Expert Systems with Applications 42 (2015) 5508–5521

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining models for student careers

Renza Campagni, Donatella Merlini *, Renzo Sprugnoli, Maria Cecilia Verri

Dipartimento di Statistica, Informatica, Applicazioni Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy

ARTICLE INFO

Article history:
Available online 6 March 2015

Keywords:
Data mining
Educational data mining
Student careers
Clustering
Frequent pattern analysis

ABSTRACT

This paper presents a data mining methodology to analyze the careers of University graduated students. We present different approaches based on clustering and sequential patterns techniques in order to identify strategies for improving the performance of students and the scheduling of exams. We introduce an *ideal career* as the career of an ideal student which has taken each examination just after the end of the corresponding course, without delays. We then compare the career of a generic student with the ideal one by using the different techniques just introduced. Finally, we apply the methodology to a real case study and interpret the results which underline that the more students follow the order given by the ideal career the more they get good performance in terms of graduation time and final grade.

1. Introduction

The large supply of data stored in the computer systems of several companies, both public and private, has given a push in the direction of the development of new technologies for data management and analysis. Data mining techniques originate in this context, with the aim of discovering hidden and non-trivial relationships among information of various nature. This collection of techniques, used in different sectors, including the educational environment, comes from the traditional methods of data analysis and have the characteristic of being able to treat large amounts of data.

In the field of education, *educational data mining* is a recent research area that explores and analyzes the information stored in student databases in order to understand and improve the performance of the student learning process. Data are analyzed by using statistical, machine learning and data mining algorithms, with the aim of resolving problems of educational research and improve the entire educational process. Recently there has been an increase in the use of educational software instruments and of databases containing students information, so we have large repositories of data reflecting how students learn. In addition, the use of Internet in education has created the context of e-learning or web-based education which continuously generates large amounts of data concerning the interactions between teaching and learning. Educational data mining tries to use all this information to better understand learners and learning, and to develop methodologies which, integrating the data with the theory, allow to improve the educational process. Educational data mining is a growing research area that involves researchers all over the world from different and related research areas and since 2008 an annual International Conference on Educational Data Mining has been established (http://www.educationaldatamining.org). Great efforts have been made in the direction of describing the state of the art of this research area and, in the recent years, several survey papers have been published on the subject (Baker, 2010, 2014; Baker & Yacef, 2009; Luan, 2002; Peña-Ayala, 2014; Romero & Ventura, 2010, 2013).

As already observed, usually data mining techniques are applied to large data sets. In the context of education, however, we are often faced with data sets corresponding to small groups of students following the same curriculum. Referring to a university context, for example, even when a degree program is frequented by many students the data of interest correspond to relatively small data sets. The recent paper (Natek & Zwilling, 2014) focuses on the study of data mining techniques applied to small data sets concerning higher education institutions and concludes that the use of these techniques in real-life situations is useful and promising and can provide administrators with precious tools for decision.

Over the years, several data mining models have been designed and implemented to analyze the performance of students. For example, in Delavari, Shirazi, and Beikzadeh (2004) and Delavari, Somnuk, and Beikzadeh (2008), a model is proposed which presents the advantages of data mining technology in higher educational systems; the authors give a sort of road map to assist the institutions to identify the ways to improve their processes. In Daimi and Miller (2009), the authors illustrate a classification

* Corresponding author.
E-mail addresses: renza.campagni@unifi.it (R. Campagni), donatella.merlini@unifi.it (D. Merlini), renzo.sprugnoli@unifi.it (R. Sprugnoli), mariacecilia.verri@unifi.it (M.C. Verri).

http://dx.doi.org/10.1016/j.eswa.2015.02.052
0957-4174/

model to investigate the profile of students which most likely leave university without ending their career. In particular, they use some classification algorithms implemented in the WEKA system (Witten, Frank, & Hall, 2011). Recommendations of suitable courses for students are analyzed with different approaches in Bydzovska and Popelínský (2014) with the aim of predicting student success. In Damaševičius (2010) a framework is proposed for mining educational data using association rules. More recently, Romero, Zafra, Luna, and Ventura (2013) proposes the application of association rule mining to improve quizzes and courses and (Saarela et al., 2014) applies frequent itemset mining and association rule learning to students previously grouped by clustering techniques. In Guruler, Istanbullu, and Karahasan (2010), in order to explore the factors having impact on the success of university students, a system based on the decision tree classification technique is presented. Clustering is used in Campagni, Merlini, and Verri (2014) for analyzing data concerning the evaluation of courses taken by students, linked to their results in the corresponding exams. The work presented in Dutt, Aghabozrgi, Ismail, and Mahroeian (2015) reviews different clustering algorithms applied to educational data mining context while (Peña-Ayala, 2014) is an interesting review of recent educational data mining development whose contents are in turn analyzed by a data mining approach. As already observed, data mining techniques have also been applied in computer-based, e-learning and web-based educational systems (Bouchet, Harley, & Trevors, 2013; Bogarín, Romero, Cerezo, & Sánchez-Santillán, 2014; Castro, Vellido, Nebot, & Mugica, 2007; Hämäläinen, Laine, & Sutinen, 2006; Koedinger, Cunningham, Skogsholm, & Leber, 2008; Mostow & Beck, 2006; Merceron & Yacef, 2005; Romero, Romero, Luna, & Ventura, 2010; Romero, Ventura, & García, 2008; Romero, López, Luna, & Ventura, 2013). The existing literature about the use of data mining in educational systems is mainly concerned with techniques such as clustering, classification and association rules (Damaševičius, 2010; Tan, Steinbach, & Kumar, 2006; Witten et al., 2011; Wu & Kumar, 2009).

An academic curriculum usually defines a specific learning program which puts some types of restrictions on how the students are required to take courses. These constraints typically describe a set of courses and a set of relationships between them. In the current practice, however, students have many degree of freedom, therefore helping students to choose courses, discovering patterns and key courses, planning future courses and refining curricula based on the feedback of students are important educational tasks, as recently pointed out in Aher and Lobo (2013), Kardan, Sadeghi, Ghidary, and Sani (2013), Méndez, Ochoa, and Chiluiza (2014) and Pechenizkiy, Trcka, Bra, and Toledo (2012). The present work fits into this context extending and unifying the results presented in Campagni, Merlini, and Sprugnoli (2012a, 2012b, 2012c). In particular, we introduce the concept of *ideal career*, that is, the career of a graduated student who takes every examination just after the end of the corresponding course, without delay, and propose a data mining methodology, based on clustering and sequential pattern analysis, to study the student behavior by comparing student careers with the ideal one. Sequential pattern analysis has been used in the context of educational data mining mainly in computer-based environments. For example, Soundranayagam and Yacef (2010) explores the order in which students access e-learning resources as they solve set assessment tasks, such as tests, assignments and exams and the links with students learning. A method to automatically detect collaborative patterns of student and tutor dialogue moves is illustrated in D'Mello, Olney, and Person (2010). Paper (Martinez, Yacef, Kay, Al-Qaraghuli, & Kharrufa, 2011) mines and clusters frequent patterns to compare distinct behaviors between low and high achievement groups around an interactive tabletop. A data mining methodology for identifying and comparing learning behaviors from students learning interaction traces is presented in Kinnebrew, Loretz, and Biswas (2013); in particular, the paper proposes an algorithm that employs a novel combination of sequence mining techniques to identify differentially frequent patterns between groups of students. Paper (Guerra, Sahebi, Brusilovsky, & Lin, 2014) models and examines patterns of student behavior with parameterized exercise. A recent research which proceeds in a direction similar to ours is illustrated in Asif, Merceron, and Pathan (2014), where the progression of a student is analyzed by defining a tuple that shows how the results of a year stay the same, increase or decrease compared to first year.

The preprocessing phase is the first step in any data mining process and allows us to transform the available data into a format suitable for the analysis. The importance of this task has been recently highlighted in Romero, Romero, and Ventura (2014). In Section 2, we illustrate the preprocessing phase necessary to organize data for our analysis. A crucial aspect during this phase is the insertion in the database of the reference to the semester in which a course has been given by a teacher and the semester in which the student has taken the corresponding exam. This information allows us to define the ideal career together with the career of each graduate student. We represent a career as a trajectory of points in the plane. In particular, the ideal career is defined by a sequence of points $\tau_{\mathcal{I}} = ((0, e_0), (1, e_1), (2, e_2), \cdots, (n, e_n), (n+1, e_{n+1}))$, where e_i is an exam identifier and i its position in the career. The position $i = 0$ denotes the starting point of the career while $i = n+1$ corresponds to the final examination given last by all students. This particular career, without loss of generality, can be represented by the bisecting line of the first quadrant (green lines in Fig. 1). The career of a generic student $\mathcal{J}$ is then represented by a broken line, corresponding to the sequence of points $t_{\mathcal{J}} = ((0, e_{\mathcal{J}_0}), (1, e_{\mathcal{J}_1}), (2, e_{\mathcal{J}_2}), \cdots, (n, e_{\mathcal{J}_n}), (n+1, e_{\mathcal{J}_{n+1}}))$, where $e_{\mathcal{J}_i}$ is the identifier of the exam given by student $\mathcal{J}$ at time i (red lines in Fig. 1). We then compute the distance between a generic career $t_{\mathcal{J}}$ and $\tau_{\mathcal{I}}$ in different ways, by using for example the *Bubblesort distance*, defined as the number of inversions in the permutation relative to $t_{\mathcal{J}}$, or the *area* between the lines $t_{\mathcal{J}}$ and $\tau_{\mathcal{I}}$. Finally, we insert these values in the database.

In Section 3, we analyse the preprocessed data with clustering and sequential pattern techniques.

For what concerns the cluster analysis, the idea is to explore the database with the aim of understanding if there exists a relation between the distance from the ideal career and the success of students. This kind of analysis, accompanied by cluster validation, can highlight different groups of students characterized by similar distances and behaviors and can give some suggestions to improve the organization of the laurea degree or to recommend precedence relations among courses.

Sequential pattern analysis aims to find relationships between occurrences of sequential events, that is, to find if any specific order of the occurrences exists. In this paper we consider as events the exams taken by a student; the temporal information is the semester in which the exam has been taken or the delay with which it has been taken. We study an organization of the university which allows students to take an exam in different sessions after the end of the course, as in Italy. The temporal information allows us to see the career of a student as a sequence $\langle s_1 s_2 \dots s_m \rangle$ where each element s_j is a collection of one or more exams taken in the same semester or having the same delay. If we use the semester as temporal information, m indicates the number of semesters in which a student takes exams; if we use the delay, m indicates the maximum number of delays, in semesters, with which a student takes one or more exams. By analyzing the sequential patterns, we can explain some behaviors which may seem