Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012061 doi:10.1088/1757-899X/551/1/012061

Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods

Samuel-Soma M Ajibade[1], Nor Bahiah Binti Ahmad[2], Siti Mariyam Shamsuddin[3]

[1,2,3]Computer Science, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia.

E-mail: samuel.soma@yahoo.com, bahiah@utm.my, mariyam@utm.my

Abstract. Nowadays, Educational Data Mining (EDM), begun as a new research area due to the broadening of numerous statistical approaches that are used to perform data exploration in educational settings. One of the applications of EDM is the prediction of performance of students. In a web based education system, the behavioural features of learners are very significant in showing the interaction between students and the LMS. In this paper, our aim is to propose a new performance prediction model for students which is based on data mining methods which includes new features known as behavioural features of students and based on sequential feature selection which is used to identify most important features. The proposed performance model is evaluated using classifiers such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Decision Tree (DT). Furthermore, so as to enhance the classifiers performance, the ensemble methods such as Bagging, Boosting and Random Forest were applied. The achieved results show that there exists a strong relationship between behaviour of students and their academic performance. An accuracy of 91.5% was gotten when the ensemble methods were applied to the classifiers to improve the academic performance. Thus, the result gotten shows the reliability of the proposed model.

1. Introduction

EDM comprises of the utilization of machine learning techniques, Data Mining (DM) techniques and diverse educational statistical methods. EDM is a discipline that is taken after to extract significant knowledge from an educational setting. EDM applications such as model development helps to predict the performances of students. As a result, driving researchers to delve profound into different techniques of mining data to enhance existing techniques [1]. Originally, EDM and LA were significant areas of research in education, but gradually student performance prediction began to become prevalent as the main aim of the study is to examine and predict the performance of students for the students to obtain better results. This research work presents ensemble methods for enhancing various classification algorithm through which the student performance can be predicted by introducing a new feature category called behavioral features. The educational dataset is gotten from an eLearning system known as Kalboard 360 [2]. The gathering of the data carried out by making use of a learner activity tracker tool known as Experience API (XAPI). The obtained features are into three different categories: academic background features, demographic features and then behavioral features. The educational process consists of a new category which is the behavioral features and it is linked to the experience of learners.

In this research work, we utilized of educational data to predict the academic performances of students. This model has therefore evaluated the effect of the learning behavioral features of students on their academic performance. A data mining technique known as classification has been used to implement this work. We made use of three classifiers: Decision Tree (DT), K-Nearest Neighbor and

Support Vector Machine (SVM). In order to improve the performance of the classifiers, some ensemble techniques like Boosting, Bagging and Random Forest were used to enhance the performance model accuracy of the students. The remainder of this paper is as follows: Literature review is presented in Section 2 and the third Section describes the methodology while the fourth section presents the result and discussion of this research and Section 5 gives the conclusion and future work.

2. Experimental

To build a predictive model, there are several DM techniques used, which are classification, regression and clustering. Decision tree is a set of conditions arranged in a hierarchical frame. Most of researchers used this technique due to their simplicity, in which it can be transformed into a set of classification rules. Some of the famed DT algorithms are C4.5 [3], [4] and CART. In the work of [5] the objective of the author is to build the prediction model for students for first and second degree students of Computer Science & Engineering and Electronics & Communication streams by making use of two classifiers: Decision Tree(DT) and Fuzzy Genetic Algorithm. Parameters like internal marks, sessional marks and admission score were chosen to carry out this research. SVM is a learning algorithm developed by [6] to handle the challenges of pattern recognition and prediction and also for analysis and mapping of both linear and non-linear functions. A hyperplane or set of hyperplanes (classes)are being created in a high dimensional space which can then be used for classification [7], [8]. SVM was selected for the research because of its generic application nature and variety of applicability. KNN is a straightforward classification algorithm that stores every single accessible case and classifies new cases centered on a similarity measure (e.g., distance functions). The training samples are defined by n dimensional numeric qualities. Each sample signifies a point in an n-dimensional space. Along this lines, the majority of the training samples are kept in an n-dimensional pattern space. At a point, when an unknown sample is provided, a KNN classifier searches the pattern space for the k training samples that are closest to the unknown sample.

Ensemble method [9], [10], [11] combines two or more classification algorithm which is known as a based learner. The base learner can be identical or non-identical. Ensemble method organizes the predicted results of used base learners to achieve more strength in the system. In most cases, ensemble techniques such as Bagging, Boosting, and Random Forest are applied to improve the performance of classification algorithm [12]. Bagging ensemble method [9], [10] randomly takes data from training dataset and put into a bag. Several numbers of bags are created where each bag contains the subset of training data set. Then each bag is trained with a classification algorithm provides a model. In boosting ensemble method [10], [11] total training set used to train the first model and next model is trained from the performance of the previous model. At first, it gives equal weights to each instance of the dataset. If the class of an instance is misclassified, then gives it more weight to focus on this instance in the later model. This process is continued until the number of added model or accuracy is obtained. Random Forest algorithm [10], [12] is a large collection of Decision Tree algorithms which are not correlated. Random Forest creates a lot of Decision Trees from the subsets of training dataset where each subset provides a decision tree. Now each Decision Tree model classified an instance in a class. Then the majority voted class will be taken as the class of the instance. In [13], authors use some common ensemble techniques (such as Bagging, Adaboosting, and Random Forest) to predict student's academic performance more accurately where Adaboosting on Artificial Neural Network gives the best accuracy of 79.1 percent.

3. Methodology

The Educational dataset that was used in this paper is collected from a LMS called Kalboard 360. The Kalboard 360 is a multi-agent LMS that was developed to enhance learning via the utilization of leading-edge technology. The data was gathered by utilizing a learner activity tracker tool known as Experience API (XAPI). XAPI is a component of the Training and Learning Architecture (TLA) that allows the tracking of learning experiences and actions of learners like reading an article or watching an educational video. In this present study, the dataset lengthens into 500 students with 16 features. Table 1 displays

the attributes/features, data description and data type of the dataset. After the data preprocessing the dataset, the feature variable consists of 480 observations with 8 features. we used cvpartition to divide data into a training set of size 336 and a test set of size of size 144. Therefore, to evaluate the model, 10-fold cross validation was used. Forward sequential selection was used in a wrapper fashion to find important features.

Table 1. Student features, description and data type.

Feature Category	Features	Description	Data Type
Demographical Features	Nationality	Student nationality	Nominal
	Gender	The gender of the student (female or male)	Nominal
	Place of Birth	Student's place of birth (Kuwait, Jordan, Lebanon, Saudi Arabia, Iran, USA)	Nominal
	Parent responsible for student	Student's parent such as (father or mum)	Nominal
Academic Background Features	Stage ID	Stage Student belongs to such as (Low level, Middle level, High level)	Nominal
	Grade ID	Grade students belongs such as (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12)	Nominal
	Section ID	Classroom student belongs to such as (A,B,C)	Nominal
	Semester	School year semester such as (First or second)	Nominal
	Topic	Course topic such as (Math, English, IT, Arabic, Science, Quran)	Nominal
	Student Absence Days	Student absence days (Above-7, Below-7)	Nominal
Parents Participation on learning process	Parent Answering Survey	Parent is answering the surveys that provided from school or not	Nominal
	Parent School Satisfaction	This attribute obtains the degree of parent satisfaction from school as follow (Good, Bad)	Nominal
Behavioural Features	Discussion groups	Student behavior during interaction with Kalboard 360 e-learning System.	Numeric
	Visited resources		Numeric
	Raised hand on class		Numeric
	Viewing announcements		Numeric

4. Results and Discussions

4.1. Evaluation results using traditional classifiers

In predicting the performances of students, numerous features have effect on the model. In this paper, the behavioral features have been considered as vital features that can have effect on the performance of students. As shown in table 2, we have shown the results through classification algorithms (DT, KNN and SVM) so as to demonstrate the impact of behavioral features. The classification results are fragmented into two different sections. Results of Classification with student's behavioral features (BF) and results of Classification without student's behavioral features (WBF). In the table, we can deduce that the DT model performs better than other data mining techniques. The DT model achieved 87.1% accuracy with BF and 84.4% accuracy WBF. For precision measure, the model achieved 85.2% with BF and 77.7% WBF. For recall measure, the results are 86.3% with BF and 78.1% WBF. For F-Measure, the results are86.0% with BF and 77.1% WBF. Hence from the above analysis, the results prove a strong effect of learner behavior on academic performance of students.

4.2. Evaluation results using ensemble methods

In this section, we made use of ensemble methods to enhance the accuracy of the evaluation results of the traditional Data Mining techniques. Table 3 shows the enhanced results using ensemble methods with three traditional classifiers (DT, KNN and SVM). Each ensemble trains the three classifiers and then now combine the results by a majority voting process in order to achieve the best prediction performance of students. The boosting techniques performs better than other ensemble methods in the case of DT, KNN and SVM, however DT gave the highest performance in which the accuracy of DT using boosting is enhanced from 0.87 to 0.89 while Precision result are improved from 0.85 to 0.87 and the Recall results are improved from 0.86 to 0.87 and F-measure result was enhanced from 0.86 to 0.89.

Table 2. Classification method results with BF and WBF.

Evaluation Measure	DT		KNN		SVM	
Behavioral Feature Extraction	BF	WBF	BF	WBF	BF	WBF
Accuracy	0.87	0.84	0.82	0.81	0.86	0.83
Precision	0.85	0.77	0.80	0.67	0.81	0.70
Recall	0.86	0.78	0.79	0.56	0.81	0.73
F-Measure	0.86	0.77	0.69	0.59	0.75	0.71

Table 3. Classification method results using ensemble methods.

Evaluation Measure	Classification Methods			Bagging			Boosting			Random Forest
Classifier	DT	KNN	SVM	DT	KNN	SVM	DT	KNN	SVM	DT
Accuracy	0.87	0.82	0.86	0.86	0.85	0.86	0.89	0.86	0.88	0.89
Precision	0.85	0.80	0.81	0.86	0.79	0.80	0.87	0.83	0.82	0.84
Recall	0.86	0.79	0.81	0.87	0.80	0.82	0.87	0.83	0.84	0.84
F-Measure	0.86	0.69	0.75	0.88	0.69	0.76	0.89	0.74	0.80	0.73

After the classification model has been trained using 10-folds cross validation, then the process of validation kicks-off. The process of validation is a very significant phase in the structuring of predictive models, it defines the accuracy of the predictive models. Table 4 displays the results of evaluation by the use of classification techniques (DT, KNN, and SVM) through the testing and validation phases.

Table 4. Classification method results through testing and validation.

Evaluation Measure	Testing Results			Validation Results		
Classifier	DT	KNN	SVM	DT	KNN	SVM
Accuracy	0.87	0.82	0.86	0.91	0. 85	0.88
Precision	0.85	0.80	0.81	0.87	0.83	0.84
Recall	0.86	0.79	0.81	0.89	0.82	0.84
F-Measure	0.86	0.69	0.75	0.89	0.75	0.82

As seen in Table 4, 91.5% accuracy is achieved in our proposed model through the validation phase. When compared to [15] et al which gave 82.2% accuracy, our model performed better. Hence, the result gotten from the validation phase proves the reliability of the proposed model.

5. Conclusion and future work

The prediction of student's academic performance has been a huge concern for higher institutions everywhere. The data gathered entails some hidden knowledge that are being used to improve the academic performance of students. In this research, a new performance prediction model for students was proposed which is based on various data mining methods which contains new features known as behavioral features. These attributes are associated with the interactivity of learners with the LMS. The predictive model is evaluated based on some classifiers like DT, KNN and SVM. Furthermore, we applied ensemble methods to enhance the performance of the classifiers. We made us of Bagging, Boosting and Random Forest. Based on forward sequential selection that was used to select the most important features, the accuracy of the predictive model is 91.5% which is a better performance than [13] that was 82.2%. In future works we will focus on analyzing the data of students to find other features that will identify the students that have weaker achievements and performances. Optimization techniques such as Differential Evolution, Genetic Algorithm and others could as well be applied to enhance the performance model of students in educational data mining.

References

[1] Amrieh E A *et al* 2015 *IEEE Jordan Conference* pg. 1-5.
[2] Quadri M M *et al* 2010 *Global Journal of Computer Science and Technology* **10** 2-5.
[3] Kumar M *et al* 2016 *International Journal of Computer Applications* **137** 2.
[4] Hamsa H *et al* 2016 *Procedia Technology* **25** 326-332.
[5] Costa E B *et al* 2017 *Computers in Human Behaviour* **73** 247-256.
[6] Subaira A *et al* 2014 *IEEE 8th Int. Conf. on Intelligent System and Control (ISCO)* **978** 274-280.
[7] Trstenjak B *et al* 2014 *IEEE Int. Conv. on Info. And Comm. Tech. Elect. And Microelectronics (MIPRO)* 1222-1227.
[8] Marquez-Vera C *et al* 2016 Expert Systems **33** 107-124.
[9] Dietterich T G *et al* 2000 *Int'l. Workshop on Multiple Classifier Systems.* 1-15.
[10] Sabzevari M *et al* 2018 *Cornell Uni. arXiv preprint arXiv:1802.07877.*
[11] Rahman M H *et al* 2017 *2nd ICEEE* 1-4.
[12] Amrieh E A *et al* 2016 *Int'l. Journal of Database Theory and Application* **9** 119-136.
[13] Moisa V *et al* 2013 *Journal of Mobile Embedded and Distributed Systems* **5** 70-77.