Predicting Grades by Principal Component Analysis

A Data Mining Approach to Learning Analyics

Alvaro Figueira
CRACS / INESCT TEC & University of Porto
Rua do Campo Alegre, 1021/1055
Porto, Portugal
arf@dcc.fc.up.pt

***Abstract*—In this paper we introduce three main features extracted from Moodle logs in order to be uses a possible means to predict future student grades. We discuss the statistical analysis on these features and show how they cannot be applied isolatedly to model our data. We then apply them as a whole and use principal component analysis to derive a decision tree based on the features. With derived tree we are able to predict grades in three intervals, namely to predict failures. Our proposed analysis methodology can be incorporated in an LMS and be used during a course. As the course unfolds, the system can to trigger alarms regarding possible failure situations.**

Keywords- Grade prediction, Data mining, Feature selection, Decision tree, Moodle logs.

I. INTRODUCTION

Predicting future grades for students, particularly when done in early stages, has been a recent concern in the academia. as it can be used to give important advises to students based on "lessons" learned from past experiences. Recent research on learning analytics has taken different approaches for that goal [1,2].

In this work we present an approach which is based on the analysis of the Moodle logs (similarly to the method adopted in [3]). This analysis is consolidated through a case study of a higher education course with the participation of more than 300 students. In the course the students had three small tests and the activities of writing part of an article, assessing a couple of articles, creating slides and presenting them orally.

Our motivation for this research is to have an expert system based on previous experiences that is able to trigger alarms whenever the systems detects, with a high percentage of confidence, that a student is following a path will lead him to a failure, so that he/she can be helped to mitigate and solve the detected problem as early as possible, as proposed in [4].

We formulate our research hypothesis as: is it possible to use the Moodle logs in order to trigger such alarms?

A. *Proposed approach*

Our methodology is to use of feature selection and a decision tree, based on the integrated analysis for a set of three features, extracted and computed from the retrieved Moodle logs [5].

The first feature is the number of accesses to the platform (activities or resources) from each student along the semester. The second feature is the percentage of coverage of the whole course online activities and given lecture notes which are stored in the Moodle platform. As a third feature we use the percentage of correct sequences of accesses to the pedagogical material made available.

B. *Case study and test data*

To assess our strategy, we used as a case study, a course from a higher education Bachelor's degree in Computer Science. The course had 332 enrolled students and lasted for one semester. The main goal of this course is to prepare students for "technical communication", namely in the areas of writing and assessing articles (or reports), in creating electronic presentations and in performing the oral presentation. Each lecture up to final presentations is followed by a small online quiz. The slides and the presentation are assessed manually. The articles and their assessment is done using the Moodle activity, the Workshop.

C. *Article structure*

In the next section we present the type of data we will be dealing with and what type of data preparation we had to adopt. In section III we present in detail the type of analysis we performed in order to have normalizes data as features of each students so that they can be compared. In section IV we confront these features with the final grade of each student and we build a decision tree in order to make predictions. Finally, in the last section we present our conclusions.

II. DATA RETRIEVAL PREPARATION

Our strategy is to use three features to classify student's interactions in order to predict grades, but, ultimately to prevent student failures in the course.

A. *The Moodle logs*

We use Moodle logs to access information regarding daily usage, action performed, resources used and sequence of using these resources. We retrieved the full log from February2015 until June2015. During this period there taken more than 55K student interactions with the platform.

From the logs we understood that from the initial 332 students enrolled in the course, only 311 interacted with the platform.

We used the logs to extract features to validate our hypothesis, as described below:

B. *Feature 1: number of accesses*

Intuitively, we would expect that the more interactions the more dedication and interest on the subject would lead to better grades. On the other hand, it is also fair to expect that students with more interactions are the ones with more difficulties. Therefore, we wanted to analyze this feature and understand is there is some correlation between the number of interactions and the final grade.

C. *Feature 2: coverage of digitally provided learning material*

A second feature would be to know if the students have covered all coursework, available resources and proposed activities in the platform. Intuitively we expect every student to access every lecture handout, all the provided slides enrolled in every proposed activity.

However, from an analysis of the logs, we discovered that this is not true. In fact, 76% is the average of accessed material in the platform. Therefore, we wanted to understand the impact that this factor may have in the final grade.

To reach this goal we created a matrix where each corresponded to a student and each column to the number of times a specific online resource/activity was accessed. Curiously, in this matrix we found many students with multiple accesses to the same activity, even if it is a single-access resource, like the handouts of one lesson (we had one student with 25 accesses to that file, another one checked his current grades for 48 times).

Therefore, we also wanted to study of the effect of this feature in the final grade.

D. *Feature 3: percent of correct sequences*

Lastly, we grouped all course mandatory activities into general activities as we list in Table 1. As we gathered actions into general groups, we analyzed the interaction graph which shows the number of interactions of each type along the semester (Figure 1). As we can see, in this case study, students spend most of the interactions viewing resources.

However, we must focus on the peaks of the graph and model that behavior according to the sequence in which actions are being made. For example, it's common for a student to skip seeing/reading some lecture slides until the very last moment before the test. Or, eventually, he skept some handouts, and later on had some regret for doing so, and tried to recover wasted time by accessing them. All in all, and to summarize these type of behaviors, what we propose as a subject of concern is that the sequence in which activities occur is not independent from any permutation of sub-set of this "normal" sequence. We stress that we are not forcing a specific behavior, but just detecting ill-situations.

For this purpose, we serialized all activities and resources made online available to students and marked the time in which it is profitable to use them. We mean profitable as having the contents, or the methodology to apply in the evaluation activities.

Table 1. Actions grouping.

View	**Submit**	**Configuration**
assign view	assign submit	calendar edit
assign view submit assignment form	choice choose	course update mod
choice view	quiz attempt	quiz editquestions
choicegroup view	quiz close attempt	quiz update
course view	workshop add assessment	**User view**
page view	workshop add submission	user view
quiz view	**Continuations**	user view all
quiz view summary	choicegroup choose again	
resource view	choicegroup remove choice	
sigarracourseinfo view	quiz continue attempt	
workshop view	workshop update assessment	
workshop view submission	workshop update submission	

We made a list of 17 different materials and did a partial order on that set. Then, we marked every interaction record for every student with the corresponding serial. Finally, we compared each student sequence with our *golden standard*. For the sake of objectivity, we present an example: the golden standard is: 1, 2, 3, 4, …, 16, 17. But, for student X, the sequence is: 2,12,2,16,12,17,8,4,7,5,4,3. It should be clear that the student's sequence may not be of the same length as the *golden standard*.

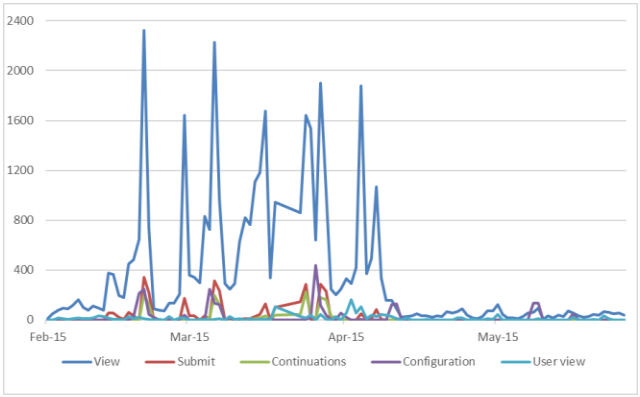

Figure 1. Action types along the semester.

We then created a function which computes the differences between the two sequences. This function is based on the Hamilton distance, as it counts the number of changes that have to be made in student's sequence in order to obtain the golden standard. We call this value the "Ordering Degree", which in use as the third feature of the analysis.

III. DATA ANALYSIS

In this section we present our preliminary analysis of each feature as seen independently from the others.

A. *Number of accesses to the platform*

We computed the correlation between the number of times one student accesses online material and the final grade he had. We found that considering all student we didn't find a clear correlation (R^2 = 0.27, which is low). When we considered only the approvals, we got R^2 = 0.23 (which is also low). However, when we consider only the failures we got R^2 = 0.56 which is a meaningful correlation, as depicted in figure 2:

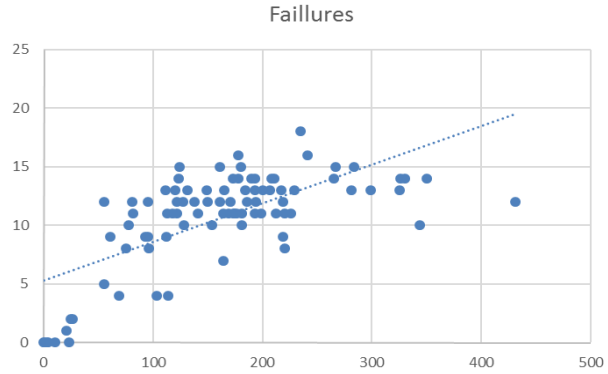

Figure 2. Accesses vs final grade.

Values of correlation increase to more than 0.65 as we switch to polynomial regression of order higher than 5.

B. *Coverage of the provided online material*

From the 14 items that each student, in theory, would need to access, only two were accessed on average only once (which would be enough for any of these 14 items). However, if we count the number of times each student does not access any of these 14 items, we compute the average of all students, we got the number 3.32 which means that, on average, each student does not access more than 3 of needed items to successfully complete the course. However, the average of covered material is 76%, and 60 out of 311 students had accessed all materials.

C. *Ordering distance*

Whilst it is fair to understand that the bigger this number, the more distant is a student sequence, to the golden standard, ie, to the natural correct sequence, this metric has an inherent problem: the shorter the sequence, smaller would be the changes to convert it to the correct sequence. Therefore, this metric can be erroneous for students with short sequences. That is, students that do not access many online resources. Nevertheless, in this case study, as the resources are regularly accessed more than once, the former cases are clearly outliers.

IV. Feature Selection and Decision Tree

As it has been discussed in the previous section, the chosen features provide meaningful descriptions of data, and particularly of student behavior. However, when assessed in an isolated way, we may find many problems. Hence, in this section we discuss the integration of these three features, trying to select the most important ones to model our case study. For this process we build a matrix with lines as records of student performance in each of the three features. Each column is one feature, and the last column is filled as the final grade for the corresponding student. We use the `rpart` function, from the Recursive Partitioning and Regression Trees library, of that R package. We then compare the final grade to each of the three features to obtain the classification tree (to predict grades) depicted in Figure 3.

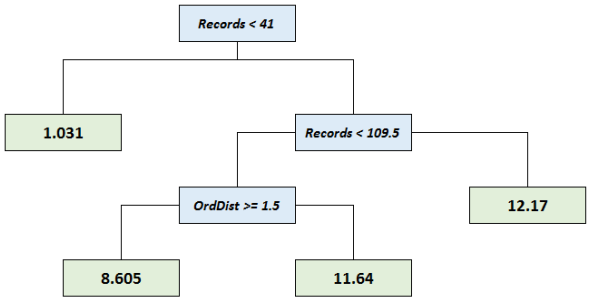

Figure 3. Decision tree.

V. Conclusions

In this study we experimented a new approach to predict grading and prevent students from failures at early stages. The model we described can be implemented within Moodle, but it can be generalized to any system with a fairly reasonable log system. In this article we discussed three types of features, that can be extracted from the logs to characterize the interaction behavior with the platform, presented an approach to data mining Moodle logs using principal component analysis to detect the relevant features to make predictions.

Acknowledgments

This work is supported by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project «Reminds/ UTAP-ICDT/EEI-CTP/0022/2014».

References

[1] Romero, C., Espejo, P. G., Zafra, A., Romero, J. R. and Ventura, S. (2013), Web usage mining for predicting final marks of students that use Moodle courses. Comput. Appl. Eng. Educ., 21: 135–146. doi: 10.1002/cae.20456

[2] Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V. and Loumos, V. (2009), Early and dynamic student achievement prediction in e-learning courses using neural networks. J. Am. Soc. Inf. Sci., 60: 372–380. doi: 10.1002/asi.20970.

[3] Figueira, A, (2014). Predicting results from interaction patterns during online group work. In proceedings of IEEE Frontiers in Education Conference. 414 - 419, 2015.

[4] Pike, G. and Saupe, J. (2004), Does High School Matter? An Analysis of Three Methods of Predicting First-Year Grades. Research in Higher Education, vol **43**(2), 187-207. Kluwer Academic Publishers. Doi: 10.1023/A:1014419724092

[5] Romero, C, Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, vol.**40**(6). 601-618. Doi: 10.1109/TSMCC.2010.2053532.