An Analysis of Courses Evaluation Through Clustering

Renza Campagni, Donatella Merlini(✉), and Maria Cecilia Verri

Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni, 65, 50134 Florence, Italy
{renza.campagni,donatella.merlini,mariacecilia.verri}@unifi.it

Abstract. Students of the University of Florence (Italy), before taking an exam, are required to assess different aspects related to the course organization and to the teaching. The data concerning the evaluation of the courses of the Computer Science Program from 2001/2002 to 2007/2008 academic years were collected and linked to the results of students: the grades obtained in the corresponding exams and the delays, with respect to the end of the courses, with which exams were taken. After this preprocessing phase, we used clustering techniques to analyze data and we highlighted a correlation between courses evaluation and the corresponding average student results, as well as regularities among groups of courses over the years. Our analysis can be used to detect possible improvements in the organization and teaching of the degree program and applied to any university context collecting similar data.

Keywords: Educational data mining · K-means clustering · Courses evaluation · Assessment

1 Introduction

The evaluation of university education is an important process whose results can be used in the programming and management of the educational activities by monitoring resources (financial, human, structural and others), services (orientation for students and administrative offices), students careers, courses and occupancy rate. In order to evaluate all these aspects, it is important to analyse the opinion of the *users* of university education, i.e. the students.

The evaluation of the learning process falls in the context of the Educational Data Mining (EDM), an emerging and interesting research area that aims to identify previously unknown regularities in educational databases, to understand and improve student performance and the assessment of their learning process. As described in [11,12], EDM uses statistical, machine learning and data mining algorithms on different types of data related to the field of education. It is concerned with developing methods for exploring these data to better understand the students and the frameworks in which they learn thus possibly enhancing some aspects of the quality of education. Data mining techniques

S. Zvacek et al. (Eds.): CSEDU 2014, CCIS 510, pp. 211–224, 2015.
DOI: 10.1007/978-3-319-25768-6_14

have also been applied in computer-based and web-based educational systems (see, e.g., [10,13]). In this paper, we use a data mining approach based on `K-means` clustering to link the evaluation of courses taken by students with their results, in terms of average grade and delay in the corresponding exams. We also analyse the evaluation of courses over the years in order to identify similar behaviors or particular trends among courses, by using an approach similar to time series clustering (see, e.g., [8]).

This study deepens the analysis presented in [5] and is analogous to that used in [2–4]. The analysis refers to a real case study concerning an Italian University but it could be applied to different scenarios, except for a possible reorganization of the involved data. The data set is not very large but allows us to illustrate a quite general methodology on a real case study. Our approach uses standard data mining techniques, but we think very interesting the concrete possibility of applying these techniques to find and analyse patterns in the context of university courses evaluation, even in large universities.

This paper represents a revised version of [6] which integrates the feedbacks from the presentation at CSEDU 2014.

2 Data for Analysis

In this section, we describe how courses are evaluated by students at the University of Florence, in Italy, with the aim of providing a methodology to search for regularities in data concerning courses evaluation. Therefore, the steps we present can be applied also in other academic contexts. In particular, we refer to a Computer Science degree of the Science School, under the Italian Ministerial Decree n. 509/1999. This academic degree was structured over three years and every academic year was organized in two semesters; there were several courses in each of these six semesters and at the end of a semester students could take their examinations. Exams could be taken in different sessions during the same year, after the end of the corresponding courses, or later.

Table 1 illustrates an example of students data after a preprocessing phase which allow us to integrate original attributes, such as the grade and the date of the exam, with both the semester in which the course was given, **Semester1**, and the semester in which the exam was taken, **Semester2**. Finally, we can

Table 1. A sample of students data: grades in thirtieths.

Student	Exam	Date	Grade	Semester1	Semester2	Delay
100	10	2001-01-14	24	1	1	0
100	20	2002-12-20	27	2	3	1
200	20	2002-06-04	21	2	2	0
300	10	2001-01-29	26	1	3	2
400	10	2002-02-15	26	1	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

compute the value **Delay** as the difference between the semester of the course and the semester in which the student took the exam. We highlight that the values of attributes **Semester1** and **Semester2** are not usually stored in the databases of the universities, therefore this preprocessing phase may be onerous.

At the University of Florence, starting from the academic year 2001/2002, a database stores information about evaluation of the courses quality of various degree programs, among which we find the degree under consideration. The results of this process are available at the address [1], under permission of the involved teacher, and show for each course several pieces of information, such as the name of the teacher who took the course and the average rating given by students on various topics. Before the end of each course (at about 2/3 of the course), students compile, anonymously, a module to express their opinion on the course just taken. This form is divided into the following five paragraphs:

- paragraph 1, concerns the organization of the degree program;
- paragraph 2, concerns the organization of the course;
- paragraph 3, concerns the teacher;
- paragraph 4, concerns classrooms and equipment;
- paragraph 5, concerns the general satisfaction about the course.

Each paragraph is composed by some questions; students can choose among four levels of answers, two negative and two positive levels (disagree, slightly disagree, slightly agree, agree). For details the interested reader can see the sample of the module in [1].

For each course of an academic year and for each paragraph, we can compute the percentage of positive answers, that is, of type *slightly agree* and *agree* by grouping together all questions belonging to the same paragraph and their average percentage value.

To relate data of students careers with courses evaluation, for each course we can compute the average grade and the average delay attained by students who took the exam in the same year. An example of this data organization is illustrated in the first four columns of Table 2. As already observed, the evaluation of courses is anonymous and is done only by students who really take the course, therefore, in this kind of organization, it may happen to consider information

Table 2. Data organization for comparing examination results and courses evaluation.

Exam	Year	AvgGrade	AvgDelay	Par*1*	…	Par*5*
10	2001	25	1	51	…	60
10	2002	26	1	50	…	61
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	2007	25	1	81	…	67
20	2001	24	0.5	56	…	77
20	2002	26	1	62	…	59
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3. Data organization for analyzing the trend over the years of courses evaluation.

Exam	Par*1*(2001)	...	Par*1*(2007)	...	Par*5*(2001)	...	Par*5*(2007)
10	51	...	81	...	60	...	67
20	56	...	84	...	77	...	84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

concerning exams of students who may not be the same students who evaluated the courses. As a consequence, we can only compare the results of courses evaluation in a specific year with the aggregate results of students who took the corresponding exams in the same period. However, this data organization does not change a lot if it was possible to identify the students involved in the courses evaluation in order to connect properly the results of the evaluation with those of exams. Obviously, in this case we should ensure the privacy of results, for example by using a differential privacy approach (see, e.g., [7]).

After a preprocessing phase, we can organize students and evaluation data into two different ways by taking into account the following fields:

- **Exam**, the code which identifies an exam;
- **Year**, the year of the evaluation;
- **AvgGrade**, the average grade of the exam;
- **AvgDelay**, the average delay, in semesters, of students exams;
- **Par$\boldsymbol{k}$(t)**, the percentage of positive evaluations of paragraph k at time t.

In particular, Table 2 illustrates a sample of the dataset which can be used to compare examination results and courses evaluation while Table 3 represents a sample of data that can be used to analyze the evolution over the years of courses evaluation. As we will illustrate in Sect. 3, data organized as in Table 2 will be clustered with `K-means` algorithm by using the Euclidean distance to separate the *multidimensional points* representing some characteristic of a course in a specific year; data organized as in Table 3 will be represented in the plane as *trajectories* corresponding to the evaluation of courses over the years and will be clustered with the Manhattan distance. Both these approaches can be used to find regularities in courses evaluations and can highlight criticalities or suggest improvements in the teaching organization.

3 K-means Clustering with Euclidean and Manhattan Distances

Among the different data mining techniques, clustering is one of the most widely used methods. The goal of cluster analysis is to group together objects that are similar or related and, at the same time, are different or unrelated to the objects in other clusters. The greater the similarity (or homogeneity) is within a group and the greater the differences between groups are the more distinct the clusters are.

`K-means` is a very simple and well-known algorithm based on a partitional approach; it was introduced in [9] and a detailed description can be found in [14]. In this algorithm, each cluster is associated with a centroid and each point is assigned to the cluster with the closest centroid by using a particular distance function. The centroids are iteratively computed until a fixed point is found. The number K of clusters must be specified. In particular, in this paper we use both the *Euclidean* and *Manhattan* distance; in the first case, the centroid of a cluster is computed as the mean of the points in the cluster while in the second case the appropriate centroid is the median of the points (see, e.g., [14]).

The evaluation of the clustering model resulting from the application of a cluster algorithm is not a well developed or commonly used part of cluster analysis; nonetheless, cluster evaluation, or cluster validation, is important to measure the goodness of the resulting clusters, for example to compare clustering algorithms or to compare two sets of clusters. In our analysis we measured cluster validity with correlation, by using the concept of proximity matrix and incidence matrix. Specifically, after obtaining the clusters by applying `K-means` to a dataset, we computed the proximity matrix $P = (P_{i,j})$ having one row and one column for each element of the dataset. In particular, each element $P_{i,j}$ represents the Euclidean, or Manhattan, distance between elements i and j in the dataset. Then, we computed the incidence matrix $I = (I_{i,j})$, where each element $I_{i,j}$ is 1 or 0 if the elements i and j belong to the same cluster or not. We finally computed the Pearson's correlation, as defined in [14, p. 77], between the linear representation by rows of matrices P and I. Correlation is always in the range -1 to 1, where a correlation of 1 (-1) means a perfect positive (negative) linear relationship.

As a first example, Table 4 illustrates the final grade and the graduation time, expressed in years, of a sample of graduated students. By applying the `K-means` algorithm to this dataset, with $K = 2$, **FinalGrade** and **Time** as clustering attributes and by using the Euclidean distance, we obtain the following two clusters, in terms of the student identifiers: $C_1 = \{100, 400, 600, 700\}$ and $C_2 = \{200, 300, 500\}$; the centroids of the clusters have coordinates $\mathcal{C}_1 = (107, 3.5)$ and $\mathcal{C}_2 = (96, 5.33)$, respectively. Tables 5 and 6 show the proximity matrix and the incidence matrix corresponding to clusters C_1 and C_2 of the data set illustrated in Table 4. The Pearson's correlation between the linear representation of these two matrices is -0.59, a medium value of correlation.

As another example, Table 7 shows a sample of data concerning courses evaluation: in particular, each row contains the exam identifier and the percentage of positive evaluation of a generic paragraph at time t_i, for $i = 1, \cdots, 4$. We can apply the `K-means` algorithm to the dataset in Table 7, with $K = 2$, $\mathbf{Par(}t_i\mathbf{)}$, for $i = 1, \cdots, 4$, as clustering attributes and by using the Manhattan distance. This means to represent each element of the data set as a broken line connecting the points $(t_i, \mathbf{Par}(t_i))$, for $i = 1, \cdots, 4$, in the cartesian plane. The Manhattan distance between two broken lines thus corresponds to the sum of the vertical distances between the ordinates. By using the `K-means` algorithm, we obtain the

Table 4. A sample data set about students.

Student	FinalGrade	Time
100	110	3
200	95	5
300	100	5
400	103	4
500	98	6
600	106	4
700	109	3

Table 5. The proximity matrix for data of Table 4.

P	100	200	300	400	500	600	700
100	0						
200	20.12	0					
300	10.25	10	0				
400	7.07	13.08	3.32	0			
500	12.41	8.06	2.24	5.48	0		
600	4.12	16.06	6.16	3	8.31	0	
700	1	19.13	9.27	6.08	11.45	3.16	0

Table 6. The incidence matrix for clustering of data of Table 4.

I	100	200	300	400	500	600	700
100	1						
200	0	1					
300	0	1	1				
400	1	0	0	1			
500	0	1	1	0	1		
600	1	0	0	1	0	1	
700	1	0	0	1	0	1	1

following two clusters in terms of course identifiers: $C_1 = \{200, 500\}$ and $C_2 = \{100, 300, 400\}$; the centroids of the clusters are represented by the sequences $\mathcal{C}_1 = [(1, 72), (2, 68), (3, 67), (4, 73)]$ and $\mathcal{C}_2 = [(1, 82.5), (2, 91), (3, 87.5), (4, 91.5)]$, respectively. Figure 1 illustrates the clustering result by evidencing the centroids $\mathcal{C}_1$ and $\mathcal{C}_2$.

Also in this case we can compute the Pearson's correlation by using the proximity and the incidence matrices computed by using the Manhattan distance.

Table 7. A sample data set about courses evaluation.

Exam	Par(t_1)	Par(t_2)	Par(t_3)	Par(t_4)
100	55	65	67	60
200	85	87	85	92
300	72	68	65	77
400	77	80	70	73
500	80	95	90	91

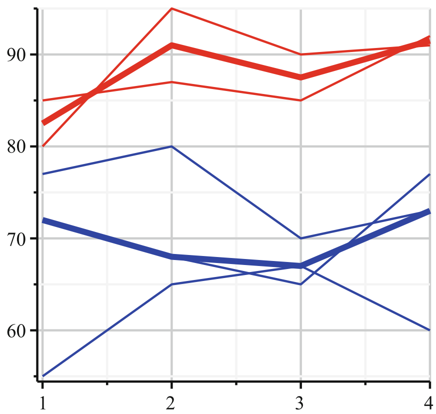

Fig. 1. `K-means` results with data of Table 7 with $K = 2$ and Manhattan distance, centroids in evidence.

3.1 The Case Study

As already observed, the real datasets we analysed concern courses and exams during the academic years from 2001/2002 to 2007/2008 at the Computer Science program of the University of Florence, in Italy. In particular, the first data set is organized as illustrated in Table 2 and refers to the evaluation of 40 courses in seven different years. We explicitly observe that we did not consider in our analysis those courses evaluated by a small number of students. For clustering, we used the `K-means` implementation of `Weka` [15], an open source software for data mining analysis. The aim was to find if there is a relation between the valuation of a course and the results obtained by students in the corresponding exam. We performed several tests with different values of the parameter K and we selected different groups of attributes. We point out that the attributes selection is an important step and should be done according to the preference of an expert of the domain, for example the coordinator of the degree program. For each choice of attributes, we applied the `K-means` algorithm with the Euclidean distance to identify the clusters; then, we computed the Pearson's correlation by using the proximity and incidence matrices. The tests we performed pointed out that the exams having good results, in terms of average grade and delay, correspond to courses having also a good evaluation from students.

In particular, we used **AvgGrade**, **AvgDelay**, **Par*1***, **Par*2***, **Par*3***, **Par*4*** and **Par*5*** as clustering attributes and $K = 2$, obtaining the clusters illustrated in Figs. 2, 3, 4 and 5; each figure represents the projection of the clusters along two dimensions corresponding to the following pairs of attributes **AvgDelay** and **Par3**, **AvgGrade** and **Par3**, **AvgDelay** and **Par4** and, finally, **AvgGrade** and **Par4**. The centroids of the resulting clusters are shown in Table 8, which also contains the average values relative to the full data set.

Table 8. The centroids of clusters in Figs. 2, 3, 4, 5.

Attribute	Full data	Cluster0	Cluster1
AvgGrade	25.31	25.85	24.58
AvgDelay	2.61	1.8	3.68
Par*1*	70.86	77.74	61.67
Par*2*	72.23	82.19	58.94
Par*3*	84.51	90.25	76.86
Par*4*	72.03	74.67	68.5
Par*5*	76.02	80.83	69.61

The cluster number 0, which correspond to 88 blue stars in the figures, contains the courses which students took with *small* delay and that they evaluated positively. On the other hand, cluster number 1, corresponding to 66 red stars, contains those courses which students took with a large delay and that they evaluated less positively. We observe that the centroids of the two clusters are very close relative to the attribute **Par4** which concerns classrooms and equipment. This is also evidenced from Figs. 4 and 5, where the blue and red stars are less separated than those in Figs. 2 and 3. The Pearson's correlation corresponding to these clusters is equal to -0.35. We obtained an improvement by excluding the attribute **Par4** from clustering, in fact in this case we find a correlation equal to -0.51. In general, our tests evidenced that the paragraphs evaluations which are more correlated with students results regard attributes **Par*2*** and **Par*3***, that is, those concerning the course organization and the teacher. We point out that the value $K = 2$ gave the best results in terms of correlation.

Among the courses considered in the previous data set, we selected those evaluated all seven years, for a total of sixteen courses, some in Mathematics and others in Computer Science. This time we are interested in analysing data organized as in Table 3, by considering the evaluation of a particular paragraph over the years. The aim was to find if there are similar behaviors among courses, that is, if we can classify courses according to their evaluations. We performed several tests, by choosing a paragraph at a time. For each choice of attributes, we applied the `K-means` algorithm with the Manhattan distance to identify the

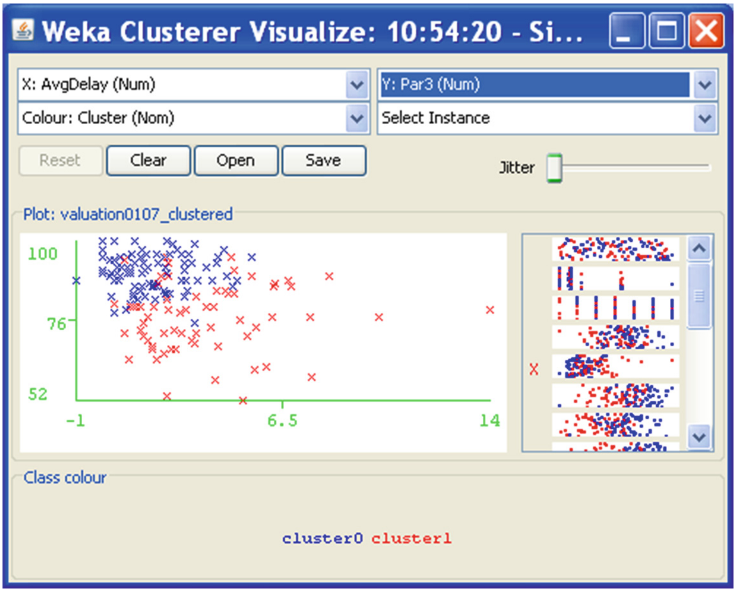

Fig. 2. Clusters of Table 8 with **AvgDelay** and **Par3** in evidence (Color figure online).

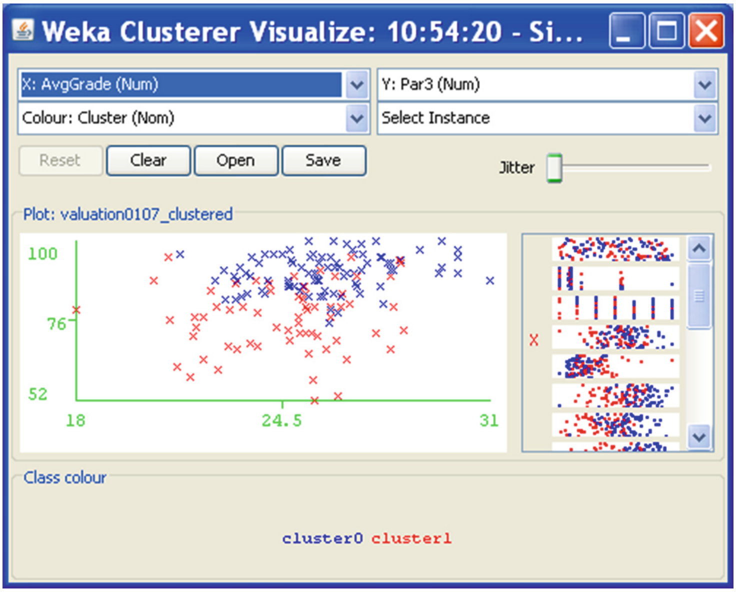

Fig. 3. Clusters of Table 8 with **AvgGrade** and **Par3** in evidence (Color figure online).

clusters; also in this case we computed the Pearson's correlation by using the proximity and incidence matrices.

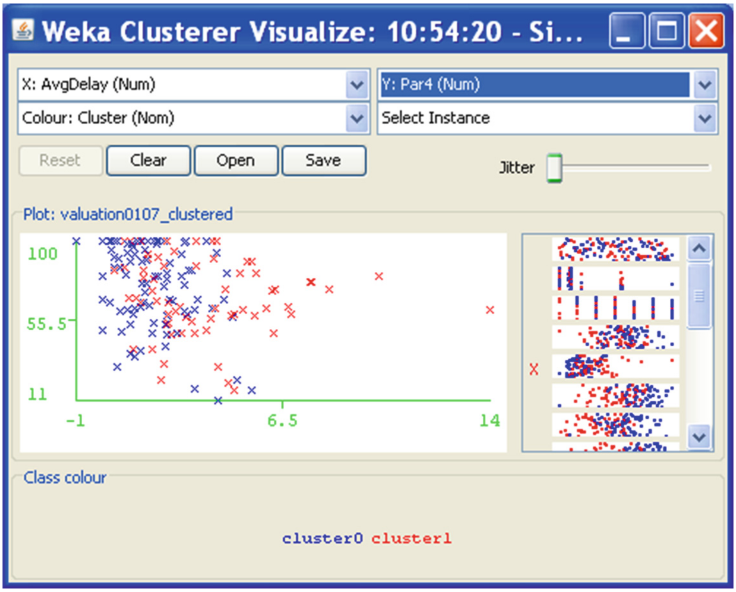

Fig. 4. Clusters of Table 8 with **AvgDelay** and **Par4** in evidence (Color figure online).

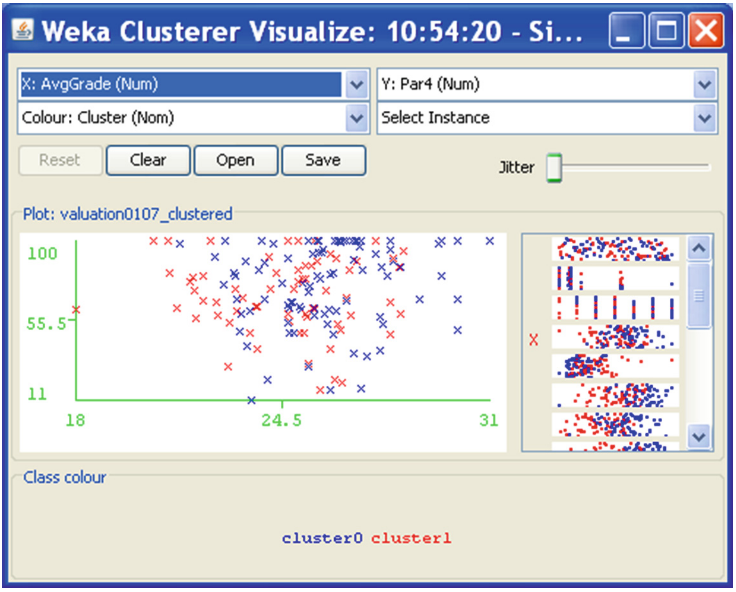

Fig. 5. Clusters of Table 8 with **AvgGrade** and **Par4** in evidence (Color figure online).

Figure 6 illustrates the result of `K-means` with $K = 2$, Manahattan distance and **Par** *2* **(2001)**, **Par** *2* **(2002)**, $\cdots$, **Par** *2* **(2007)** as clustering attributes.

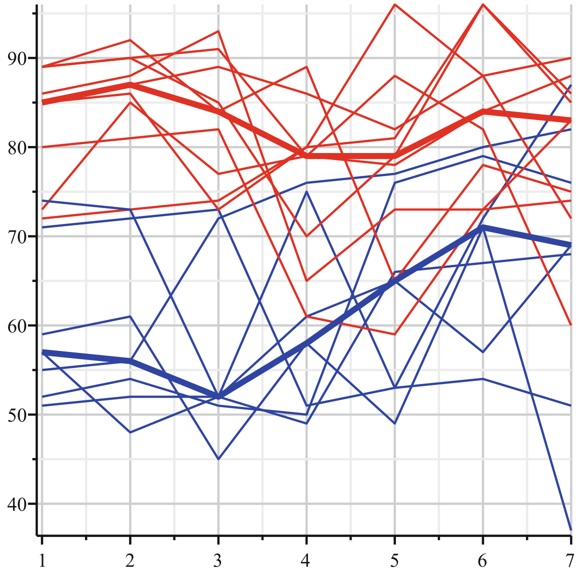

Fig. 6. Clusters of Table 9 with centroids in evidence: each line represents the percentage of positive evaluations about the organization of a course (paragraph 2) over the years 2001–2007 (Color figure online).

The points defining the centroid trajectories of the resulting clusters are shown in Table 9, which also contains the median values relative to the full data set. The Pearson's correlation corresponding to these clusters is equal to -0.64.

The figure puts well in evidence that the courses are divided into two clusters with well distinct centroids. The red cluster contains courses that have been evaluated better over the years while the blue cluster corresponds to courses that students rated worse. What is interesting, though not surprising, is that all courses in the red cluster are Computer Science courses while the blue cluster contains many Mathematics courses. We highlight that the centroids show rather clearly the behavior of the assessment over the years. In particular, the evaluation of the courses in the blue cluster has improved over the years while that of courses in the red cluster has remained more stable.

Also in this case the best results in terms of correlation were found with $K = 2$; however, with $K = 4$ we found the courses rated worse distributed into two clusters, one of which contains only the Mathematics courses. The corresponding centroid illustrates a gradual improvement of the assessment for this type of courses during the years under examination.

Figure 6 illustrates the results of the clustering corresponding to Table 9 with a graphical representation that refers to the natural temporal order, i.e., the values on the horizontal axis ranges from 2001 to 2007. However, since we used the Manhattan distance, the sum of the vertical distances between the ordinates, the trajectories could be represented by taking into account a different time ordering

Table 9. The points defining the centroid trajectories of clusters in Fig. 6.

Attribute	Full Data	Cluster0	Cluster1
Par2(2001)	73.5	85	57
Par2(2002)	77	87	56
Par2(2003)	73.5	84	52
Par2(2004)	72.5	79	58
Par2(2005)	74.5	79	65
Par2(2006)	78.5	84	71
Par2(2007)	75.5	83	69

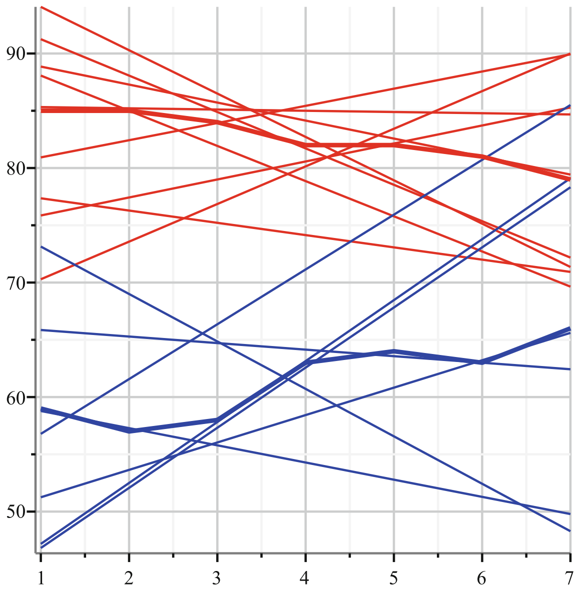

Fig. 7. The regression lines for the trajectories of Fig. 6 with the centroids of the corresponding clustering in evidence.

and the results of clustering would not change. To keep into consideration the temporal values, we can think to compute the regression line corresponding to the points that identify each broken line and then perform clustering with the Manhattan distance. Figure 7 shows the results of clustering with $K = 2$ on the regression lines corresponding to the broken lines of Fig. 6, with the centroids in evidence: the corresponding courses match exactly the same subdivision of the clustering shown in Fig. 6. This new representation provides a better highlight of the courses evaluation over the years and confirms the previous remarks on Fig. 6. Moreover, the analysis could be deepened by clustering courses with respect to the slope values of the regression lines, thus grouping together courses with the same evaluation trend.

4 Conclusions and Future Work

The results of the previous sections show, in a formal way with data mining techniques, that there is a relationship between the evaluation of the courses from students and the results they obtained in the corresponding examinations. In particular, the analysis performed on data related to the Computer Science degree program under examination illustrates that the courses which received a positive evaluation correspond to exams in which students obtained a good average mark and that they took with a small delay. Conversely, the worst evaluations were given to those courses which do not match good achievements by students.

The analysis based on clustering with Manhattan distance allows us to classify courses according to the assessment received by students and can highlight some regularities that emerge over the years or points out some trend reversals due to changes of teachers. In the Computer Science degree program just considered, for example, we observe the trend to give not so good evaluation to Mathematics courses. Results of this type point out a critical issue in the involved courses and can be used to implement improvement strategies.

The correlations highlighted in this paper could be considered predictable, however, we wish to emphasize that our analysis refers to the courses evaluation that students make before taking the exams and before knowing their grades. In fact, as already observed, the evaluation module is given to students before the end of the course. Surely, there is the risk that their judgment is influenced by the inherent difficulty of the course or by the comments made by students of the previous years. To this purpose, it is important that during the module compilation the teacher explains that a serious assessment of the course can increase the quality level of the involved services. Students represent the end-users as well as the principal actors of the formative services offered by the University and the measure of their perceived quality is essential for planning changes. However, the results of courses evaluation should always be considered in a critical way and should not be used for simplifying the contents and to get best ratings. Moreover, we wish to point out again the formal proof of the results with a data mining approach. From this point of view, other cluster algorithms and goodness evaluation measures could be used; for example, an interesting future work could include the use of hierarchical clustering (see, e.g., [14]).

In general, many other factors should be considered for evaluating courses and student success, as addressed in [11]. The approach used in this work could be refined and deepened if it was possible to identify the students involved in the courses evaluation in order to connect properly the results of the evaluation with those of exams. Moreover, it would be interesting to connect the assessment of students with other information such as the gender of students and teachers or the kind of high school attended by students. Starting from the academic year 2011/2012, the University of Florence began to manage on line the evaluation module described in Sect. 2. Therefore, in a next future, it might be possible to proceed in this direction, taking into account appropriate strategies to maintain privacy.

An interesting additional source of information could be given by social media sites, such as *Facebook* or *Twitter*, used by students to post comments about courses and teachers. It would be useful to link this information with the results of students and their official evaluations about teaching, in order to take into account more feedbacks. In such a context, it might be interesting to use text mining techniques to classify the student comments and enrich the database for an analysis similar to that illustrated in this work.

References

1. Progetto SISValDidat. https://valmon.disia.unifi.it/sisvaldidat/unifi/index.php
2. Campagni, R., Merlini, D., Sprugnoli, R.: Analyzing paths in a student database. In: The 5th International Conference on Educational Data Mining, Chania, Greece, pp. 208–209 (2012)
3. Campagni, R., Merlini, D., Sprugnoli, R.: Data mining for a student database. In: 13th Italian Conference on Theoretical Computer Science, ICTCS 2012, Varese, Italy (2012)
4. Campagni, R., Merlini, D., Sprugnoli, R.: Sequential patterns analysis in a student database. In: ECML-PKDD Workshop: Mining and Exploiting Interpretable Local Patterns (I-Pat 2012), Bristol (2012)
5. Campagni, R., Merlini, D., Sprugnoli, R., Verri, M.C.: Comparing examination results and courses evaluation: a data mining approach. In: Didamatica 2013, Pisa, Area della Ricerca CNR, AICA, pp. 893–902 (2013)
6. Campagni, R., Merlini, D., Verri, M.C.: Finding regularities in courses evaluation with k-means clustering. In: CSEDU 2014, Barcelona, Spain, Scitepress, vol. 2, pp. 26–33 (2014)
7. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
8. Liao, T.W.: Clustering of time series data: a survey. Pattern Recognit. **38**(11), 1857–1874 (2005)
9. MacQueen, J.: Some methods for classifications and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
10. Romero, C., Romero, J.R., Luna, J.M., Ventura, S.: Mining rare association rules from e-learning data. In: The 3rd International Conference on Educational Data Mining, pp. 171–180 (2010)
11. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. **40**(6), 601–618 (2010)
12. Romero, C., Ventura, S.: Data mining in education. WIREs Data Min. Knowl. Disc. **3**, 12–27 (2013)
13. Romero, C., Ventura, S., García, E.: Data mining in course management systems: moodle case study and tutorial. Comput. Educ. **51**(1), 368–384 (2008)
14. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Boston (2006)
15. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)