Knowledge Discovery Through Time Series Applied to Students' Grades

Joaquim Assunção[1], **Fernando L. Oliveira**[2], **Claiton M. Correa**[2]

[1]Department of Applied Computing – UFSM
Santa Maria – RS – Brazil

[2]Lardev Research Group – Instituto Federal Farroupilha (IFFar)
São Borja – RS – Brazil

joaquim@inf.ufsm.br, {fernando.oliveira,claiton.correa}@iffarroupilha.edu.br

Abstract. *Assessment is a constant activity in education, in the school system, and the teaching-learning process. The traditional approach classifies the students learning level through grades. This paper shows an application of knowledge discovery and data mining through classification and clustering via time series modeling on students' grades from high school. We collected historical data from an institute of technology, from this, we created models that can be used to extract patterns to help teachers to understand the profile of the students and provide early warns about possible poor results.*

1. Introduction

The Brazilian school system, in general, uses the quantitative evaluation process, which consists of a system of grades that classifies student learning after the application of some evaluation instrument [Luckesi 2014]. Assessment is an activity that does not allow neutrality which is a process of an ethical nature that considers values and social implications [Richmond et al. 2019]. Accordingly, any activity that can assist the teacher as an educator becomes valuable for understanding the teaching-learning process.

Federal Institutes (IFs) are an educational institution that comprises different levels of education. They vary from high school to undergraduate programs. Although the IFs have diverse campuses, with several graduate programs, their primary focus is on the high school education boosted by integrated technical courses. Graduation courses are also broadly offered, and the teaching staff is usually the same for high school, graduation, and undergraduate courses. Therefore, a professor can be the adviser for a graduate student, as well as be teaching in the high-school. Hence, they have to evaluate dozens of technical Course Conclusion Papers per year.

Technical courses in computer science, usually, are focused on the creation of professionals able to understand and create primary products for the internet. In order to graduate, the students are required to create software as their final work. Regarding the required software, the technology used is often the same; PHP, Bootstrap, HTML, CSS, and MySQL database.

Regardless of the level of study, we would like to identify which are the disciplines that most collaborate in the formation of the student, in particular, those that directly impact on Course Conclusion Papers. Through classification algorithms, we investigated the current technical curriculum in order to get the disciplines that have a bigger impact on the students' final work. Through time series representations and clustering, we created groups of students, thus extract similarities in their grades, *i.e.,* which disciplines are unexpected correlated, both for lower or higher grades.

The rest of this paper is organized as follows. Section 2 details some related work. Section 3 presents some aspects of the time series. Section 4 presents the methodology. Section 5 shows the process of data mining, do extract data, and finally, Section 6 concludes the paper and shows some future works.

2. Related Work

Using data mining to predict students' performance, or detect causes of poor performance, is not a novel approach. Many papers show a similar methodology to predict bad grades or try to avoid problems, such as a high rate of evasion.

[Abu Tair and El-Halees 2012] collected five-years-period data from the College of Science and Technology Khanyounis. The primary purpose of the work was to investigate the domain of data mining in an educational environment. The work used some of the most common group of techniques to discover knowledge from databases; association rules, classification, and clustering. The authors showed their discoveries as well as the advantages of each Data Mining technique, for instance, using classification rules, they reached a 71% accuracy to detect the attributes that influence the category of the target class. Thus, it showed how useful data mining could improve graduate student's performance.

In a similar work, [Osmanbegović and Suljić 2012] developed a model for a decision support system in higher education. Different techniques of data mining were used in order to identify the most accurate and comprehensible approach among different supervised algorithms and methods of classification. The data were collected through questionnaires conducted during the summer semester at the University of Tuzla. These questionnaires had questions about socio-demographic variables, achieved results from high school, from the entrance exam, and attitudes towards that can affect the success.

Others, such as [Othman et al. 2016] focus on traditional clustering techniques for large amounts of data, such as data from MOOCs. Alternative approaches, such as [de Paula Santos et al. 2016], uses sentiment Analysis to identify positive and negative teaching practices, which can generate a complementary model for better pedagogical practices through data mining.

We performed two actions: *(i)* collect years of student data, and *(ii)* create a model for decision support (See Section 4), an approach that is fashion similar to the ones adopted by Abu Tair [Abu Tair and El-Halees 2012] and Osmanbegović [Osmanbegović and Suljić 2012]. Furthermore, our work used traditional classification learners such as Random Forests and Naive Bayes.

However, our methodology differs from the others concerning the use of time series clustering, which is a straightforward approach for temporal data, unlike the traditional clustering algorithms (e.g., k- means, k-medoids, DBScan). Therefore, we use time series to represent data as well as weight and measure distances between the students' grades in order to obtain profiles regarding the student's performance.

3. Time Series Background

Time series have been long used as a mean for data representation and data mining. It is useful to reduce data dimensionality and measure the distance between objects. We can use time series for representation or as a pseudo-series as a way to store and mining unstructured data [Keogh et al. 2006].

There are many methods to represent and measure time series distances; however, only a few have efficient representations for index and mining data using different approaches. Among these methods, we can emphasize Symbolic Aggregate approximation, SAX [Lin et al. 2003], which is efficient to adapt data, generate, and lower bounding symbolic representations.

In this work, we used SAX as a representation method in order to index Students Grades and then perform clustering to extract students profile. However, first, we obtained the essential features from the dataset to reduce data dimensionality and get weights for each variable in order to generate more precise results. Tree-based models give the set of weights.

4. Methodology

The data used for the analysis were retrieved from a sample of four years, ranging from 2014 to 2018, extracted and loaded into a Business Intelligence tool. The scores of 142 students were analyzed, which recorded a total of 4.430 grades in different subjects throughout their technical courses. The information was contained in files in PDF format. Therefore, the first step was to extract, transform, and normalize the records (Figure 1).

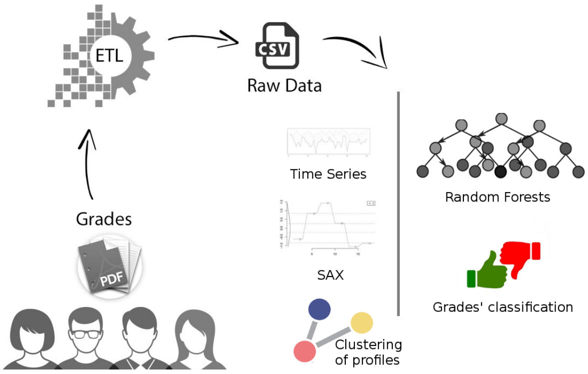

Figure 1. Classifying and clustering students' profile. The classification model aims to alert for a possible ideal or unsatisfactory grade (ideal ¿ 8.5). The clustering model seeks to identify common characteristics among groups with a specific range in the grades of their final work.

Besides, it was essential to adjust the data points with their respective weights. We used the five more essential variables, based on a tree-based model, to adjust the series (Section 4.1). Then, we proceed to transform the series with Piecewise Aggregate Approximation (PAA), which is a step to get a SAX representation.

4.1. Feature Extractions

To get the essential variables to our model, we used a boosting tree-based model being evaluated by the negative log-likelihood and validated by a test set. In near 84 iterations, we got the critical coefficients with a log-loss of 0.31, which partially answer the question, "what are the most relevant disciplines for our final work?". The result is the following:

- Development III (0.4)
- Operational Systems (0.4)
- Computer Networks (0.08)
- Databases (0.03)
- Development I (0.02)
- Others (0.07)

The same test was performed with the ten most important disciplines; however, the $6th$ already got a very small percentage; therefore, we decided to stay with the best five.

After having these values, we proceed to use the weights as a boost for modeling the time series. Thus, non-important variables cannot interfere negatively with the results. For instance, in a time series clustering, a distance between two series defines, and the set of distances determine the clusters. However, if all variables have the same weight, it would incorrectly classify students due to disciplines that have no impact on the final work. It is expected that technical disciplines, such as programming and databases, have a significant impact on the development of software, as well as humanities have lower or no impact on the development.

The question to be answered is, "based on the course history and the final grades, how important is each technical discipline?". As shown in this Section, the answer is not apparent and, in some cases, might be surprising. For instance, "computer networks" is not used in most of the works, yet it stands with a high correlation with good grades.

5. Data Mining

Once we collected the essential features for our analysis, we scattered the dataset into two different formats: i) a proper set for classification aiming the use of random forests, and; ii) a pseudo time series set for clustering through distance measurements. Figure 1 illustrates the process.

5.1. Classification Analysis

This analysis aims to identify the natural conditions for a student to have a good final work. More specifically, the cut was 8.5, which is the average grade for those approved. In this dataset, we created a binary class attribute based on the final work evaluation. So, the aim is to identify patterns that lead to grades higher than 8.5.

We used Random Forests and Naive Bayes for the classification set. As expected, the models with Naive Bayes got a more reduced accuracy compared with the ones using Random Forests. Nonetheless, we still can use Naive Bayes' outputs in order to make inferences. For instance, a Naive Bayes using all disciplines as working data, and 25% split for a test set, retrieved 100% correct on the cases where the student will have a grade higher than 8.5 in its final work and 55% correct on the cases where the student will have a grade lower than 8.5 in its last work. Thus, despite its poor performance for the false negatives, we still can use true positives as an indication of good performance.

We achieved better results using Random Forests as a default miner. Using our importance formula, and 500 trees, we got an estimate of the error rate of 19.5%. The confusion matrix is exposed to Table 1.

Table 1. Confusion Matrix using Random Forest as classifier.

	A	B	Class error
A	22	4	0.15
B	4	11	0.27

The strongest rule for a grade higher than 8.5 is the following, *"A student with more than 8.1 in computer networks and more than 7 in computer programming I"*. The other path is, *"A student with less than 8.2 in computer networks and more than 8 in computer programming III"*.

The results, regarding computer programming, are expected, since it is the bases for all the final works, and almost all of them use the same technology and templates given in the discipline. However, computer networks is an unexpected result, since it is, usually, not directly related to any final work.

5.2. Clustering through time series

After getting the decision tree, we proceeded to create clusters of students based on their final grades. This approach can help us to better understand the student's profile, as well as to direct attention for groups that will probably have a poor performance in the final work.

Figure 2 shows a process, previously proposed, to extract knowledge from stochastic models. In this case, we use part of the process and generate measurements instead of creating a stochastic model. The output is three clusters based on the students' grades in their final work.

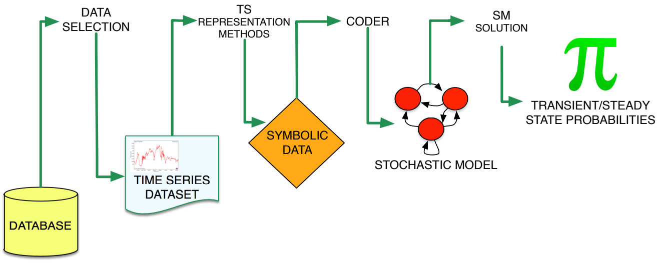

Figure 2. A model to forecast events from stochastic models [Assunção et al. 2014] partially applied in mining students grades.

Figure 3 shows the students' profile measured against each other. Each time series represents an average student from that category. For instance, the time series A given by the mean of all points regarding the students that had more than 9 in their final grade. The series is then measured against each other using a simple Euclidean Distance.

Figure 3 also shows a Dynamic Time Warping (DTW) alignment, which is another measurement considered. However, since the variables are independent, the DTW alignment is not appropriated. Therefore, we consider the ordinary Euclidean distance as the most appropriated measure.

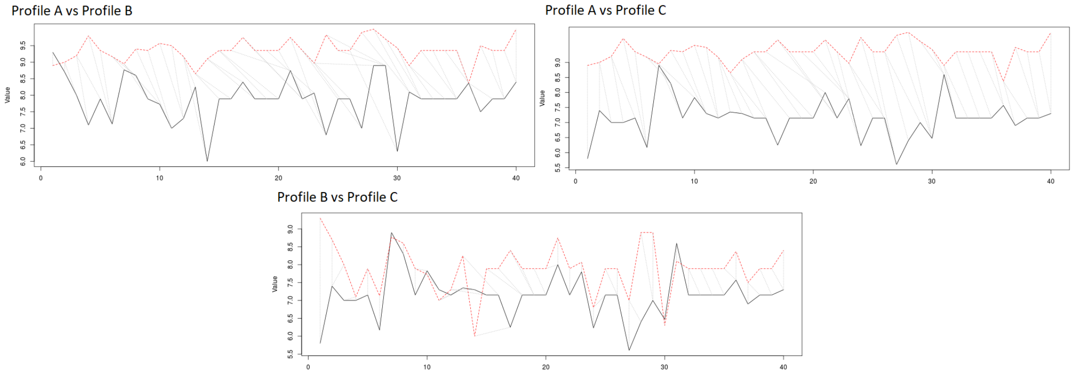

Figure 3. Students' profile measured by category (disciplines vs grades). a. shows an average student A measured with an average student B. b. shows a student A measured against an average student C. c. shows a student B measured against an average student C.

Figure 4 shows three clusters according to the students' profile. Profile C is characterized by two or three purple pixels. Profile B is characterized by four to five blue pixels, and profile A is characterized by one purple pixel.

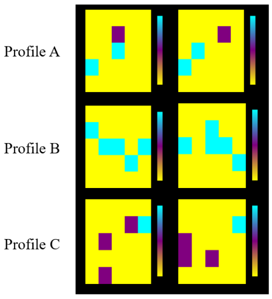

Figure 4. Clusters showing the students' profile according to their grades and the weights assigned by the data mining process. Profile A, B, and C are for students with grades, respectively, above 9, between 8 and 9, and less than 7.9.

A bitmap, extracted from the time series measurements, is formed by 25 pixels describing the grades' distribution, blue shows the maximum frequency, and yellow the lowest. Hence, the blue and yellow pixels always will have one or more elements. For instance, a student that had the following, weighted, grades: $6, 7, 7, 9$, would generate a four pixels bitmap with one yellow (6), one blue (9), and two purple pixels (7,7). In this case, grades with lower weight were interpolated in order to make one measurement. Thus the 40 records became 25 pixels.

Although these colors do not mean a universal value, they still are representatives of the variation. For instance, yellow in C class is lower than yellow in A class; however, by comparing the number of blue and yellows, we can indicate how much the student tends to vary in its grades. Looking at the profiles seems that Bs are the ones with a higher variation. As, on the other hand, tend to have a more specific variation, average grades on the humanities, and very good grades on the specifics. Cs leads to have poor grades in almost all disciplines with a few exceptions.

Finally, we performed a correlation analysis to see how similar the profiles are. Figure 5 shows the correlation between the three profiles. Here we can see that students that have the highest grades in their final work are close to the ones that have good grades. Both A and B are almost equally good students; however, their difference in technical classes are determinant to their final performance. Our best hypothesis, in this case, is that the ones that are more likely to pursue graduation in computer science are the ones that do better in their final work. Profile C is different, and they tend to have average and bad grades at all classes, with a few exceptions in the humanities.

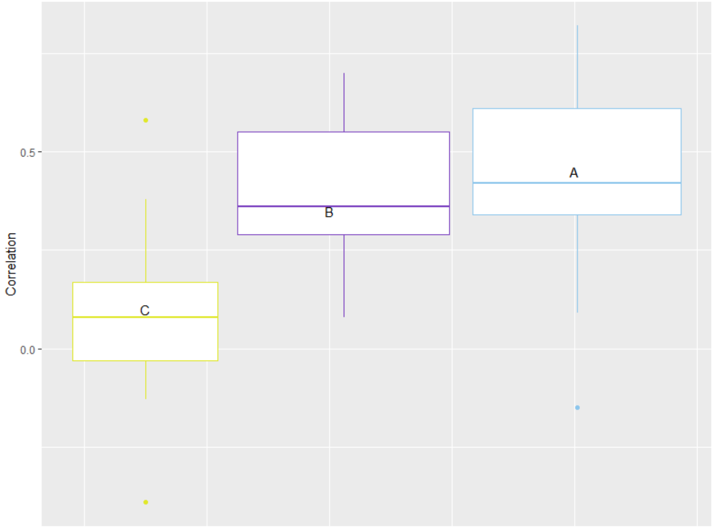

Figure 5. Correlation of an average A profile against all the others.

6. Final Remarks

This work used a knowledge discovery process to investigate the main curricular components that directly influence the students' final work results. Although it may seem obvious that some disciplines, such as programming and databases, are essential, it is vital to analyze all items in the curriculum matrix and identify which are impacting the work. Based on the grades obtained in each subject, we used Naive Bayes and Random Forests to generate common patterns for students that have higher grades in their final work. The results pointed out to some disciplines not directly related to the development of software, such as operating systems and networks. The reason for such a phenomenon is unknown and reserved for future work.

Through the measurement and correlation of time series, we generated clusters to identify groups of students regarding their performance in their final work. However, many other variables should be considered, such as advisers and evaluators, because they have slightly different criteria to evaluate work. From an educational point of view, there are many things to be discussed and analyzed; however, the knowledge obtained with this work might be part of important decisions.

We are aware that other circumstances, such as different teachers for the same discipline, can have a strong influence on the students' grades. Nonetheless, we have powerful techniques in the big domain of machine learning and data mining, and even a small improvement is worth to be tested since it can help to improve the performance of many students.

As future work, we intend to use metrics from the produced software to get a better model. We hypothesize that the grades have a strong positive correlation with the size and complexity of the produced software since the document is merely a technical report. Depending on the correlation obtained, a model can be used as a tool for helping in the evaluation, which can be used to save the time of the professors involved. If a model is accurate enough, it might be a replacement for one member, which would drastically decrease the time in which the professors are allocated for technical works, allowing them to focus on graduate works.

References

Abu Tair, M. M. and El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*, 2(2).

Assunção, J., Fernandes, P., Lopes, L., and Normey, S. (2014). A dimensionality reduction process to forecast events through stochastic models. In *The 26th International Conference on Software Engineering & Knowledge Engineering, SEKE 2014*, pages 534–539. ISBN-13: 978-1-891706-35-7.

de Paula Santos, F., Lechugo, C. P., and Silveira-Mackenzie, I. F. (2016). "speak well" or "complain" about your teacher: A contribution of education data mining in the evaluation of teaching practices. In *2016 International Symposium on Computers in Education (SIIE)*, pages 1–4.

Keogh, E., Wei, L., Xi, X., Lee, S.-H., and Vlachos, M. (2006). Lb_keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *Proceedings of the 32nd international conference on Very large data bases*, VLDB '06, pages 882–893. VLDB Endowment.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, DMKD '03, pages 2–11, New York, NY, USA. ACM.

Luckesi, C. C. (2014). *Avaliação da aprendizagem escolar: estudos e proposições*. Cortez editora.

Osmanbegović, E. and Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1):3–12.

Othman, E. H., Abdelali, S., and Jaber, E. B. (2016). Education data mining: Mining moocs videos using metadata based approach. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 531–534.

Richmond, G., Salazar, M. d. C., and Jones, N. (2019). Assessment and the future of teacher education.