systems in predicting the student performance and to explore the prediction accuracy of these techniques as well. For this purpose very many researches that have successfully implemented these techniques were carefully studied and their contribution to predicting the performance were analysed. **Findings and Novelty:** It came to light that ensembles created by combining classifiers performed well and their accuracy in predicting the performance was commendable compared to the individual performance of the classification, regression and recommender techniques. The nuance of this study is the incorporation of recommender systems along with conventional techniques since these are not commonly used in performance prediction. Tensor factorization in particular has desirable effect in prediction since it takes the time factor into consideration. It is a fact that performance of students increases over time.

Keywords: Ensemble Classifiers, Neural Networks, Performance Prediction, Recommender Systems, Regression

1. Introduction

The significance of extracting useful information from educational data has gained momentum because of the impact it can have on the teaching and learning paradigm. Large repositories of data produced by the educational sector can be quantitatively explored to provide meaningful insights for enhancing teaching and learning, predicting the performance of students, identifying irregular learning process, grouping students, predicting dropouts to name a few. These large data come from a varied source such as establishment of state databases, data from MOOCs, ITS (Intelligent Tutoring Systems) and so forth. By analysing and exploring past data performance of the students can be forecasted and methods can be devised to help the students to perform well in the course. By analysing and exploring past data performance of the students can be forecasted and methods can be devised to help the students to perform well in the course.

The contribution of data mining to education can be viewed in two perspectives one is research and the other is societal. The research perspective brings about implementation of various data mining techniques and methods that would help in understanding and enhancing the pedagogical principles, while societal perspective aims at helping various stakeholders like teachers, students, administrators and researchers to put into use the results of research thereby envisaging a learned and civilised society. Since universities and colleges think of devising curriculum in par with the industry and on-going trends in the world there is a lot of emphasize in understanding the needs of the education community.

There are many popular data mining methods used in understanding educational data such as classification, clustering, outlier detection, relationship mining, Social Network Analysis, Process mining and text mining[1]. This study is on steps involved in data science but focuses mainly on the data mining technique, prediction of students' performance using machine learning techniques and

* *Author for correspondence*

compares and contrasts various supervised algorithms and techniques employed in predicting the performance of students. The supervised learning techniques classification, regression and the recommender systems are explored here along with algorithms and techniques used in these with proven study by many research articles.

2. Need for Data Science in Education

The contribution of data science to business world is manifold and the success rate would be tangible hence its use in business modelling and prediction has increased in terms of maximizing the profit of the businesses. Though the outcome of data science implementation in education produces intangible benefits its application has been embraced by the education community over the last few years. Data Science process involves Data Selection, Pre-processing, Transformation, Data mining, interpretation and evaluation as stated in an edx MOOC. It is an iterative process. The data science process by Joe Blitztein and Hanspeter pfister created for the Harvard data science course visualizes this iteration. The Figure 1 presented below is the combination of the two statements mentioned above and can be mentioned as Data Science – Iterative Process (based on our interpretation).

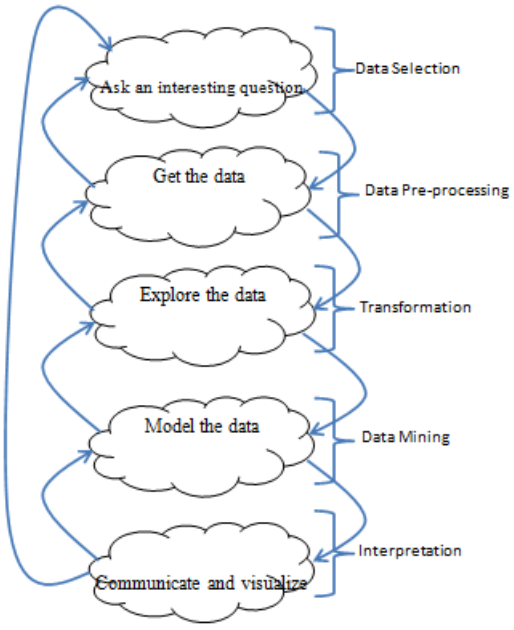

Figure 1. Data Science – Iterative.

The data selection and pre-processing tasks this not discussed in this article though accuracy in prediction or any task related to data analysis has significant dependence on the type of data. The more relevant the datasets are the more accurate would be the results of any data mining task. One of the processes of data science which is data mining is discussed in terms of the tasks involved in prediction of performance.

From the architectural viewpoint as shown in Figure 2, educational data mining can be divided into operational and analytical. Operational view deals with the functional entities of any educational environment like students, teachers, administrators, demographics, motivation, background, psychometry and the like. Analytical view deals with knowledge inferences from the data stored which is implementing technology for extraction and analysis of knowledge from the historical data obtained by the operational process.

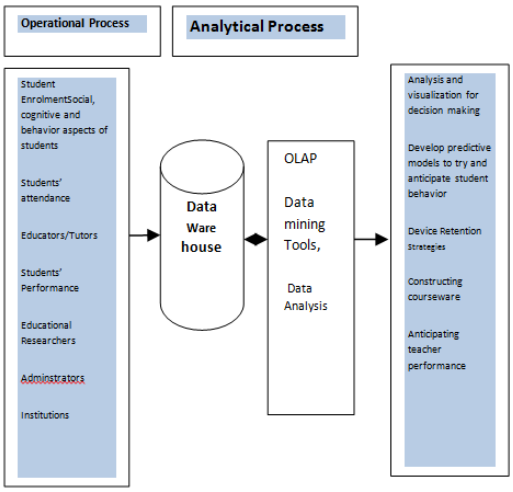

Figure 2. Architectural framework for educational data mining.

The knowledge that is discovered can be used for analysing and visualizing which would lead to decision making, predicting performance of students, devising strategies to thwart retention, understanding student behaviour and change teaching methodology accordingly by instructors, utilizing available resources more efficiently by administrators, personalized learning and to recommend courses to the students, foreseeing teachers' performance based on analysing students' assessment[2]. Educational environment can be broadly classified into

traditional classroom and technology driven learning. In both the setup students' role is considered to be crucial.

3. Data Mining Tasks in Building Prediction Models

Forecasting students' performance has become crucial in understanding and enhancing the behaviour of students, learning capability and teaching strategy. The aim of prediction is to deduce a target attribute from many combinations of other features of data. Many different types of prediction methods are employed like classification, regression and density estimation to name a few. These methods can be used depending on the type of variable marked for prediction. Classification can be employed if the predicted variable is a categorical value or a, regression method if a predicted variable is a continuous value and density estimation if a predicted variable is a probability density function[3].

While classification is a supervised learning technique that is used to predict a Boolean true or false value for an object with a given set of features, regression also a supervised learning technique predicts real numeric label values (y) from a vector of one or many known feature values (x). In Classification and regression the goal is to produce function that calculates the known label values in the training dataset accurately and should also generalize the known values in the test dataset for accurate prediction. These functions would generate over-fitting or under-fitting problem. This should be eliminated for accurately predicting. Apart from the methods stated above there are other techniques and ensemble methods used for prediction. The next sections dwell upon analysing the various methods. The main traits of data mining approaches in predicting the performance of students has been summarized in Figure 3. Also unsupervised learning technique clustering is used in one of the studies along with classification.

4. Classifiers and Ensemble Models

An ensemble model of classifiers was created using three online algorithms namely 1-NN, NB and WINNOW to predict student performance in distance education[4]. Online algorithms have been used instead of batch algorithms since the scenario dealt with continuous and large dataset. So, the need for storing and reprocessing of each instance can be eliminated as it would be expensive[5]. An ensemble of classifiers using several linear models including simple averaging, linear SVM, linear regression and logistic regression was created by a team of researchers. This was to predict the performance of students in an online environment. The team found that regularized linear regression directly minimizes RMSE (Root Mean Square Error), which was their evaluation criterion[6].

The classifiers, Quadratic Bayesian classifier, 1-Nearest Neighbour (1-NN), k-Nearest Neighbor (k-NN), Parzen-window, multilayer perceptron (MLP), and Decision Tree were employed by another team. They first tried to model prediction capability using these classifiers separately and later found that combining all the classifiers improved the prediction process significantly. Additionally by learning an appropriate weighting of the features used through a Genetic Algorithm (GA) the performance accuracy of combined classifier was further increased to about 10 to 12 % accuracy of the ensemble of classifiers[7]. Linear classifier support vector machine provided higher percentage of accurate predictions and specifically was found to predict the performance of individual students[8]. The final performance of first year computer science students was predicted from their participation in on-line discussion forums, using traditional classifiers and classification through clustering[9]. Many clustering algorithms were chosen and finally EM (Expectation Maximization) Algorithm was found to produce higher accuracy and F-measure (Harmonic mean of precision and recall) similar to their classifier counterparts.

Artificial Neural Network (ANN) is capable of modelling very complex non-linear functions. The MLP (Multilayer Perceptron) architecture of ANN with back propagation was used to predict secondary education placement test scores. Though MLP can be used to produce both classification and regression prediction models classification is used. Many ANN, decision trees and SVM were compared and contrasted. ANN provided accurate prediction while trying to predict the CGPA (Cumulative grade Point Average) of the students upon graduation[10]. In another study C5 decision tree algorithm predicted the placement test scores with an accuracy of

0.95[11] among the other models that are SVM (Support Vector Machine), ANN Logistic regression. They ranked these models based on the prediction accuracy as follows: C5, SVM, ANN, and Logistic Regression. The evaluation criteria employed were k- fold cross validation and sensitivity analysis. MLP in Neural Networks, Random Forests and Decision Tree and Linear discriminant Analysis were compared and contrasted for university students' performance prediction from various factors that influence their achievement. To the modified data by using ensemble model Random Forest, CART algorithm was used to build decision trees. These models were ranked based on their accuracy in the following order: discriminant Analysis and closely followed by neural networks and random forests[12]. Multiple Instance learning paradigms was explored by using many classifiers[13]. It focussed on Decision Trees, Logistic Regression, SVM, and Neural Networks for predicting performance of student considering various factors like reading forums, writing forums, quizzes and assignments.

5. Regression

Multiple linear regression models have been recommended to find the average academic performance of an entire class with the cumulative GPA of the students as the only predictor variable[14]. Regression models were developed based on multivariate linear regression technique to predict the academic performance of engineering students. Four models were developed based on the predictor variables determined through method of least squares[15]. At each point in the linear representation of data as shown in the Figure 3 each data point will have an error associated with its distance from the regression line.

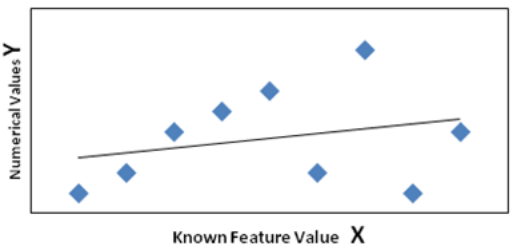

Figure 3. Regression line.

Stepwise linear regression model was identified to predict the MCAS (Massachusetts Comprehensive Assessment System) scores of students[16] using the ASSITment system to assess school students performance based on which they would be inducted into graduation programs.

Regression models use various features that include past course performance of students. Linear multi-regression model performed on a dataset extracted from the Moodle installation of University of Minnesota produced higher RMSE (Root Mean Square Error) of 0.147 compared to linear regressions whose RMSE was 0.177. RMSE kept minimal indicates the prediction accuracy of the algorithm or model since it is difference between the measure of the differences between value predicted and the values actually observed by the model. This speaks about the efficacy of the algorithm or model. Thus adding multiple linear regressions reduces RMSE and improves the prediction accuracy[17]. The over-fitting problem encountered by this study was overcome by solving the minimization process. Performance of students based on CGPA was predicted using a model that used linear regression. This model used RASE (Root of Average Squared Error) for estimating the accuracy of the validation dataset and was found to be 0.1848[18].

6. Recommender Systems

Recommender systems most commonly used in e-commerce applications have found its way into learning environment. A good example for recommender systems was the Netflix contest for creating a recommender system for recommending movies for movie goers. This formed the basis for few researchers to try recommender systems in prediction.

6.1 Matrix Factorization

The recommender system technique of matrix factorization is used to predict the performance of students in the intelligent tutoring systems since it was believed that recommender systems was not explored much in this domain[19]. This was compared with the traditional approaches like regression and has been proved that recommender systems perform well.

Factorization techniques were studied that belong to latent factor models and further tensor factorization was proposed in predicting student performance by taking into account the temporal effect since the knowledge of the students improves over time which is a natural fact[20].

6.2 Collaborative Filtering

A model-based Collaborative Filtering (CF) technique with IRT to analyse the students responses and there by predicting their performance was carried out[21]. Here CF has been used to identify a best log-linear model that has higher prediction accuracy by training a class of these models with CF on the data. It has been suggested that owing to the speed and easy generalizability of collaborative filtering it can be used in analysing student performance.

Collaborative filtering technique of the recommender systems was applied in another study to predict the capability of students to choose the right answer based on historic results. KNN known as user based CF and matrix factorization was adopted[22].

In addition to the above methods classification, Association rule mining, and clustering were employed to analyse the performance of students involved in virtual learning[23]. In a comparative analysis of classification algorithms to predict the performance of students, it was found that there was no uniformity in the prediction rates among the classifiers used such as J48, KNN, OneR and JRip[24].

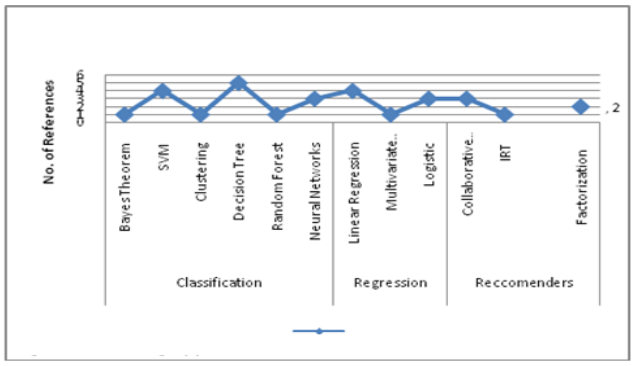

Figure 4. Data mining approaches referred.

7. Discussion and Conclusion

The paper extensively covers recent studies in predicting student performance implementing data mining tasks and employing the supervised learning tasks classification, regression and recommender systems. Each of the techniques in their own ways influenced the outcomes of the prediction task. Recommender systems apart from recommending objects are used in performance prediction and tend to outperform the

Table 1. Traits of data mining approaches in performance prediction

Task	M: Method T: Technique	A: Algorithm F: Function E: Equation
Classification	M: Bayes Theorem	Naïve Bayes, Quadratic Bayesian Classifier A: Genetic Algorithm
	M: SVM	F: Non- Linear Kernel Function A: SMO
	M. Clustering	Expectation Maximization,
	M: Decision Tree, T: Random Forest	A: CART, C5, J48, RepTree
	M: Neural Networks T: MLP	A: Back Propagation A: RBF
Regression	T: Linear Regression T: Multivariate linear Regression T: Logistic	A: Multiple Linear Regression A: Stepwise linear regression A: MILR
Recommenders	M: Collaborative Filtering T: M: IRT M: Factorization	E: Regularized Logistic Regression A: KNN A: Tensor Factorization A: Matrix Factorization

traditional classification and regression models in certain datasets. In many studies ensemble models were found to yield accurate prediction results than their individual counterparts. The Table 1 displays the implementation of various methods in predicting student performance. It has been inferred that traditional methods like classifiers and regression are most commonly used than the recently explored recommender systems as inferred from the graph shown in Figure 4. But recommender systems have been employed lately in many researches related to mining educational data.

8. References

1. Romero C, Espejo PG, Zafra A, Romero JR, Ventura S. Web usage mining for predicting final marks of students that use Moodle courses. Comput Appl Eng Educ. 2013; 21(1):135–46.
2. Barracosa J, Antunes C. Anticipating teachers performance. KDD 2011 Workshop: Knowledge Discovery in Educational Data; 2011. p. 77–82.
3. Romero C, Ventura S. Data mining in education. WIREs Data Mining Knowl Discov. 2013 Jan-Feb; 3(1):12–27.
4. Kotsiantis S, Patriarcheas K, Xenos M. A combinational incremental ensemble of classifiers as a technique for predicting students, performance in distance education. Knowledge-Based Systems. 2010 Aug; 23(6):529–35.
5. Littlestone N, Warmuth M. The weighted majority algorithm, Information and Computation. 1994 Feb; 108(2):212–61.
6. Yu H-F, Lo H-Y, Hsieh H-P, Lou J-K, McKenzie TG, Chou J-W, Chung P-H, Ho C-H, et al. Feature engineering and classifier ensemble for KDD cup 2010. Workshop and Conference Proceedings. 2010; 1:1–16.
7. Minaei-Bidgoli B, Kash DA, Kortemeyer G, Punch WF. Predicting student performance: an application of data mining methods with the educational web-based system Lon-capa. 33rd ASEE/IEEE Frontiers in Education Conference; 2003 Nov.
8. Huang S, Fang N. Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Computers and Education. 2013 Feb; 61:133–45.
9. Romero. Predicting students final performance from participation in on-line discussion forums. Computers and Education. 2013 Oct; 68:458–72.
10. Ibrahim Z, Rusli D. Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. 21st Annual SAS Malaysia Forum; 2007 Sep.
11. Sen B, Uçar E, Delen D. Predicting and analyzing secondary education placement-test scores: a data mining approach. Expert Systems with Applications. 2012 Aug; 39(10):9468–76.
12. Superby JF, Vandamme JP, Meskens N, Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of International Conference Intel. Tutoring Systems Workshop Educational Data Mining; 2006. p. 1–8.
13. Zafara A, Romero C. Multiple instance learning for classifying students in learning Management Systems. Expert Systems with Applications. 2011 Nov-Dec; 38(12):15020–15031.
14. Huang S, Fang N. Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Computers and Education. 2013 Feb; 61:133–45.
15. Huan S, Fang N. Regression models of predicting student academic performance in an engineering dynamics course. American Society for Engineering Education. 2010. p. 1–17.
16. Feng M, Heffernan NT, Koedinger KR. Addressing the assessment challenge in an online system that tutors as it assesses. User Modeling and User-Adapted Interaction: The Journal of Personalization Research. 2009 Aug; 19(3):243–66.
17. Elbadrawy A, Studham RS, Karypis G. Collaborative multi-Regression models for predicting students, performance in course activities. Proceedings of the Fifth International Conference on Learning Analytics and Knowledge; 2015 Mar. p. 103–7.
18. Ibrahim Z, Rusli D. Predicting students academic performance: comparing artificial neural network, decision tree and linear regression. 21st Annual SAS Malaysia Forum; 2007 Sep.
19. Nguyen T-N, Drumond L, Krohn-Grimberghe A, Schmidt-Thieme L. Recommender system for predicting student performance. Procedia Computer Science; 2010. p. 2811–9.
20. Nguyen T-N, Drumond L, Krohn-Grimberghe A, Schmidt-Thieme L. Factorization techniques for predicting student performance. Educational Recommender Systems and Technologies: Practices and Challenges. 2010; 1(2):2811–9.
21. Bergner Y, Dröschler S, Kortemeyer G, Rayyan S, Seaton D, Pritchard DE. Model-based collaborative filtering analysis of student response data: machine-learning item response theory. Proceedings of the 5th international conference on educational data mining; 2012 Jun. p. 95–102.
22. Oscher AT, Jahrer M. Collaborative filtering applied to educational data mining, KDD Cup 2010. Journal of Machine Learning Research. 2010.
23. Provide a method for increasing the efficiency of learning management systems using educational data mining. Available from: http://www.indjst.org/index.php/indjst/article/view/82454/63612
24. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. Available from: http://www.indjst.org/index.php/indjst/article/view/74555/58051