CLASSIFYING NORTHERN ITALIAN STUDENTS IN THEIR TRANSITION TO MASTER DEGREE

Alfonzetti Giuseppe [1], Grassetti Luca[1] and Rizzi Laura[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: giuseppe.alfonzetti@uniud.it, luca.grassetti@uniud.it, laura.rizzi@uniud.it)

ABSTRACT: The university students' behaviour represents a relevant field of study from the management point of view. Given the availability of large administrative data on students' careers, the chance to discover students' profiles in terms of behavioural patterns could be interesting. However, the identification of students' clusters that are informative, feasible and robust at the same time could be complex. The present work aims to define a feasible student clusterisation, adopting an empirical algorithm to treat mixed data and large sample sizes and borrow the syncytial clustering idea developed in the machine learning framework. The proposal is a generalisation of the original algorithm to mixed data cases. Finally, the importance of finding a prototype of students' behaviours is discussed.

KEYWORDS: Two-stage clustering, Hierarchical clustering, Partitioning clustering, Student profiling, Students careers

1 Introduction and aims

A relevant task in education is analysing students' behaviour during their careers. In particular, finding some structured patterns helps defining actions to optimise the supply and organisation of second-level university (Masters) courses and, more in general, of the third-level educational system. Analysing the identified patterns makes it possible to point out both opportunities and shortages in the university education provision.

The availability of individual-level administrative data and their integration with contextual information on the university students (such as the secondary school track) can be considered fundamental for the development of a detailed data mining process able to extract the relevant signals from the humongous set of available data. In particular, the adoption of feasible and robust clustering behavioural has a crucial role in the detection of students' prototype characteristics that can be relevant in providing better services and management.

The dataset considered in the present study comes from a population database regarding Italian university students. We decide to focus on the subset of students from the North of Italy. Even in this restricted framework, the size of the dataset is very large (close to 400,000). Clustering under these settings is not straightforward. The classical hierarchical clustering is unfeasible given the size of the distance matrix, and the partitive solutions (k-means and other machine learning algorithms) are typically difficult to manage. For instance, determining the optimal number of groups or developing a diagnostic for the obtained solution is complex and time-consuming.

In the present work, we propose a solution which inherits the two-stage clustering idea of alternating a hierarchical and a partitive algorithm to reach a more interpretable solution. The typical two-stage clustering algorithm involves a hierarchical clustering first and a partitive approach then, which results unfeasible in large dataset settings. Our approach, instead, connects to the syncytial algorithms outlined in Peterson *et al.* (2018) and Almodóvar-Rivera & Maitra (2020), where the output of a partitive algorithm is used as input for a second agglomerative step.

The main contribution of the present work is the definition of a practical tool for students' prototype recognition based on a syncytial clustering algorithm accommodating mixed data types. The proposal provides enhanced interpretability compared with the classical unsupervised solutions usually adopted in large dataset frameworks.

The structure of the paper is as follows. First, Section 2 details the proposed methodology and presents the data. Then, Section 3 summarizes the results of the empirical analysis.

2 Data and Methods

The characteristics of the Northern Italian students are collected from the Italian Ministry of University's administrative databases (Mobysu.it, 2016, update 2022). In this preliminary analysis, variables considered in the clustering procedures are only some available measures of students' career performance. in particular, the analysis involves a set of dummy variables identifying: Italian students, private secondary schools, and public universities. In addition, factors for the kind of secondary school attended and the gender of students are also included. Finally, some quantitative variables are introduced, including students' age at bachelor's degree, bachelor's course duration, diploma and bachelor's degree marks, years between diploma and bachelor's degree enrolment, and the distance between the first-level university and secondary school municipalities.

As anticipated, the proposed methodology can be framed as a syncytial clustering algorithm (Peterson *et al.*, 2018; Almodóvar-Rivera & Maitra, 2020). Furthermore, due to the mixed nature of our dataset, where many dummy variables and factors are observed along with some numerical measures, the two steps accommodate algorithms suitable to deal with mixed data. Specifically, the first step implements a k-prototypes clustering (Huang, 1998; Szepannek, 2018), while the second one is a hierarchical clustering procedure based on Gower's distance (Gower, 1971; Maechler *et al.*, 2022). It is worth stressing that the proposed method enjoys easy identification of the optimal clustering solution, along with enhanced interpretability and a robust cluster selection.

3 The empirical analysis

In this section, we analyse a specific clustering solution focusing on students' prototyping and interpret the obtained results.

The optimal number of clusters selected by the procedure is four, as the dendrogram in Figure 1 suggests. Figure 1 also shows the flow of the students through their career characteristics represented for the different clusters. First, Gender is used to describe the population, and then the student patterns are observed over the schooling period. The choice of the kind of school is the first step (for the sake of readability, we reduced the factor to a dummy variable identifying the Liceo secondary school). A fundamental variable affecting the students' career is the diploma mark which is here classified into three levels (≤ 80, $81-90$, and $91+$). The choice of the University is also linked to the opportunity to move from home. The distance variable is a categorized version of the Euclidean distance between the first-level university and secondary school municipalities (0, $1-50$, $51-100$, $101+$ km). The Degree Age is a categorical variable collecting the regular, young, and Late students defined based on their age at bachelor's degree (≤ 22, $23-25$, and $26+$). In the plot, the final collected aspect before the master's degree choice is defined here by Degree Mark, a factor presenting low, mid-low, mid-high, and high levels (≤ 90, $91-100$, $101-110$, $summacumlaude$). All these characteristics flow into the choice of continuing the career or dropping from the university system.

We finally link the students' prototypes to the dropout phenomenon as an example of university outcomes. However, while the clustering approach allows to point out some prototypes of students on the basis of their high-school and first-level university tracks, the identified groups are not connected with

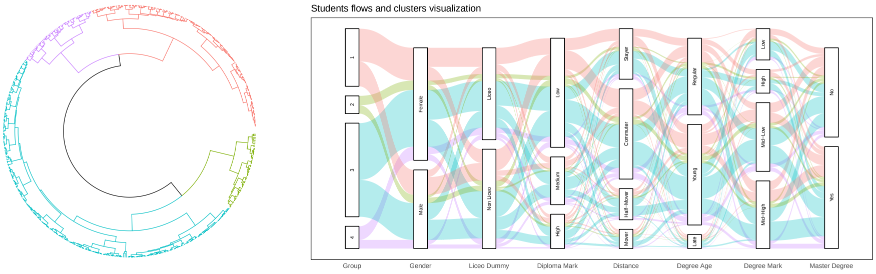

Figure 1. *Behavioural flow of students clustered in the four groups discussed in the text: From birth to master's degree enrolment.*

the propensity to enrol on a master's program. Other factors, such as the field of study, family background and social-economic aspects, may play a relevant role in the choice of transition.

References

ALMODÓVAR-RIVERA, I.A., & MAITRA, R. 2020. Kernel-estimated nonparametric overlap-based syncytial clustering. *The Journal of Machine Learning Research*, **21**(1), 4808–4861.

GOWER, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

HUANG, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, **2**(3), 283–304.

MAECHLER, M., *et al.* 2022. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4.

MOBYSU.IT, DATABASE. 2016. *Database Mobysu.it degli studi universitari in Italia.* In Protocollo di ricerca MIUR-Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II. Fonte dei dati: ANS-MIUR/CINECA.

PETERSON, A.D., GHOSH, A.P., & MAITRA, R. 2018. Merging K-means with hierarchical clustering for identifying general-shaped groups. *Stat*, **7**(1), e172.

SZEPANNEK, G. 2018. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal*, **10**(2), 200–208.