Data mining strategies to explore and analyze student data of secondary school

Stéfany L. Esteves, Patrícia De Nardi, Lucas C. Oliveira
Departamento de Ciência da Computação
Federal University of Lavras
Lavras, MG, Brazil
lealesteves@sistemas.ufla.br, patynardi@sistemas.ufla.br
lucacharles@sistemas.ufla.br

Vinícius R. P. Borges
Departamento de Ciência da Computação
University of Brasília
Brasília, DF, Brazil
viniciusrpb@unb.br

ABSTRACT

The analysis of student's data and the identification of relevant knowledge in educational databases are relevant tasks in schools, universities and other institutes. They can be useful for planning an academical year, for improving student's learning and also to avoid dropouts. Educational databases present many records, which are described by a large number of attributes, which is unfeasible to be performed by human analysts. This paper proposes to use data mining to investigate and identify relevant patterns in student's data of a secondary school level in Portugal. The strategy employed in this study consists of formulating hypothesis over students' characteristics, modeling appropriate data mining techniques and applying them to a subset of features. Experimental results indicated that students interested to attend college in the future, the parent's educational level and their past experiences and grades are important factors when analyzing their performances.

Categories and Subject Descriptors

Computing methodologies [**Machine learning**]: Machine learning approaches

Keywords

Student's performance, data mining, feature selection, clustering, classification

1. INTRODUCTION

Handling and investigating academical and social data of students is an important task to support analysts of educational institutions and associated researchers. Such information is extremely relevant for planning a school year, offering specific courses of a semester, understanding the student's learning process and providing directions to teachers, professors and lecturers for teaching purposes. Institutions and educational centers using that strategy are likely to attract motivated students which are looking for courses and institutions offering education with high quality [9]. As a result, student's performance tend to improve, leading to increase the rate of students concluding their courses and reducing occasional dropouts [14].

The modern computers are capable of storing and processing a large amount of data, which is unfeasible for a human specialist to perform similar tasks. To convey the obtaining of the most relevant data patterns, automated systems employing data mining have been devised in the past years. For instance, data mining can be applied to predict student's performance [1] or to identify patterns of students' clusters [5] using several variables, such as social, familiar, financial etc. These methodologies have been extensively studied in educational data mining [2, 3], allowing researchers and analysts to discover new, interesting and useful knowledge about students, which can potentially improve the quality of education [7].

Several works have been proposed for analyzing student's data and related tasks. Shahiri et al. [11] provide a systematic literature review concerning the application of data mining techniques for predicting performance of students in Malaysia. Dutt et al. [6] survey several studies and applications in educational data mining, such as assessing tools for improving teaching methodologies, using resources on student's learning, understanding why some students fail in an academical year, among others. Cortez and Silva [4] approached the prediction of students' performances using traditional classifiers (decision tree, random forest, neural network and support vector machines) and considering three combinations of input settings. In particular, as the dataset employed in Cortez and Silva's work is available in public domain, we decided to use it in our investigations on analyzing student's data, though using different research methodologies from those reported in their work.

In general, the studies above consider datasets that describe students of a specific region or presenting a particular educational level. Applying the proposed methodologies to different datasets can lead to unexpected results, since student's patterns vary between datasets. Our strategy relies on formulating hypothesis over student's information (for instance, if Internet is relevant for the student's performance) and adjusting appropriate data mining techniques (such as classifiers or clustering techniques) in order to find relevant patterns.

This paper is organized as follows: Section 2 presents previous studies on the analysis and prediction of student's

SBSI 2017, June 5th-8th, 2017, Lavras, Minas Gerais, Brazil

performance and provides the motivation of this research. Section 3 describes the proposed methodology to analyze a public dataset of student's performance, detailing the preprocessing, feature selection and the adopted data mining techniques. Section 4 reports the experimental results and discusses the discovered patterns on the student's dataset. Section 5 presents the final considerations and alludes to possibilities for future work.

2. RELATED WORK

Data mining have emerged with potential tools to analyze student's performance data. It has been subject of research particularly in educational data mining [2], which is a process that extracts relevant patterns and useful information from large educational datasets.

Ramaswami and Bhaskaran [10] surveyed an experimental methodology to analyze the factors which affect student performances in Indian high schools. A dataset of 1000 students were obtained from two sources: the primary data was collected from the regular students, while the secondary data was gathered from the school and Chief Educational Officer (CEO). After preprocessing and preparing raw data, 772 student records were employed in the prediction model. The accuracy of the present model was compared with other model and they reported to be satisfactory considering mostly social and personal factors.

Yukselturk et al. [14] investigated by means of data mining techniques the main factors of student dropouts in an online program. Their studies considered a dataset constituting of 189 students and ten variables, such as gender, age, educational level and previous online experience. Taking the dropout status as the class label, the key idea was to train classifiers and predict through 10-fold cross validation if a student instance dropout or not. Four classifiers were employed: K-Nearest Neighbors, Naive Bayes, C4.5 decision tree and a feed-forward multilayer perceptron (MLP) with backpropagation. According to the classification results, MLP achieved the highest sensitivity. In addition, a feature selection method based on a genetic algorithm reported that online technologies self-efficacy, online learning readiness, and previous online experience were the most relevant factors for predicting student dropouts.

Cortez and Silva [4] employed data mining techniques to investigate student's performance in secondary education considering two Portuguese secondary schools. Student data were collected by means of questionnaires and school reports. The target classes were Portuguese and Math courses and the data analysis were conducted by performing binary and five-level classifications, and regression. Data preprocessing was required prior to such tasks due to the presence of nominal attributes. Three configurations that take into account past school grades, demographic and social attributes were tested for predicting the student's performance. Experimental results reported a high predictive confidence if the first and/or secondary school period grades are known, i.e., the student's performance is highly influenced by past performances.

It is worth noting that the studies mentioned above used particular datasets according to their school, institution, university, region or country. Moreover, such datasets are not available in public domain, which constraints additional and extended researches using other methodologies or new data mining techniques. Thus, we chose to analyze Cortez and Silva's dataset once it is available for download at the University of California, Irvine (UCI) repository. However, our research differs from Cortez and Silva's, because we formulate questions about social, financial and familiar information of students, their behavioral characteristics, and attempts to associate them with their performances (for instance, the final grades, absences, failures in the academical year). Appropriate data mining techniques are applied in order to answer them or to find relationships that can support analysts in further investigations.

3. PROPOSED METHODOLOGY

3.1 Dataset

A public dataset [1] [4] describing student's achievement in secondary education of two Portuguese schools was employed in the experiments. The data attributes comprise student's grades, their demographic, social, financial, personal and school characteristics and other related features. Such data was collected by means of school reports and questionnaires in two Portuguese schools.

The dataset employed in our studies refers to 395 students of the Math subject and it is described by 33 attributes, as reported on Table 1. Originally, the final grade ("G3" on Table 1) is considered as the class label attribute for predicting students performance. However, in this work, as the focus is not only to predict student's performance, the target class may vary according to the proposed analysis.

3.2 Data preprocessing

A preliminary view of the data attributes motivated us to perform a preprocessing step prior to the application of the data mining techniques, since attributes are from different types (numerical and nominal) are present and due to the presence of irrelevant attributes.

The dataset does not present missing or inconsistent values. The nominal and numeric attributes are present, demanding a standardization on data attributes to facilitate comparison between instances in further steps. We follow the strategy described in [12] to convert nominal attributes to numerical ones so that well-known metrics (such as the Euclidean distance) can be applied to compare data instances. For example, attribute Mjob is mapped from nominal to numerical by representing the value "mother" by zero, "father" by one and "other" as 2. Furthermore, we follow the strategy described by Cortez and Silva to discretize the class label attribute "G3". Such attribute is transformed to binary, in a way that students with grades greater or equal to 10 passes in the course, and fails otherwise.

The raw data is described by 33 attributes, which some of them are redundant or irrelevant depending on the target task. Thus, a dimensionality reduction is performed using the gain ration measure with an entropy criterion [8] and considering "G3" as the class label. We rank all the attributes according to their obtained gain ratio measures, in which the top ranked are those with the highest gain ratio values. We chose to select the top-15 attributes, which

[1]Dataset available at `http://archive.ics.uci.edu/ml/datasets/Student+Performance`

[2]0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education

[3]'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'athome' or 'other')

Table 1: The complete description of the dataset and their attributes. G3, the final grade, is the target attribute (class label).

attribute name	description
school	student's school (binary: "Gabriel Pereira" or "Mousinho da Silveira")
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: "urban" or "rural")
famsize	family size (binary: "less or equal to 3" or "greater than 3")
Pstatus	parent's cohabitation status (binary: "living together" or "apart")
Medu	mother's education (numeric: 0 to 3[2])
Fedu	father's education (numeric: 0 to 3)
Mjob	mother's job (nominal)[3]
Fjob	father's job (nominal)
reason	to choose school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	travel time to school (numeric: 1. <15 min., 2. 15 to 30 min., 3. 30 min. to 1 hour, or 4. >1 hour)
studytime	weekly study time (numeric: 1. <2 hours, 2. 2 to 5 hours, 3.& 5 to 10 hours, or 4. >10 hours)
failures	number of past class failures (numeric: n if 1≤n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

are sorted in the rank as: G2, G1, failures, higher, absences, goout, age, schoolsup, guardian, Dalc, romantic, Fedu, paid, Medu, Mjob. A discussion over these selected attributes are provided next.

3.3 Data mining techniques

The strategy of this work consists of formulating some questions and applying appropriate data mining techniques to address them. In this research, three questions were formulated and are described below:

1. Is the interest in attending college relevant for the student's performance?
2. Is it possible to predict the living area of students using their social attributes and past performances?
3. Can we associate some assort of familiar factors and the free time after school to the student's performance?

Questions 1 and 2 are addressed by using their associated attributes and by creating a classification model. We attempt to answer such questions by viewing them as separated binary classification problems. For each question, the target class can be adjusted to an appropriate attribute: for questions 1 and 2, the attributes `higher` and `address` are set as the class label, respectively.

The classification model relies on a feed-forward multilayer perceptron with backtracking. The input layer of such neural network is constituted of 15 nodes and the output layer is set with one node, because in both cases, the class label present only two values. Two hidden layers, each one containing two nodes, obtain the best correct classification rate among some experimentations. For that purpose, we tested the neural network comprising from one to two layers, and varying the number of nodes in each layer from one to three. The final grade (G3) was set as the class label and was transformed to binary, in which grades greater or equal than 10 are set as '1' (student is approved) and '0' (student is not approved) otherwise.

Answering Question 3 requires a cluster analysis, which has been performed using K-means algorithm. Its key idea is to define K centroids, each one associated to a cluster, and then assign each of the dataset instances to the most similar centroid according to its features and a dissimilarity measure. More details about the process of setting parameter K are provided in the next section. After performing clustering, we analyze the instance patterns within the clusters and discuss their local relationships.

4. EXPERIMENTAL RESULTS

The experiments were conducted using the Weka environ-

ment [13]. The strategy for evaluating the classification was the leave-one-out strategy, in which for a dataset containing N instances, $N-1$ instances are used for training and the remaining one is used for test. Each instance of the dataset is employed as a test instance in this evaluation strategy. A confusion matrix is obtained from the experiments and the correct classification rate is reported in each case for measuring the correct predictions.

Our idea when attempting to address question 1 is to verify if using student's past grades, failures, absences and other attributes, their interest in attending college can be predicted. The strategy is to use the attribute `higher` as the class label and the remaining attributes as input to the classifier. The interest in attending college could be correctly predicted in 92.40% of the cases, indicating the relevance of this factor and that this information can be successfully predicted from other student attributes.

In question 2, we would like to verify if the student's home (if the student lives in the city or rural areas) affects the performance. The same strategy is employed: we take the attribute `address` as the class label and the remaining attributes as input to the classifier. Experiments reported a correct classification rate of 73.67%, meaning that the student's address can be predicted considering his final performances, past grades e the other attributes. Thus, the way students go to school can be relevant for their performances and learning process.

In K-Means, we considered 5 attributes of the dataset as the input of the clustering which were manually selected: father's education (Fedu), mother's education (Medu), family educational support (famsup), extra-curricular activities (activities) and free time (freetime). The final grade was discretized as 5-levels as proposed by Cortez and Silva, in which the ranges are: [16-20], [14-15], [12-13], [10-11] and [0-9]. The Euclidean distance is chosen as the dissimilarity measure. K-Means were performed by setting $K = 5$ and the results showed that the cluster associated with the higher grades presented students whose parents have high educational levels and provide support during their learning at school. This group is composed of 14% of the instances, with a grade average of 75%. On the other hand, the cluster associated with the lower grades presents students whose parents have only the basic education. In addition, free time after school showed to be irrelevant since diverse values were encountered along the clusters.

Feature selection based on the gain ratio measures suggests that the parent's education has a fundamental role for motivating students, which somehow affects their performances. It is important to note that the student's dedication and his personal life are also relevant factors. The student's past grades and past failures are extremely relevant for predicting their future performances as Cortez and Silva also reported.

5. CONCLUSIONS

This paper presented a study which consisted of the application of data mining techniques to a student's performance dataset. A public dataset containing information of two Portuguese schools has been employed concerning the Mathematics subject. The goal of this study was to explore the relations between the student's performance with several aspects (social, personal, financial and geographical), as well as to identify relevant patterns in student's data.

The experimental results allowed us to identify that the most relevant factors that affects student's performance are their past grades, the absences and past failures. Moreover, the conditions which students go to school are reasonably important, while their motivations to attend college in the future showed to be correlated with their performances. The parent's education also has a fundamental role for students, in which higher grades could be observed when parents have higher education.

Possibilities for future work comprise the use of datasets of Brazilian students, so that we can understand a scenario that is more familiar to our reality. Furthermore, it is also interesting to investigate in more details the benefits of applying feature selection methods to these data.

6. REFERENCES

[1] A. B. E. D. Ahmed and I. S. Elaraby. Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, 2(2):43–47, 2014.

[2] A. P. Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[3] R. S. Baker. Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, 29(3):78–82, 2014.

[4] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. *Universidade do Minho, Portugal*, 2008.

[5] A. Dutt, S. Aghabozrgi, M. A. B. Ismail, and H. Mahroeian. Clustering algorithms applied in educational data mining. *Int. Journal of Information and Electronics Engineering*, 5(2):112, 2015.

[6] A. Dutt, M. A. Ismail, and T. Herawan. A systematic review on educational data mining. *IEEE Access*, 2017.

[7] I. E. Livieris, T. A. Mikropoulos, and P. Pintelas. A decision support system for predicting students' performance. *Themes in Science and Technology Education*, 9(1):43–57, 2016.

[8] D. Oreski, S. Oreski, and B. Klicek. Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, 52:109–119, 2017.

[9] Z. K. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

[10] M. Ramaswami and R. Bhaskaran. A chaid based performance prediction model in educational data mining. *arXiv preprint arXiv:1002.1144*, 2010.

[11] A. M. Shahiri and W. Husain. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015.

[12] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, us ed edition, 2005.

[13] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[14] E. Yukselturk, S. Ozekes, and Y. K. Türel. Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 17(1):118–133, 2014.