Multi-Dimensional Analysis to Predict Students' Grades in Higher Education

https://doi.org/10.3991/ijet.v14i02.9905

Eslam Abou Gamie, M. Samir Abou El-Seoud(✉), Mostafa A. Salama, Walid Hussein
British University in Egypt, Cairo, Egypt
selseoud@yahoo.com

Abstract—This work enhances the analysis of the student performance in the high education level. This model categorizes the features according to their relativeness to the teaching style and to the student activities on an Electronic Learning system. Several new features are proposed and calculated in each of these two categories/dimensions. This approach applies an extra level of machine learning that analyses the data based on a set of dimensions, and each dimensions includes a set of features. The prediction analysis is applied on each dimension separately based on a different classifiers. The best fitting classifier to each dimension ensures the enhancement of the local analysis accuracy and though enhances overall global accuracy. The accuracy of prediction of the student is enhanced to 87%. This study allows the detection of the correlation the features in different dimension. Furthermore, a study is applied on the scanned text documents for extracting and utilizing the features that represent the student uploads.

Keywords—Data mining, education data mining; MOODLE; feature selection, correlation analysis; learning activities; pedagogical approaches; classification

1 Introduction

The analysis of the interaction between the student and the electronic learning systems has increased exponentially. The prediction of the student success is a common requirement for early detection of students-at-risk, and for providing a list of recommendations to decision makers of the high education levels, like adding more attracting materials. Systems like Moodle uses the web usage mining in analyzing the performance of students in high education levels. Educational Data Mining (EDM) is the process of exploring the knowledge existing the several forms of educational data that is required in the critical decision making process. The aim of this process is to detect the relation between the performance of the student and the several inputs in the education process. The educational inputs could involve several types like the nature of the taught modules, the teaching style and different activities applied on E-Learning. Every type of these educational inputs includes a set of patterns and rules that can be used in predicting the success opportunity of the students. Several approaches are applied to reach a high accuracy in predicting the success rate of the students. define the problems that prevents the accurate generations of these and set approaching are proposed to solve these problems. The work in [1] studies the data imbalance problem, the number of failure students are usually less than the number of succeeded students. The work in [2] studies the variation of the types of the module and the variation of the student activities. Most important eLearning activities can be stated under the following topics: online assignments, online quizzes, SCORM packages, Library logs, and other resources like online labels and online files. The problem investigated in this work is enhance the accuracy of predicting the final grades of the students. The dependency on student performance only is not accuracy enough for classifying students according to their grades. The current research did not evaluate the tutor performance or the quality of the delivered contents. Students could perform badly in the module, not because the students are not diligent, but because the materials of the module via the e-learning or the teaching style are not attractive. Also students could lose the interest or feel disappointed if the resources of module are not added probably or added lately. On the other hand, every module could have a different type relative to the other modules. Where modules can be grouped according to some common characteristics, such that every category can be classified according to its existing patterns. For example, modules like “entrepreneurship”, “legal and professional issues” and “computer Interaction” are of the same category. This category is dependent more on the reading capabilities of the students and the English writing talent on expressing their ideas. While modules like “Programming”, “Algorithms” and “data structure” are of different category. This category is dependent on the mathematics and computing capabilities together with the systematic thinking talent. The patterns in both categories are different from each other, and applying data analysis and mining on each one will have a positive impact on the accuracy of classification.

2 Previous Work

The work in [5] focused on the extraction of rules from eLearning systems using rare association mining techniques (RARM). Four mining association algorithms were compared; Apriori- frequent, Apriori –Infrequent, Apriori-Inverse and finally Apriori-Rare. The paper explored applying RARM to detect infrequent student behavior, it also stated that normal association roles (like Apriori algorithms) do not take infrequent associations into consideration, despite the fact that relatively infrequent associations could be of significant interest. Three students' online activities were counted; assignments, quizzes and forums and the predicted final course grade. The paper receded roles extraction on variable types of association roles only, in addition no behavioral attributes were taken into consideration, which for sure could enhance the accuracy of predication. In [6] the author deals with variance of courses types and number of activities generated from eLearning systems, he detects the relationship between activities and resources in a certain course along with students' final grades, he did so by applying different Multiple Instance learning techniques and results were compared. Although the research is well organized the main focus was on the techniques, and not the data attributes, without mentioning the reason behind choosing only three specific students' online activities. The author in [7] started with a question if it is possible to predict student's success enrolled in a course with a small dataset? And that datasets associated with students are considered small even if with a big number of students. Student attributes considered in this paper for prediction: gender, year of birth, Employment, status, registration, type of study, Exam condition and activities.

Most of the previous work in studying the behaviour of the students over the academic period are based on several perspectives. The first perspective is partitioning the factors that are affecting the student behaviour according to the institutional and family support and degree the student awareness [8]. The second perspective studies the students who perform an improvement during their study in the university [9]. Another perspective is the addition of the external factorials like the economic status [10]. Finally, the current perspective adds, to the known factors, the interaction to the electronic educational systems [11].The current research trendsin this area examines the different activities performed by the student on the Electronic learning systems. The work in [11] studies the frequency of online interaction of the students. It studies the percentage of accessing the virtual classroom and discussion boards. The work in [12] provides an evaluation to the E-learning systems by categorizing the different factors that may affect the student performance. These factors are divided into six dimensions: system quality, service quality, content quality, learner perspective, instructor attitudes, and supportive issues. The purpose of the previous work is to gain the benefit of all factors that affect the student performance and build a machine learning model that enables the decision makers in altering the teaching methodology. None of the previous work builds a model that simulates the multiple dimensions of these factors.

3 Proposed Solution

The problem investigated in this work is to enhance the accuracy of predicting the final grades of the students. The dependency on students' eLearning logs only is not sufficient enough for classifying students according to their grades. The current researches did not evaluate the pedagogical attributes along with eLearning analysis. Students could perform badly in the module, not because they are not diligent, but because the materials of the modules via the e-learning or the teaching style are not attractive. Also students could lose the interest or feel disappointed if the resources of module are not added probably or added lately. On the other hand, every module could have a different type relative to the other modules. This work extends the current research in the area of e-learning data mining process by detecting the correlation between three general areas. The first area is the module delivery on the e-learning system, the second area is the student performance and interaction to the delivered contents, the third area is the students' performance in the final marks of the module.

This work considering the fact that the factors that affects the student performance is semantically categorized into several dimensions. Although various research is applied make use of these dimensions [12], the predication model of the student performance do not consider this important fact. Ensemble methods that combines the results of various heterogeneous classifiers is a fitting tool to guarantee the enhancement of the accuracy of utilized machine learning models [13]. Each classifier in the ensemble model could target one the dimensions of the learning factors. Which classifier to be for each dimension is dependent on the nature of the data in this dimension? In this work, an assumption is considered that all classifiers could perform equally to all the dimensions. Figure 1 provides a picture of the proposed model that distribute the classifiers used in the ensemble technique to the data sets of each dimensions. This process of distribution is applied randomly in this work, the future work that is applied here is to use an evolutionary model like genetic algorithms to perform this distribution.

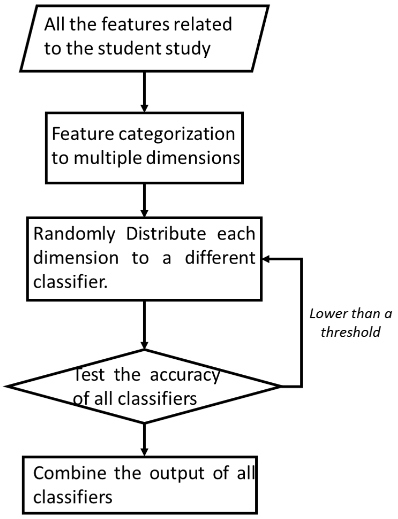
All the features related to the student study
Feature categorization to multiple dimensions
Randomly Distribute each dimension to a different classifier.
Lower than a threshold
Test the accuracy of all classifiers
Combine the output of all classifiers

Fig. 1. The proposed model for predicting the student performance based on the E-Learning log analysis.

3.1 Population and Data collection

This work considers two main dimensions that are influencing the performance of the students. These dimensions can be named as; student activities via eLearning (abbreviated as A), teaching style from pedagogical approach (abbreviated as T). Each dimension includes a set of features (factors); that describes the different characteristics of each dimension. The aim of this work is to study the effect of these factors on determining the results of the students by utilizing data mining techniques. The features and the corresponding values of each factor can be listed as follows:

E-Learning activities A: While trying to test the effect of e-learning frameworks on instructive results, this exposition endeavors to examine the instance of the local private university. A recently established university in 2005 is considered that has one of the oldest Learning Management Systems. The aggregate enrolment at the college was 9228 understudies in 2016, and the enrolment extended nearly around this level in the earlier years. The e-learning framework at the University covers all modules over all resources. Materials are transferred week by week, and chiefly incorporate PowerPoint introductions covering the sessions about that week, and an instructional exercise sheet. Each student has a specific ID number, and the e-learning system uses that ID to track students' usage, collecting data such as the number of times a student accessed a certain module, his/her enrolment date, the duration of using e-learning, and which materials the student accessed and/or downloaded. Also students could lose the interest or dispassionate if module resources are not added properly or lately. For academic purposes this data was made available, and coupled with the fact that the British University in Egypt has a well-established e-learning system, the university was chosen as the case study of this research. A total of 5 variables were selected to be included in the analysis, based on previous literature on the topic. The variables, their explanation, and previous literature in which they were used can be viewed in table 1. All variables were obtained from the British University in Egypt's e-learning server, with the help of the E-learning Department.

- Delay in enrolment to the module [0-100 days]
- Number of accessing the module and resources in the semester [0-100 times]
- Average number of accessing the module per week [0-10 times]
- Average time delay in accessing the lectures starting from the upload time[0-100 days],
- Average time delay in accessing the lectures starting from the upload time [0-100 days]
- Average time of uploading the assignment answers subtracted from deadline time [0-100 days].

Such data is collected from eLearning system log file and university database.

Teaching style T: The style of teaching presented by the instructor could not be suitable for all students. For example, some student may like the use of the marker during the lecture, while others prefer using power points presentation. This type of data will be collected from the student evaluation form provided by the students by the end of the semester. The current researches have not evaluated the tutor performance or the quality of the delivered contents along with the analysis of student data. Materials presentation is an important factor where the instructor could deploy the lectures properly, for instance if the lectures are uploaded simultaneously with labs or weekly or within the first week, or the lectures of the contents are not up to date with each session. The style of teaching presented by the instructor may not be suitable for all students. For example, some student could like the use of the marker during the lecture, while others will prefer using power points presentation. This type of data will be collected from the student evaluation form provided by the students in the end of the semester.

Module nature M: The module nature reflects the direction of the module, whether it is scientific, mathematical, programming, or theoretical module. The module specification of the module includes the detailed information about the module including the nature of the assessment like whether the assessments focus on the lab tests, projects or unseen exams. The module specification also may include the topics covered by the module, the number of hours of the labs, lectures and tutorials. The distribution of the mark also reflects the focus of the module about the theoretical contents, laboratory contents or the mathematical contents. The module reading list can be recognized according to the category of the materials in the famous websites like Amazon, or from the local library like the library of the university or the school. The module nature could reflect the interest of the student and the points of strength and weakness of each student. Students may perform badly in a module, not because the students are not diligent, but because the materials of the module via the e-learning or teaching style are not attractive. This will help the university decision makers to adjust the module according to the student needs and capabilities without ignoring the need of the industry.

3.2 Data analysis

The study is divided into two phases as shown in figure. The first phase combines the features of the two dimensions along with the student results in one data set. Then apply a set of feature selection techniques to detect the most discriminating features, the correlation among features, and patterns that infers the final performance of the student.

The second phase in this study is to construct three different data sets based on the features of the three factors: student activities, the teaching style, and the content categorization. The features of the student characteristics factor (S) are common among the three data sets. For each data set, a set of classifiers are tested to select the most appropriate one whose classification accuracy percentage is the maximum. For testing a new student, the trained classifier specific to each data set is applied to predict the final performance of the student. If two classifiers lead to a certain prediction, while the third classifier got a different result, the final prediction goes to the majority. The utilized classifiers are neural networks, decision tree, support vector machine and Bayesian belief network.

Finally, a comparison is conducted between the two phases according to the classification accuracy percentage. If the first phase shows a better accuracy percentage, this concludes that the correlation between the features of different factors is highly important in prediction of the students' final performance.

Table 1. E-Learning activities variables

Variable	Explanation
Student Grade	The final grades for each module, across the 243 students. This variable is used as a proxy for measuring educational outcomes
Number of Course Log Ins	The total number of times a student logged into a module's page on e-learning during the whole year. This variable is used as a proxy for e-learning usage
School Leaving Grade	The final high school grade for each student, in percentage terms
Module Type	Specifies whether the grade of the student belongs to a mathematical or a theoretical module
Attendance	Measures the overall attendance level of students by specifying whether the mandatory attendance policy existed

The summary statistics of the data can be viewed in table 2 below. The statistics show that the total number of observations is not constant across all variables, due to the fact that some data is missing across some of the variables, however this is not a problem, since the used software STATA is designed to automatically drop missing data. The total number of observations ranges between 3455 and 3000. The mean of the student grade variable is 53, which is equal to a C in the grading scale of the British University in Egypt. The values of the grades have a very wide range between 0 and 98, given that the highest possible grade is 100. The average number of course log-in's is close to that of the average of the grade, standing at nearly 49 times. However, the standard deviation differs greatly being 37, showing that the number of course log-in's is less concentrated around the mean. This can be due to the fact that the data range of the number of log-in's is wider, ranging between 1 and 379. The mean of the school leaving grade is significantly higher than that of the student grade, at 80% suggesting that students tend to perform better in high school than in university. The mean of the school leaving grade is much smaller than the others, showing that the majority of the data centers around the high mean, suggesting that the majority of the students scored relatively high grades. The lowest reported grade is 63% and the highest is 105%.

Table 2. Summary statistics

Variable	Observation	Mean	Standard deviation	Min	Max
Student grade	3,092	53.00356	14.80982	0	98
Number of course log ins	3,000	48.88733	37.55451	1	379
School leaving grade	3,247	80.58451	8.959488	62.9	104.9
Module type	3,455	1.449493	.4975146	1	2

4 Result Analysis

This work studies the factors that affects the performance of students in the last year. The experimental work address a collective data set related to high education students. The collective data set includes all the subclasses/dimensions that categorizes these factor into four main dimensions. The student, model, teaching style and student activity on E-Learning represents the categories of all the factors that may affect the student performance.

The first step in this study it find out the main factors that are highly discriminative to the student rank classification. The statistical correlation between the ranking feature and the rest of the factors are measured as follows.

```
Ranked attributes:
0.23085   5 School leaving Grade
0.07562   4 No of Log ins
0.01428   1 Faculty
0.01317   3 Module
0.00753   6 Module Type
0.001    2 Cohort
  0.001    7 Attendance
```

It appears that the school leaving grades possess the highest correlation to the ranking feature, followed by the number of log in to the educational system [E-learning]. On the other hand, the student attendance appears to be of the lowest discriminating effect. The School leaving grades reflect the dimension of the student characteristics and original behaviour. While the no of log in reflects the degree of the student interaction related to the E-Learning division. This proves that a single dimension is not enough to reach an accurate prediction of the student performance. And provides an evidence that the data analysis of this field must compromise all the factors but with putting into consideration that these factors are categorized.

The second step is the classification of the data features against the ranking feature. The classifier that shows the highest classification accuracy is the Naïve Bayesian network. Baysian models consider the univariate model of the input data set. This behavior shows that the data attributes are gathered from the different resources, so the dependency or the correlation between these attributes are decreases to minimum. Otherwise another technique like neural network, support vector machine or even the evolutionary algorithms. The accuracy percentage of this classifier appears as follows:

```
Correctly Classified Instances       2647               87.0724 %
Incorrectly Classified Instances      393               12.9276 %
```

The accuracy of the classifier is calculated based on the number of instances predicted correctly to the total number of instances. The number of instances are equally categorized to ensure the fairness of the classification process. The confusion matrix of the classifier is as follows:

```
=== Confusion Matrix ===
   a    b    c   <-- classified as
 702   29   16 |    a = Top 25%
  88 1293  132 |    b = Middle 50%
  28  100  652 |    c = Bottom 25%
```

The confusion matrix shows that the error rate in each category, where the error rate in the middle category is the highest. This is because the attribute values lies between the top and bottom class and this increases the confusion of the applied classifier. When the features Faculty, cohort, module and module type features are removed, the classification accuracy appears to be the same. When removing the No of Log ins feature, the classification accuracy decreases to 86.25%, while when removing the high School leaving Grade, the accuracy is decreased to 49.76%. This provides an evidence that the student ranking is dependent mainly on this evaluation before joining the high education stage.

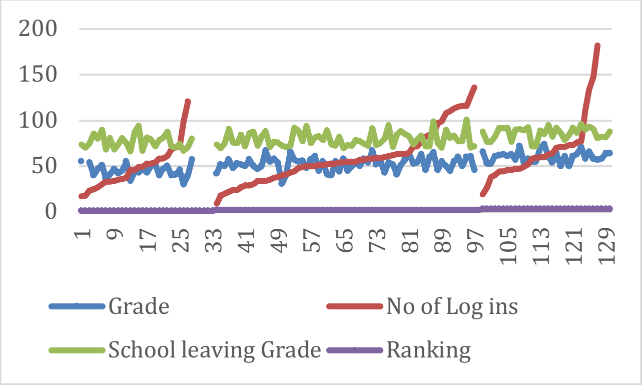

Fig. 2. The proposed model for predicting the student performance based on the E-Learning log analysis.

Figure 2 presents the values of the current grades of the students and the corresponding grades in high school and the number of login activities to the E-Learning system. These values are distributed over three ranking values (Top, Medium, and Bottom ranking). This chart shows that the high school grades are always higher than the current grades of the students. A low correlation appears between the number of login activities and the current and high school grades.

5 Extended Results for Integration

Another dimension that could have a great contribution to the Educational systems is the handwritten document detection [14,15]. The documents scanned and uploaded by students are saved in the Educational system as black-box where no use of the contents of these documents. The detection of the text could help on automatic marking of the contents and recording the results by the student records. This ensure online evaluation of the uploaded document, providing an appropriate feedback to each student separately.

Here in this work, the Line detachment regularly is put off to the division step. Division calculations endeavor to part an archive into pieces: pages into lines, lines into words, words into characters. These calculations create hopeful districts for recognition. Each word is validated by discovering its standard and turn it on its focus of gravity so that the standard winds up noticeably even. The resulted text is compared to the model answer and a reasoned-mark is resulted automatically online to the students. The resulted mark is a feature, other several features could be detected like the time submitting the answers, the handwriting clarity, and the answer organization.

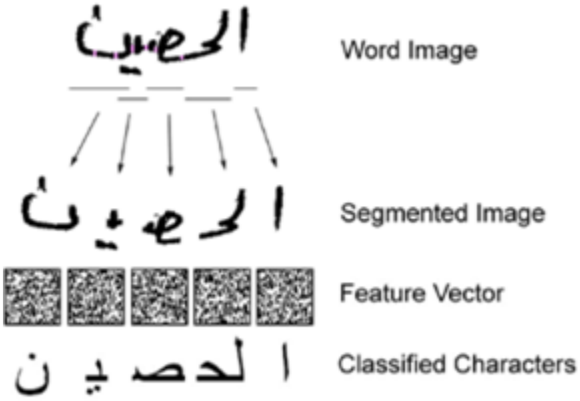

Fig. 3. A sample for the detection of an image of a handwritten word in Arabic.

A general accuracy of about 95 % was found with this technique on an inside gathered dataset. Preparing was done on 50 words each of manually written and machine printed pictures, and the testing information comprised of reports containing an aggregate of 286 machine printed and 104 manually written words.

6 Conclusion and Future Work

This work extends the current researches in the area of learning analytics and data mining process by detecting the correlation between three general dimensions; the first dimension is the student activities via eLearning; the second dimension is teaching style and finally student result. Data gathering phase is done for eLearning dimension and ready for analysis, Future work will include analysis of these dimensions and their results, then other dimensions like demographic data and pedagogical attributes from open linked data will be included for further accuracy results.

7 References

[1] Koichiro Ishikawa, M. F. (2013, December 6). Log Data Analysis of Learning Histories in an e-Learning. International Journal of Information and Education Technology.

[2] Beth Dietz-Uhler, J. E. (2013, spring). Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. Journal of Interactive Online Learning.

[3] Nor Bahiah Hj Ahmad, S. M. (2010). A Comparative Analysis of Mining Techniques for Automatic Detection of Student's Learning Style. 10th International Conference on Intelligent Systems Design and Applications (pp. 877-882).

[4] Fatos Xhafa, S. C. (2011). Using Massive Processing and Mining for Modelling and Decision Making in Online Learning Systems. 2011 International Conference on Emerging Intelligent Data and Web Technologies, (pp. 94-98).https://doi.org/10.1109/EIDWT.2011.22

[5] Romero, C. R. (2010). Mining rare association rules from e-learning data. Proceedings of 3rd International Conference on Educational Data Mining, International Educational Data Mining Society, (pp. 171–180). Pittsburgh.

[6] Ventura, A. Z. (2009). Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming. Proceedings of the 2nd International Conference on Educational Data Mining, (pp. 307-314). Cordoba.

[7] Srecˇko Natek, M. Z. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications.

[8] KHURSHID, FAUZIA. (2014) Factors Affecting Higher Education Students' Success, Asia Pacific Journal of Education, Arts and Sciences, Vol. 1, No. 5, November 2014.

[9] Hijazi, S. T., & Naqvi, S. M. M. (2006). FACTORS AFFECTING STUDENTS'PERFORMANCE. Bangladesh e-journal of Sociology, 3(1).

[10] Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: a case of secondary school level. Journal of quality and technology management, 7(2), 1-14.

[11] Davies, J., & Graff, M. (2005). Performance in e-learning: online participation and student grades. British Journal of Educational Technology, 36(4), 657-663. https://doi.org/10.1111/j.1467-8535.2005.00542.x

[12] Ozkan, S., & Koseler, R. (2009). Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. Computers & Education, 53(4), 1285-1296. https://doi.org/10.1016/j.compedu.2009.06.011

[13] Whalen, S., & Pandey, G. (2013, December). A comparative analysis of ensemble classifiers: case studies in genomics. In Data Mining (ICDM), 2013 IEEE 13th International Conference on (pp. 807-816). IEEE.

[14] Topaloglu, M., & Ekmekci, S. (2017). Gender detection and identifying one's handwriting with handwriting analysis. Expert Systems With Applications, 79, 236-243. http://doi.org/10.1016/j.eswa.2017.03.001

[15] Rao, Z., Zeng, C., Wu, M., Wang, Z., Zhao, N., Liu, M., & Wan, X. (2018). Research on a handwritten character recognition algorithm based on an extended nonlinear kernel residual network. KSII Transactions on Internet & Information Systems, 12(1).

8 Authors

Eslam Abou Gamie is master student in the British University in Egypt, in Web sciences Master Program. His Email account is eslam.gamie@bue.edu.eg.

M. Samir Abou El-Seoud Professor Samir Abou El-Seoud received his BSc degree in Physics, Electronics and Mathematics from Cairo University in 1967, his Higher Diplom in Computing from Technical University of Darmstadt (TUD) /Germany in 1975 and his Doctor of Science from the same University (TUD) in 1979. His Email account is samir.elseoud@bue.edu.eg

Mostafa A. Salama Dr. Mostafa is an associate professor in the British University in Egypt since 2018, and a lecturer since 2011. His teaching experience since 2007 is following the UK quality and validation standards. He worked in industry for 8 years in embedded system and business administration projects. His M.Sc. is related to securing the electronic cash and hi Ph.D. in works on enhancing the mining techniques required in medical informatics. His Email account is mostafa.salama@bue.edu.eg.

Walid Hussein is a lecturer in Department of Computer Science, Faculty of Informatics and Computer Science, The British University in Egypt, Cairo, Egypt, since 2014. His research concentrates on develop, analysis, and implementation of these provably good numerical methods. Throughout my studies, he have worked intensively in solving pattern recognition problems using the advances in Signal/Image Processing and Artificial Intelligence, as non-destructive monitoring and testing approaches.. His Email account is walid.hussein@bue.edu.eg.

Article submitted 23 November 2018. Final acceptance 27 December 2018. Final version published as submitted by the authors.