Analyzing paths in a student database

Renza Campagni, Donatella Merlini, Renzo Sprugnoli
Dipartimento di Sistemi e Informatica
viale Morgagni 65, 50134, Firenze, Italia
[renza.campagni,donatella.merlini,renzo.sprugnoli]@unifi.it

1. INTRODUCTION

In the recent literature, several data mining models have been proposed to understand and improve the educational performance and assessment of a student learning process. For example, in [1] the authors illustrate a classification model to investigate the profile of students which most likely leave university without ending their study; in [2], the author uses association rule mining for assessing student results. Data mining techniques have been also applied in computer-based educational systems (see, e.g., [3]). In our research we present a data mining approach to analyze a database containing information about students, in particular, their personal but anonymous data and their exams. We consider the *path* of a student, that is, the way the student implement her or his exams over the degree-learning time: a student can take an exam immediately after a course, the *ideal* choice, or later.[1] The aim of this work is to understand how this order affects the performance of the students in terms of graduation time and final grade. See [4] for a recent research in the field of choices in learning. We consider the *ideal path*, i.e., the path of a student which has taken each examination just after the end of the corresponding course, without delay. Therefore, the ideal path matches the curriculum settled by the academic degree. The path of a generic student is then compared with the ideal path by using two different approaches: the first one uses the *Bubblesort* distance while the second is based on the computation of the *area* between the two paths. Students who have taken the exams in the same order, that is, students with the same path, can have different final grade and graduation time. The idea is to understand if there exists a relation between these distances and the success of students. If the students having small distances achieve good performance, then we may conclude that the academic degree is well structured but if there exist many good students with large distances, then can mean that the organization should be modified. Once the distances have been computed, they can be inserted in the database, as new attributes of each student, and a clustering analysis can be performed, for example by using the `K-means` implementation of `WEKA` (see, e.g., [5]). By using this methodology on our database, with both approaches and $\mathtt{K} = 2$, we obtained two clusters characterized by *small* and *large* distances. The first one corresponds to the group of students who graduated relatively quickly and with high grades; the second cluster corresponds to students who obtained worst results. Our analysis shows that the more students follow the order given by the ideal path the more they get good performance in terms of graduation time and final grade. In conclusion, no student with a large distance achieves good results; we can conclude that the academic degree under consideration was well scheduled.

[1]We refer to an organization which allows students to take an exam in different sessions after the end of the course, as in Italy. A drawback is that students end up graduating with a significant delay. Some constrains between exams can be fixed in order to force students to take exams in a specific order, however, usually students have many degrees of freedom.

2. THE METHODOLOGY

We consider a database containing the data of N students, each student characterized by a sequence of n exams identifiers. We consider a particular path $\mathcal{I} = (e_1, e_2, \cdots, e_n)$, the *ideal path*, corresponding to a student which has taken every examination just after the end of the corresponding course, without delay. Without loss of generality, we can assume that $e_i = i$, $i = 1, \cdots, n$, that is, $\mathcal{I} = (1, 2, \cdots, n)$. The path of a generic student J can be seen as a sequence $\mathcal{J} = (e_{J_1}, e_{J_2}, \cdots, e_{J_n})$ of n exams, where e_{J_i}, $i = 1, \cdots, n$, is the identifier of the exam taken by the student at time i. Therefore, $\mathcal{J}$ can be seen as a permutation of the integers 1 through n. In order to understand how the order of the exams affects the final result of students, we compare a path $\mathcal{J}$ with $\mathcal{I}$ by using two different approaches. The first approach uses the *Bubblesort distance*, which is defined as the number of exchanges performed by the Bubble sort algorithm to sort an array containing the numbers from 1 to n. The number of exchanges, bounded above by $n(n-1)/2$, can be computed easily since it is exactly the number of inversions in the permutation. Our second approach concerns the graphical representation of the paths; we represent them in the integer lattice, the x-axis denotes the number of exam and the y-axis the exam identifier, according to the order of the ideal student. The ideal path is defined by the 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 $e_0 = 0$ denotes the starting point of the path and $e_{n+1} = n + 1$ denotes the final examination taken last by all students. Therefore, $\tau_{\mathcal{I}}$ can be represented as the bisecting line of the first quadrant. The path of a generic student J, is then represented by a broken line corresponding to the sequence of points $t_J = ((0, e_{J_0}), (1, e_{J_1}), (2, e_{J_2}), \cdots, (n, e_{J_n}), (n + 1, e_{J_{n+1}}))$. By convention, we have $e_{J_0} = 0$ and $e_{J_{n+1}} = n+1$ for every student J, that is, the resulting trajectory begins at $(0, 0)$ and finishes in the point $(n + 1, n + 1)$. We then compare a path $\mathcal{J}$ with $\mathcal{I}$ by computing the area $\mathcal{A}_{\mathcal{J},\mathcal{I}}$ between

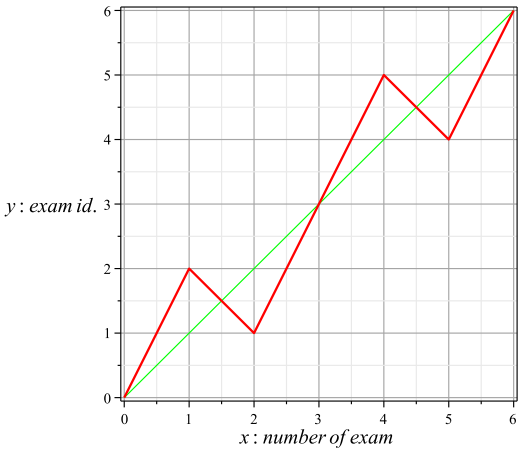

Figure 1: The path $(2, 1, 3, 5, 4) : \sigma(\mathcal{J}) = 2$ and $\mathcal{A}_{\mathcal{J},\mathcal{I}} = 3$.

$\tau_{\mathcal{I}}$ and t_J. In Figure 1 we illustrate the path $(2, 1, 3, 5, 4)$ and its distances from the ideal path $(1, 2, 3, 4, 5)$.

2.1 Cluster analysis

The database we analyze contains data of students in Computer Science at the University of Florence beginning their studies during the years 2001-2004 and graduated up to now. The academic degree in Computer Science at the University of Florence is structured in three years (Laurea triennale) and students can choose among different curricula. Every year is organized in two semesters; there are several courses in each semester and at the end of a semester students can take their examinations. Exams can be taken in different sessions during the year and students can try to pass their exams in any of these sessions, after the end of the course. In the years under consideration, no constrains between exams were fixed, so students could take their exams almost in any order. For each student, the database contains the identifier of the student, the grade obtained at the high school level, the year of enrollment at the university, the date and the mark of final examination and the sequence of exams; each exam is described by an identifier, a date and a grade. We considered students belonging to two slightly different curricula: databases and distributed systems. In particular, we analyzed the paths of $N = 100$ students characterized by a sequence of $n = 25$ exams. For each curriculum we computed the ideal path through an important pre-processing phase, which allowed us to identify the semester in which courses were originally hold by the teacher. In fact, the original database did not contain the information about the semester, which is fundamental for our purposes. In the ideal path, courses relative to the same semester were sorted by taking into account the preference of students. Therefore, we obtained two different ideal paths of length 25. Then, for each student J of both curricula we computed the distances $\sigma(\mathcal{J})$ and $\mathcal{A}_{\mathcal{J},\mathcal{I}}$ from the ideal path and inserted these values into two fields `Bubblesort` and `Area` of the database. We tried to sort our data according to both fields and we found some pairs of paths having values of `Bubblesort` and `Area` in reverse order. Therefore, these two distances are not completely equivalent; however, this difference seems not to be important for the clustering analysis, as we will see later. To understand how the order of the exams affects the path of the students, we have performed several tests by using the `K-means` implementation of `WEKA`. We first analyzed the paths of students separately for the two curricula. In particular, in both cases, we obtained significant result with $K = 2$ and by selecting as clustering attributes the graduation time, `Time`, expressed in days, and the final grade, `Grade`, an integer between 66 and 110. In fact, with these parameters we can see that students are well divided into two groups: the group of students who graduated relatively quickly and with high grades and the group of students who obtained worst results, respectively. Luckily, we observed that students in the first group are characterized by *small* values of `Bubblesort` and `Area` while students in the second group have *larger* values. Our analysis shows also that the path of a student seems not to be affected by the results achieved at the high school level. We performed similar tests by adding the distance values as attributes of clustering, and we obtained two more distinct clusters, which divide students more precisely in terms of `Time`, `Grade` and `Bubblesort` (or `Area`) distance. We finally analyzed together the students belonging to the two curricula obtaining 2 clusters with the following characteristics:

Attribute	Full data	Cluster 1	Cluster 2
Bubblesort	89.7	71.8	119.2
Area	121.5	96.8	162.2
Time	2156.7	1841.9	2678.1
Grade	98	101	94

This result confirms that, regardless of the curriculum, the more students follow the order given by the ideal path, the more they obtain good performance in terms of graduation time and final grade. We point out that we obtained similar results by using either distance. In conclusion, the methodology proposed in this research can be used to valuate the organization of an academic degree in terms of the scheduling of the courses.

ACKNOWLEDGEMENTS

The authors would like to thank Dino Pedreschi for interesting discussions about the model presented in this paper.

3. REFERENCES

[1] K. Daimi and R. Miller. Analyzing student retention with data mining. In *Proceedings of the 2009 International Conference on Data Mining*, pages 55–60, 2009.

[2] R. Damaševičius. Analysis of academic results for informatics course improvement using association rule mining. In *Information Systems Development*, pages 357–363. Springer, 2010.

[3] C. Romero, J. R. Romero, J. M. Luna, and S. Ventura. Mining rare association rules from e-learning data. In *The 3rd International Conference on Educational Data Mining*, pages 171–180, 2010.

[4] H. Soundranayagam and K. Yacef. Can order of access to learning resources predict success? In *The 3rd International Conference on Educational Data Mining*, pages 323–324, 2010.

[5] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.