Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance

Elaf Abu Amrieh
Computer Information Systems Department
The University of Jordan
Amman, Jordan
Ilef.kram@hotmail.com

Thair Hamtini
Computer Information Systems Department
The University of Jordan
Amman, Jordan
thamtini@ju.edu.jo

Ibrahim Aljarah
Business information Technology Department
The University of Jordan
Amman, Jordan
i.aljarah@ju.edu.

Abstract— **Educational data mining concerns of developing methods to discover hidden patterns from educational data. The quality of data mining techniques depends on the collected data and features. In this paper, we proposed a new student performance model with a new category of features, which called behavioral features. This type of features is related to the learner interactivity with e-learning system. We collect the data from an e-Learning system called Kalboard 360 using Experience API web service (XAPI). After that, we use some data mining techniques such as Artificial Neural Network, Naïve Bayesian, and Decision Tree classifiers to evaluate the impact of such features on student's academic performance. The results reveal that there is a strong relationship between learner behaviors and its academic achievement. Results with different classification methods using behavioral features achieved up to 29% improvement in the classification accuracy compared to the same data set when removing such features.**

Keywords— Educational Data mining, E-learning, Student Performance Prediction, Classification, Behavioral Factors.

I. INTRODUCTION

Nowadays we are noticing that the amount of educational data is increasing rapidly. Educational data mining is a new emerging field that concerns with developing methods for solving educational problems by discovering the hidden knowledge from data that come from educational environments [1]. Educational data are collected from different sources such as educational institute databases, e-learning systems and traditional surveys. This data contains valuable hidden information that can be extracted by educational data mining such as Decision Tree, Naïve Bayes, and many others [2 and 3]. The discovered knowledge from educational data can be used to help the decision makers at any educational institutions to enhance the educational systems and produce high quality outcomes.

In this paper, we introduce a new student performance model with a new category of features, which called behavioral features. The educational dataset is collected from e-learning system that called Kalboard 360 [4]. This model applies some of data mining techniques on such data set, for evaluating student's behavioral features impact on student academic performance. Furthermore, we try to understand the nature of this kind of features by expanding data collection and pre-processing steps.

The data collection process is accomplished using a learner activity tracker tool, which called experience API (XAPI). The collected features are classified into three categories: demographic features, academic background features and behavioral features. The behavioral features are a new feature category that related to the leaner experience during educational process. To the best of our knowledge, this is the first work that employs this type of features/attributes. After that, we use three of the most common data mining methods in this area to construct the academic performance model: Artificial Neural Network (ANN) [5], Decision Tree [6], and Naïve Bayes [7].

The rest of this paper is organized as follows: Section II presents related works in educational data mining. Section III presents the data collection and preprocessing. Section IV presents the proposed methodology. Section V reports an experiments and results. Finally, we conclude this paper with summary remarks and future work in Section VI.

II. RELATED WORKS

In recent years, some researches have applied data mining techniques to help instructors and administrators to improve e-learning systems. In [8], the authors used data mining techniques to explore some factors having an impact on students' success in Istanbul University. This research developed a model to extract the features that affects students' achievement by using the path analyses. The authors to focus more on the factors that have impact on students' academic success.

The authors in [9] deuced that the student's success is related to the school environment and school management. While in [10], the authors discovered that the teacher is playing the main role in student success. In [1], the authors introduced a case study to analyze students' learning styles using educational data mining. The goal of this study was to show how data mining techniques can help in improving students' performance in higher education. The data set is collected from course database that include personal records and academic records of students. The authors in [11] categorize the students into five groups according to their performance using Expectation-Maximization Algorithm (EM-clustering) which it is depends on the maximum likelihood estimates of parameters in probabilistic models.

978-1-4799-7431-3/15/$31.00 ©2015 IEEE

Shannaq et al. in [12] applied classification technique to predict the numbers of enrolled students, by studying the main attributes that may affect the students' loyalty. The authors collected 2069 sample records from student's database, and then build a classification model using a decision tree method to select the main attributes that may have impact on students. This research allows the university management to prepare necessary resources for the new enrolled students in higher education systems.

Ayesha et al. in [13] used k-means clustering algorithm as a data mining technique to predict students' learning activities in a students' database, which includes class quizzes and exams. After that, the collected information will be transmitted to the class teacher before the conduction of final exam. This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

In summary, various researches have been investigated to solve the educational problems using data mining techniques. However, very few researches shed light on student's behavior during learning process and its impact on the student's academic success. This research will focus on the impact of student interaction with the e-learning system. Furthermore, the extracted knowledge will help schools to enhance student's academic success. In addition, to help administrators to improve learning systems.

III. DATA COLLECTION AND PREPROCESSING

The data set used in this paper is collected from Kalboard 360 E-Learning system using Experience API (XAPI) [14]. XAPI is a component of the Training and Learning Architecture (TLA) that enables to track learning experiences and learner's actions like reading an article or watching a training video. The Experience API helps the learning activity providers to define the learner, activity and objects that describe a learning experience [14]. The goal of X-API in this research is to track student behavior during the educational process for evaluating the features that may have an impact on student's academic performance.

The collected data set consists of 150 student's record with 11 features. The features are classified into three main categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as stage, grade and section. (3) Behavioral factors such as raised hand on class, opening resources, participating in discussions groups, viewing messages and announcements.

Improving student's academic achievement start with improving student's behavior and encourage the student to engage in the classroom. Student engagement is one of the important researches in educational psychology field. Student engagement was defined by Gunuc and Kuzu [15] as "the quality and quantity of students' psychological, cognitive, emotional and behavioral reactions to the learning process as well as to in-class/out-of-class academic and social activities to achieve successful learning outcomes".

Kuk [16] refer to the student engagement as spending time by students in educational activities. According to Stovall [17], student engagement includes not only the spending time on tasks but also their desire to participate in some activities. There are many definitions and researches that defined student's behavior and engagement. All of these researches confirm the strong relationship between students' behavior and student's academic achievement. Table I shows the dataset's attributes/features and their description. As shown in the table, we can notice a new feature category which is a behavioral factor. This type of attributes is related to the learning experiences and learner behavior during the educational process.

After the data collection task, we apply some preprocessing mechanisms to improve the quality of the data set. Data preprocessing is considered an important step in the knowledge discovery process, which includes data cleaning, feature selection, data reduction and data transformation.

TABLE I: STUDENT FEATURES AND THEIR DESCRIPTION

Feature Category	Feature	Description
Demographical Features	Nationality	Student nationality
	Gender	The gender of the student (female or male)
	Place Of Birth	Place of birth for the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Relation	Student's contact parent such as (father or mum)
Academic Background Features	Stage ID	Stage student belongs such as (Lower level , Middle level , and high level)
	Grade ID	Grade student 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)
	Section ID	Section student belongs such as (A, B, C).
	Semester	School year semester such as (First or second).
	Topic	Course topic such as (Math, English, IT, Arabic, Science, Quran)
	Teacher ID	Teacher who teach this particular course.
Behavioral Features	Raised hand on class	Student Behavior during interaction with Kalboard 360 e-learning system.
	Opening resources	
	discussion groups	
	Viewing announc4ments	

Data cleaning is applied on this data set to reduce the noise, and missing values. The data set contains 17 missing values in various features from 150 records, the records with missing values are removed from the data set, the data set after cleaning becomes 133 records. The data set includes 85 males and 48 females. Stage ID includes 64 lower level, 47 Middle level, and 22 High level. Also, the students are distributed to three sections such as: 69 students from section A, 49 students from section B, and students 15 from section C. Topic attribute includes: 92 students are related to IT topic, 15 to Math topic, 15 to English topic, 4 to Quran topic, 4 for Science topic and 3 students for Arabic topic. Relation attribute includes 111 students, their contact person is the father and 22 students, the contact person is their Mother.

Feature selection is the process that focus on reducing the number of attributes that appearing in the patterns, in which it reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data, and increase the comprehensibility of the mining results [23, 24]. In this paper, we applied filter-based approach using information gain based selection algorithm to evaluate the features ranks to check which features are most important to use them to build students' performance model [18, 25].

IV. METHDOLOGY

In this paper, we introduce a student's performance model using classification techniques, to evaluate the features that may have an impact on student's academic success. Fig.1 shows the main steps in the proposed methodology. This methodology starts by collecting data from Kalboard 360 E-Learning system using Experience API (XAPI) as mentioned in section III. This step is followed by data preprocessing step, which concerns with transforming the collected data into a suitable format.

After that, we use discretization mechanism to transform the students' performance from numerical values into nominal values, which represents the class labels of the classification problem. To accomplish this step, we divide the data set into three nominal intervals (High Level, Medium Level and Low Level) based on student's total grade/mark such as: Low Level interval includes values from 0 to 69, Middle Level interval includes values from 70 to 89 and High Level interval includes values from 90-100. The data set after discretization consist of 58 students with Low Level, 53 students with Middle Level and 22 students with High Level.

After that, feature selection process is applied to choose the best feature set with higher ranks. As shown in Fig.1, we applied filter-based technique for feature selection. After all these steps, we construct classification model by using three different classification techniques.

In this paper, the classification is applied to evaluate the features that may have an impact on the performance/grade level of the students. The classification techniques used to evaluate the student's performance are Naïve Bayesian [BN] classifier, Decision Tree [DT], and Artificial Neural Network (ANN).

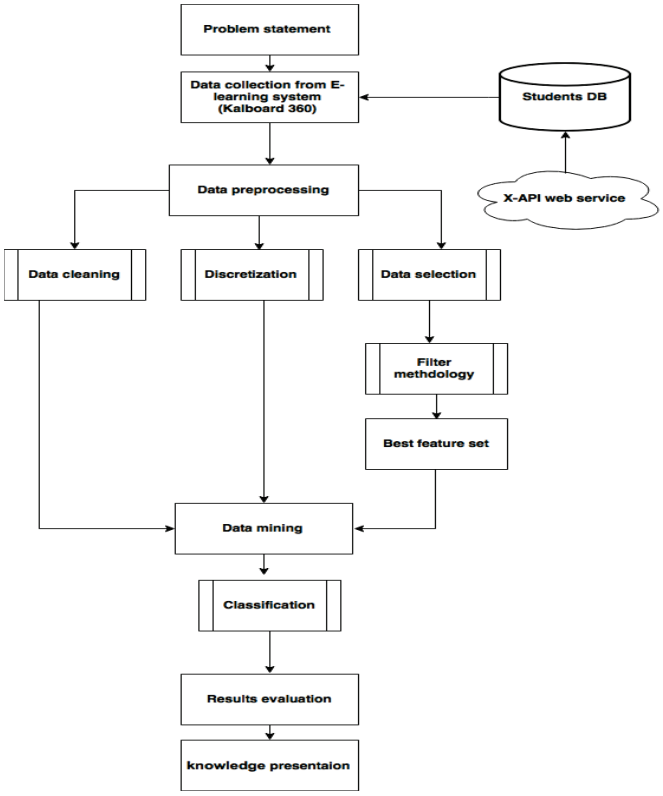

Fig. 1: Architectural diagram of the student performance

NB classifier [7] is a technique to estimate the probabilities of attributes values, given a class, from training data and then use these probabilities to classify new entities. DT [6] is an automatic rule discovery technique that produces a set of branching decisions that end in a classification; it works best on nominal attributes, so numeric ones need to be split into bins.

ANN [5] is an application of artificial neural network that concerns on training data inputs for achieving the best accuracy. ANN model consist of three layers: (1) input layer, (2) hidden layer and (3) output layer. The input layer receive input from the user program and output layer send output to user program too. Between the input layer and output layer are hidden layers. Hidden layer neurons are only connected to neurons and never directly interact with the user program. Then the process will be followed by the evaluation of results and patterns to generate the knowledge representation.

V. EXPERIMENTS AND RESULTS

A. *Enviroment*

We ran the experiments on the PC containing 6GB of RAM, 4 Intel cores (2.67GHz each). For our experiments, we used WEKA [20] to evaluate the proposed classification models and comparisons. Furthermore, we used 5-fold cross validation to divide the dataset into training and testing partitions.

B. Evaluation Measures

In our experiments, we use four common different measures for the evaluation of the classification quality: Accuracy, Precision, Recall, and F-Measure [21, 22]. Measures calculated using Table II, which shows classification confusion matrix based on the Equations 1, 2, 3 and 4, respectively.

TABLE II: CONFUSION MATRIX

		Detected	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy is the proportion of the total number of predictions where correctly calculated. Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases. Also, we used the F-measure to combine the recall and precision which is considered a good indicator of the relationship between them [22].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = 2\frac{Precision * Recall}{Precision + Recall} \quad (4)$$

C. Results

After applying classification techniques on the data set, the results are distinct based on different data mining measurements. Table III presents the results using different classification algorithms (ANN, NB and DT). Each classification algorithm introduces two classification results: (1) classification results with student's behavioral features (BF) and (2) classification results without behavioral features (WBF). As shown in the Table III, we can see good classification results for different classification measures with behavioral features comparing with the results when removing behavioral features, which proves the strong relationship between student's academic achievement and the learner behavior in the classroom.

TABLE III: CLASSIFICATION METHOD RESULTS WITH BEHAVIORAL FEATURES (BF) AND RESULTS WITHOUT BEHAVIORAL FEATURES (WBF)

Evaluation Measures	**DT**		**ANN**		**NB**	
	BF	**WBF**	**BF**	**WBF**	**BF**	**WBH**
Accuracy	61.3	55.6	**73.8**	**45.8**	72.5	50.4
Recall	61.3	55.6	**73.8**	**45.9**	72.5	50.4
Precision	60.9	56.2	**73.9**	**45.2**	72.7	49.6
F-Measure	60.1	53.4	**73.2**	**44.8**	71.9	49.4

In addition, we can notice that ANN model outperform other data mining techniques. ANN gives 73.8 accuracy with BF and 55.6 without behavioral features, 73.8 means 98 of 133 students are classified correctly to the right class labels (High, Medium and Low). Recall results 73.8 with BF and 45.9 without behavioral features, 73.8 means that 98 students correctly classified to the total number of unclassified cases and correctly classified cases. Precision results 73.9 with BF and 45.2 without behavioral features, 73.9 means 98 of 133 students are classified correctly and 35 students are misclassified. F-Measure results 73.2 with BF and 44.8 without behavioral features.

Furthermore, we applied filter-based feature selection approach using information gain based to check which the most important features in the data set [18, 26]. Table IV shows the best feature set after evaluation. As shown in the above table the behavioral factors got the highest feature ranks, which means learner behavior during the educational process have an impact on student's academic success.

The experimental results prove that the strong effect of learner behavior on student's academic achievement. We can get results that are more accurate by increasing the training set, training time, classification methods, expanding data set with more distinctive attributes.

TABLE IV: FILTER FEATURE SET EVALUATION.

Feature Name	**Feature Rank**
Raised hands	0.57
Visited Resources	0.52
Announcements View	0.40
Discussion Groups	0.24
Relation	0.24

VI. CONCLUSIN AND FUTURE WORK

In this paper, we have presented a new student performance classification model that study the behavioral actions of the learner during learning process. The data was collected from an e-Learning system called Kalboard 360 using Experience API (XAPI). This is may the first time in researches to integrate student's behavior with their academic success, by applying data preprocessing techniques on data we find that behavioral factors have the good impact on students' academic success. The discovered knowledge that results by applying classification techniques obtain that learner's actions played a main role in learning process, and this is proved by the good accuracy results. The accuracy enhancement when using behavioral features are: ANN obtains 25% to 29% improvement, NB obtains 22% improvement, and DT obtains 6% to 7% improvement.

Our future work includes applying data mining techniques on an expanded data set with more distinctive attributes to get more accurate results. Also, experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbor, and others.

REFERENCES

[1] El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study', The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18 , 2008 .

[2] Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students' Performance: A Case Study, International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.

[3] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

[4] Kalboard360-E-learning system, http://cloud.kalboard360.com/User/Login#home/index/ (accessed July 31, 2015).

[5] Naser, S. A., Zaqout, I.,Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2), 221-228.

[6] Kevin Swingler "Data Mining Classification" http://www.cs.stir.ac.uk/courses/ITNP60/lectures/1%20Data%20Mining /3%20-%20Classification.pdf (accessed July 10, 2015).

[7] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[8] Burcu a.m, A path model for analyzing undergraduate students' achievement, Journal of WEI Business and Economics-December 2013 ,Volume 2 Number 3.

[9] Harris, A. (1999). Teaching and Learning in the Effective School. Aldershot: Ashgate .

[10] Concordia Online Education, "Strategies to Improve Classroom Behavior and Academic Outcomes" http://education.cu-portland.edu/blog/classroom-resources/strategies-to-improve-classroom-behavior-and-academic-outcomes/ (accessed September 19, 2015).

[11] Bradley, P. Fayyad, U. and Renia C., "Scaling EM clustering to large databases". Technical Report. Microsoft Research. 1999.

[12] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

[13] Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) 'Data Mining Model for Higher Education System', European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[14] Moisa, V. (2013). Adaptive Learning Management System. Journal of Mobile, Embedded and Distributed Systems, 5(2), 70-77.

[15] Gunuc, S., & Kuzu, A. (2015). Student engagement scale: development, reliability and validity. Assessment & Evaluation in Higher Education, 40(4), 587-610.

[16] Kuk, G. D. (2001). Assessing what really matters to student learning. Change, 33(3), 10-17.

[17] Stovall, I. (2003). Engagement and Online Learning. UIS Community of Practice for E-Learning. http://otel.uis.edu/copel/EngagementandOnlineLearning.ppt

[18] Ron Kohavi, George H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273-324.

[19] Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2006). A comparison of classification techniques for the P300 Speller. Journal of neural engineering, 3(4), 299.

[20] Arora, R., & Suman, S. (2012). Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications, 54(13), 21-25.

[21] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[22] Chen, T. Y., Kuo, F. C., & Merkel, R. (2004, September). On the statistical properties of the f-measure. In Quality Software, 2004. QSIC 2004. Proceedings. Fourth International Conference on (pp. 146-153). IEEE.

[23] Asha.G.Karegowda1, A. S. Manjunath2 & M.A.Jayaram3, Comparative study of attribute selection using gain ratio and correlation based feature selection, International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277.

[24] Blum, A.I., and Langley, P. "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol 97, 1997, 245- 271.

[25] Jasmina Novakovic, Using Information Gain Attribute Evaluation to Classify Sonar Targets, 17th Telecommunications forum TELFOR 2009, Serbia, Belgrade, November 24-26, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.