Finding the best paths in university curricula of graduates to improve academic guidance services

Individuare i migliori percorsi di carriera universitaria per migliorare l'efficacia dei servizi d'orientamento d'Ateneo

Silvia Bacci and Bruno Bertaccini

Abstract Within the more general quality assurance pathways undertaken by the universities, inbound and on going guidance activities are assuming an increasingly strategic role with the aim to reduce the dropout rate and the time to qualification. In this contribution, the usefulness of some typical data mining solutions is evaluated in the context of students' careers analysis. More in detail, an analysis of graduates' careers paths is proposed, mainly based on the application and comparison of clustering procedures. With our proposal we aim at identifying career paths that are particularly virtuous in terms of average scores and time to qualification. Such type of information can be used by the university management for planning the career paths of freshmen.

Abstract *All'interno dei più generali percorsi di assicurazione della qualità intrapresi dagli Atenei, le attività di orientamento in ingresso e itinere stanno assumendo un ruolo sempre più strategico in ottica di riduzione del tasso d'abbandono e contenimento dei tempi di conseguimento del titolo. Questo lavoro intende valutare l'applicabilità di alcune soluzioni di data mining nel campo dell'analisi dei dati di carriera dei laureati; in particolare i percorsi scelti da coloro che hanno completato gli studi saranno analizzati attraverso alcune tecniche di clustering per l'identificazione di carriere virtuose (in termini di votazione media e tempo richiesto per il completamento del percorso) che possano essere proposte dagli organi di governo dei corsi di studio quale modello di riferimento per i nuovi iscritti.*

Key words: academic guidance services, cluster analysis, data mining, education, hidden Markov models, mixture models

---

Silvia Bacci
Dipartimento di Economia, Università di Perugia, Via A. Pascoli 20, 06123 Perugia (IT), e-mail: silvia.bacci@unipg.it

Bruno Bertaccini
Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti", Università degli Studi di Firenze,Viale Morgagni 59, Firenze (IT), e-mail: bruno.bertaccini@unifi.it

1 Introduction

The Italian university system is characterized by some peculiarities, *in primis* that students can freely decide when taking an exam and the specific sequence of enrolled exams. The main relevant consequence of this organizative approach are the long times to qualification. In such a context identifying career paths that are particularly virtuous in terms of average grades and time to qualification as well as the exams representing "bottlenecks" in the career flows is especially relevant for the university management in order to plan the career paths of freshmen and to improve the academic guidance services.

In this contribution we aim at studying the sequences of exams taken by a cohort of graduated students, through the comparison of some clustering approaches.

2 Data

The analysis here proposed is based on a cohort of 189 students in Business Economics at the University of Florence that enrolled the degree course in year 2012 and completed it within year 2017. In Figures 1 and 2 the sequences of first-year exams are shown, by average grade (low-high; Figure 1) and time to qualification (low-high; Figure 2). In both cases, attention is captured by differences in the observed sequences: in particular, the first-year exams that are more problematic in terms of tendency to postpone are Private Law and Mathematics.

3 Clustering approaches

Behavioral homogeneity of the cohort of students at issue is investigated through some clustering approaches:

Hierarchical cluster analysis. This a well-known approach that does not require any model specification, even if some choices are necessary (e.g., number of clusters, grouping method, type of distance). The analysis is performed on the exams' cumulative average score and on the cumulative time to qualification. This approach suffers for the presence of sparse data due to a large number of exams that are taken only by few students: as a consequence a preliminary cleanup of data is necessary.

Latent class model-based cluster analysis [4]. This is the simplest model-based clustering approach. The analysis is performed on the same variables as the previous approach. This approach has the same drawbacks as the hierarchical clustering.

Mixture Luce-Plackett model-based analysis [5]. Differently from the two above approaches, this approach works on the students' partial ranking of exams, that

is for each student the corresponding sequence of exams is formulated in terms of a rank, where missing values correspond to exams that are not in the student's degree curriculum. In such a way the sequence of exams is explicitly taken into account. As main result, this approach provides a measure of liking toward each exam, separately for each cluster. The main practical drawback of this approach is that the clustering process does not account for the exam grades and, only partially, for the time gaps between exams, so that clusters are likely to overlap a lot in terms of exam grades and times to qualification. Moreover, the liking measures tend to assume high values for the very first enrolled exams and values around zero for the following exams, such that differences between clusters in terms of sequences may not be well defined.

Mixture Hidden Markov (MHM) model-based cluster analysis [2]. Similarly to the mixture Luce-Plackett model-based approach, this approach directly models the sequences of exams. More in details, a multiple multichannel MHM model is specially suitable for data at issue, as it will be cleared in the next section. Terms "multiple" and "multichannel" refer to the presence of more than one individual (i.e., cohort of students) and of more than one sequence of data for each individ-

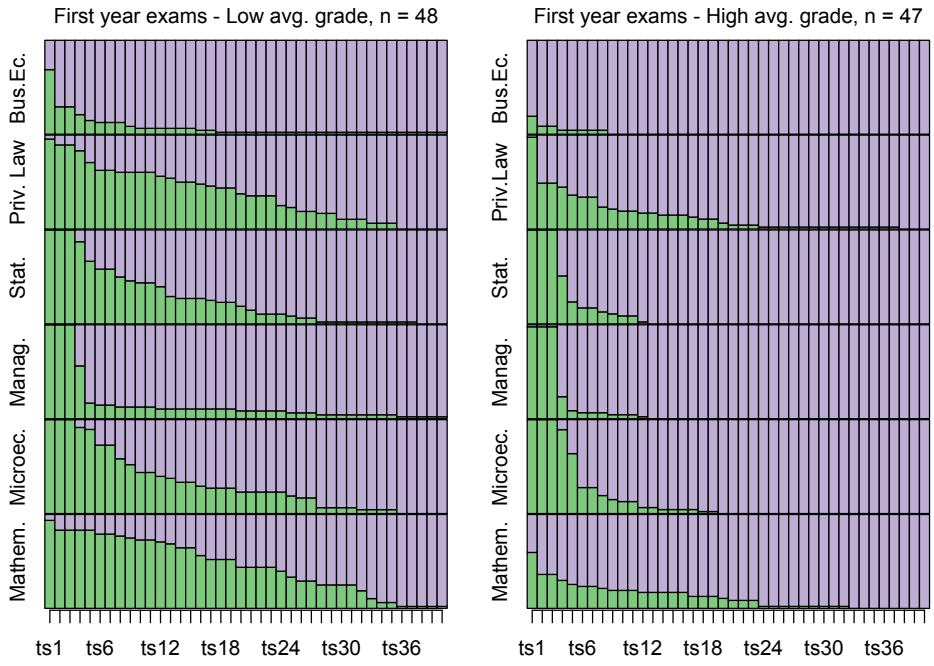

Fig. 1 Sequences of first-year exams by average grade: average grade less than first quartile (left panel) and average grade greater than third quartile (right panel). Legend: green = exam not taken; purple = exam taken.

ual (i.e., one sequence for each exam), respectively. In practice, the MHM model presents some specific advantages with respect to the other approaches: (*i*) the presence of sparse data is not a problem (exams that are not present in the degree curriculum of a student correspond to a sequence of 0s), (*ii*) the observed sequences of exams are explicitly modelled and individuals are clustered on the basis of these sequences, (*iii*) in addition to the other finite mixture approaches, it allows us an in-depth analysis of every exam within each cluster, such that weaknesses of the different typologies of students are highlighted, (*iv*) differently from the other approaches, it allows us to enclose in the analysis students that dropped out or that did not yet finished the exams of the degree course.

3.1 Mixture Hidden Markov model

HM models [3, 1] represent a nice frame to analyse sequence data. In such a context, a sequence of 0s and 1s represents the observed states, which are interpreted as

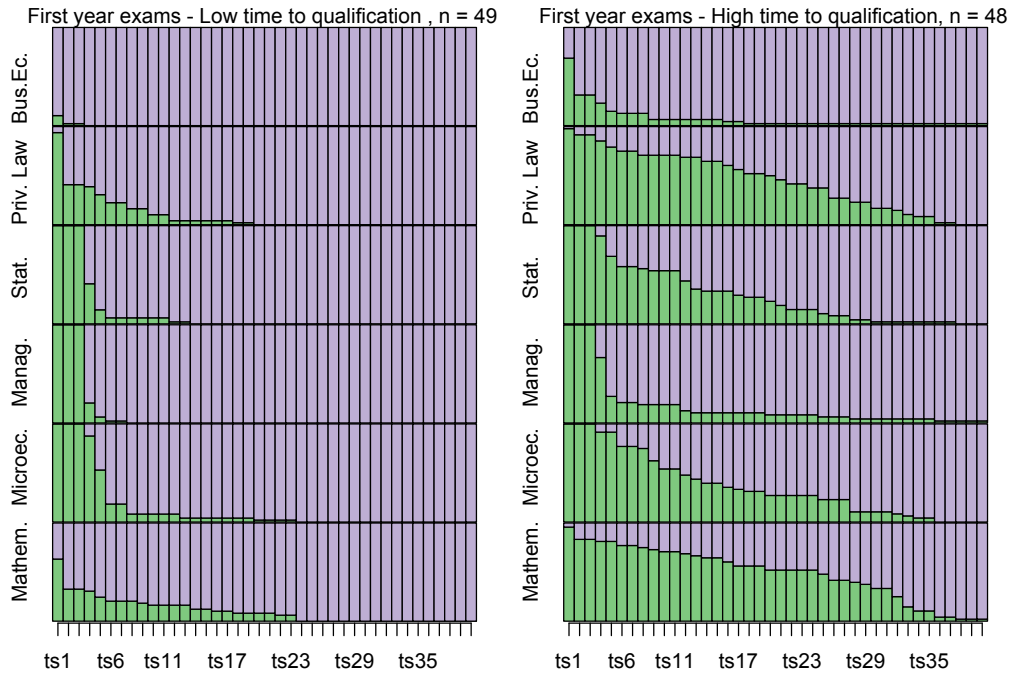

Fig. 2 Sequences of first-year exams by time to qualification: time less than first quartile (left panel) and time greater than third quartile (right panel). Legend: green = exam not taken; purple = exam taken.

probabilistic manifestations of a certain number of unobservable (i.e., hidden or latent) states. In our contribution, we assume that student i in any time point may belong to one of two hidden states: state $u_{it} = 1$ denotes a low propensity to take exams at time t and state $u_{it} = 2$ denotes a high propensity to take exams at time t. As usual in the HM models, students can move from a state to another one. In addition, we introduce the observed state $y_{itj} = y$, with $y = 1$ if student i takes exam j at time t or before, and $y = 0$ if student i did not yet taken exam j at time t; $j = 1,\ldots,J$, with J denoting the total number of exams (it does not matter how many students choose an exam for their own degree curriculum), and $t = 1,\ldots,T$ with T denoting the length of any sequence and corresponds to the number of exam sessions scheduled in the years 2012-2017.

Hence, we have two types of vectors for each student: vector $\mathbf{u}_i = (u_{i1},\ldots,u_{it},\ldots,u_{iT})$ of hidden state sequence and vector $\mathbf{y}_{ij} = (y_{i1j},\ldots,y_{itj},\ldots,y_{iTj})$ of observed state sequence; note that we have one vector $\mathbf{y}_{ij}$ for each exam.

The probability of observed sequence of data is formulated according to a time-homogeneous multivariate HM model:

$$p(\mathbf{Y}_{ij} = \mathbf{y}_{ij}) = \sum_{u=1}^{2} p(\mathbf{y}_{ij}|\mathbf{u}_i)p(\mathbf{u}_i) = \sum_{u=1}^{2} \left[p(y_{i1j}|u_{i1})p(u_{i1}) \prod_{t=2}^{T} p(y_{itj}|u_{it})p(u_{it}|u_{i,t-1}) \right], \tag{1}$$

with: $p(y_{itj}|u_{it})$ conditional probability of observed state given the hidden state (emission probability), $p(u_{i1})$ initial probability of starting from hidden state u_{i1}, and $p(u_{it}|u_{i,t-1})$ transition probability of moving from hidden state $u_{i,t-1}$ to hidden state u_{it}. This model is usually estimated through the maximization of the log-likelihood function, using the forward-backward algorithm.

A generalization of the HM model is represented by the MHM model, which is based on the assumption that the population is composed by homogenous groups of individuals and each of these groups follow a specific HM model. In such a context, model in equation 1 modifies as

$$p(\mathbf{Y}_{ij} = \mathbf{y}_{ij}) = \sum_{k=1}^{K} \pi_k \left\{ \sum_{u=1}^{2} p(\mathbf{y}_{ij}|\mathbf{u}_i)p(\mathbf{u}_i) \right\}, \tag{2}$$

where π_k denotes the prior probability that the sequence of observed states of an individual belongs to cluster k.

To synthetize, the mixture part of the model allows us to cluster students in groups that are homogenous in terms of observed sequences of exams and propensity to take exams. In this way a classification of students is obtained, which is comparable with those obtained in the previously described approaches. In addition, the HM part of the model allows us for an in-depth analysis of the performance of students on single exams.

4 Main results

To define homogenous sub-groups of students, we selected a number of clusters equal to three and we applied the clustering procedures above described. In Table 1 the results in terms of average grade, average time to qualification and cluster size are illustrated.

Table 1 Average grade and average time to qualification by cluster and clustering approach. In bold the best performances, in italic the worst performances.

Approach	Variables	Cluster 1	Cluster 2	Cluster 3
Hierarchical clustering				
	avg. grade (out of 30)	**26.14**	*23.69*	24.99
	avg. time (days)	**1037.04**	*1598.90*	1262.98
	# of students	54	48	87
Latent class clustering				
	avg. grade (out of 30)	25.35	*24.05*	**26.74**
	avg. time (days)	1171.83	*1436.12*	**1079.16**
	# of students	59	93	37
Mix. Luce-Plackett clust.				
	avg. grade (out of 30)	**25.59**	24.97	*24.13*
	avg. time (days)	**1305.47**	1382.06	*1447.91*
	# of students	30	148	11
MHM clustering				
	avg. grade (out of 30)	25.25	*23.99*	**25.81**
	avg. time (days)	1324.83	*1577.25*	**1219.21**
	# of students	46	68	75

It is worth to be noted that all approaches provide a cluster of best performers (in bold) and a cluster of worst performers (in italic) with respect to both criteria of average grade and time to qualification. The clustering approach based on partially ranked data is the least satisfactorily as concerns the level of separability among clusters.

Additional details about the cluster characteristics are provided by the estimated parameters of the MHM model. In Table 2 the emission probabilities of the first-year exams are shown, which describe the probability of taking an exam - in any time point - given the hidden state, that is, $p(Y_{itj} = 1|u_{it})$; these probabilities are cluster-specific. For the sake of clarity we remind that state 1 denotes a low propensity to take exams and state 2 denotes a high propensity to take exams; then, the lower $p(Y_{itj} = 1|u_{it})$, the higher the tendency to postpone exam j.

As shown in Table 1, the worst performers are allocated in cluster 2. In more detail 2, students of cluster 2 belonging to state 1 have a high tendency to postpone Mathematics and Private Law (emission probabilities equal to 26.6% and 33.3%, respectively), followed by Microeconomics (46.0%) and Statistics (57.8%). Mathematics and Private Law represent bottlenecks also for the colleagues in state 2 (emission probabilities equal to 79.1% and 83.8%, respectively).

Table 2 MHM model: Estimated emission probabilities by cluster (only first-year exams)

Exam	Hidden state	Cluster 1	Cluster 2	Cluster 3
Business Economics				
	State 1	0.893	0.925	0.983
	State 2	0.984	1.000	1.000
Private Law				
	State 1	0.636	0.333	0.655
	State 2	0.962	0.838	0.987
Statistics				
	State 1	0.609	0.578	0.740
	State 2	0.991	0.963	0.998
Management				
	State 1	0.704	0.763	0.804
	State 2	0.977	1.000	1.000
Microeconomics				
	State 1	0.493	0.460	0.677
	State 2	0.956	0.929	1.000
Mathematics				
	State 1	0.438	0.266	0.783
	State 2	0.885	0.791	0.989

On the opposite, the best performers are allocated in cluster 3. In such a cluster the main problems are observed for Private Law and Microeconomics: for both the exams the emission probabilities for state 1 are strongly smaller than 1 (65.5% for Private Law and 67.7% for Microeconomics).

Finally, cluster 1 lies in an intermediate position with respect to clusters 2 and 3 for all (the first-year) exams, with the exception of Business Economics, for which a certain tendency to postpone is observed for both the hidden states (emission probabilities are higher than those of cluster 2).

A synthetic representation of the three clusters is provided by Figure 3, where the estimated sequence of hidden states is displayed for each cluster. It is worth to be noted that around the centre of the time line cluster 2 presents a tail for state 1 that is heavier with respect to cluster 1 and, mainly, cluster 3.

5 Conclusions

The illustrated approaches of cluster analysis, with a special attention for the mixture Hidden Markov model, represent useful instruments for the academic management to detect critical exams for specific sub-groups of students and, more in general, to plan the career paths of freshmen and to improve the academic guidance services.

For the future development of this contribution we intend to extend the analysis to the entire cohort of students, including students that dropped out from the university

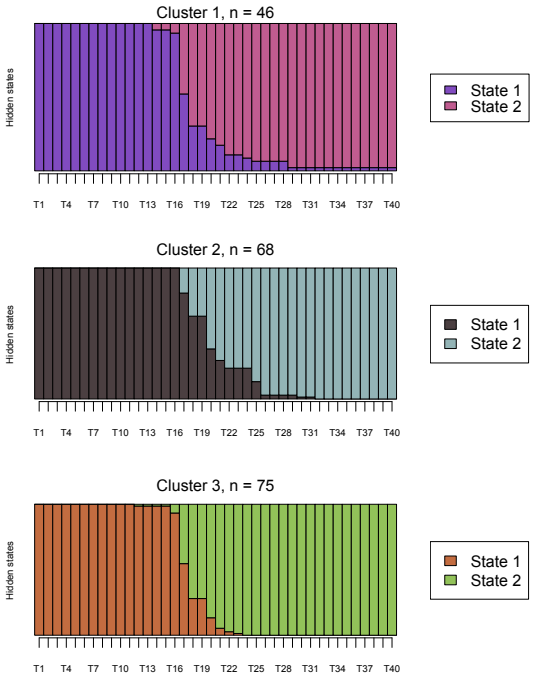

Fig. 3 MHM model: Sequences of hidden states by cluster.

and students that have to finish the exams of the degree course. We also intend to extend the analysis to account for individual characteristics (e.g., type of high school, high school graduation mark).

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov models for longitudinal data. Chapman & Hall/CRC, Boca Raton, FL (2012)
2. Helske, S., Helske, J.: Mixture Hidden Markov models for sequence data: the seqHMM package in R. Available via https://arxiv.org/pdf/1704.00543.pdf. Cited 25 Apr 2018
3. Zucchini, W., MacDonald, I. L.: Hidden Markov Models for Time Series: an Introduction using R. Springer, New York (2009)
4. McLachlan, G., Peel, D.: Finite mixture models. Wiley, New York (2000)
5. Mollica, C., Tardella, L.: PLMIX: An R package for modeling and clustering partially ranked data. Available via https://arxiv.org/pdf/1612.08141.pdf. Cited 25 Apr 2018