Cluster Analysis, a powerful tool for data analysis in Education

Vasconcelos, Rita
Universidade da Madeira, Department of Mathematics and Engeneering
Caminho da Penteada
9000-390 Funchal, Portugal
E-mail: rita@uma.pt

Baptista, Márcia
Direcção Regional de Saúde Pública
Rua das Pretas
9000 Funchal, Portugal
E-mail: marciabatista12@gmail.com

1. Introduction

A database was created after an inquiry to 14-15 - year old students, which was developed with the purpose of identifying the factors that could socially and pedagogically frame the results in Mathematics. The data was collected in eight schools in Funchal (Madeira Island), and we performed a Cluster Analysis as a first multivariate statistical approach to this database. We also developed a logistic regression analysis, as the study was carried out as a contribution to explain the success/failure in Mathematics.

As a final step, the responses of both statistical analysis were studied.

2. Cluster Analysis approach

The questions that arise when we try to frame socially and pedagogically the results in Mathematics of 14-15 - year old students, are concerned with the types of decisive factors in those results. It is somehow underlying our objectives to classify the students according to the factors understood by us as being decisive in students' results. This is exactly the aim of Cluster Analysis.

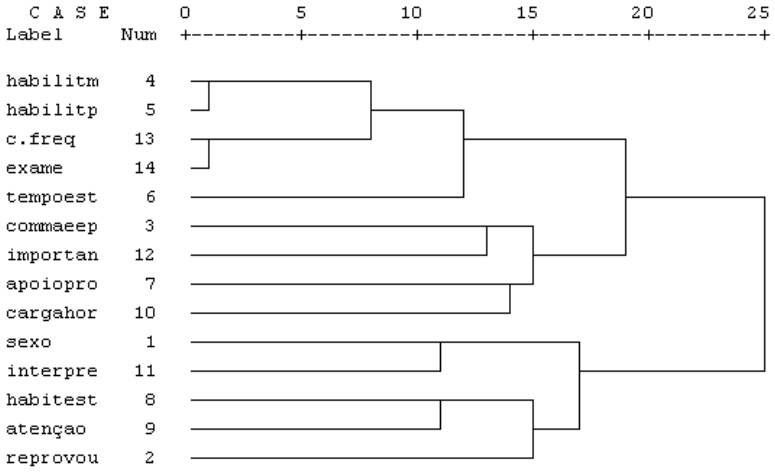
The hierarchical solution that can be observed in the dendogram presented in the next page, suggests that we should consider the 3 following clusters, since the distances increase substantially after it:

Variables in **Cluster1**: mother qualifications; father qualifications; student's results in Mathematics as classified by the school teacher; student's results in the exam of Mathematics; time spent studying. This result enhances the influence of parents' qualifications in the results in Mathematics and the time spent on studying, although the last one is less influent. It shows the influence of parents' qualifications in their children results in Mathematics and in the time spent on studying.

Variables in **Cluster2**: the student lives with his parents; what students think as more important for improving results in Mathematics; special aid of Mathematics school teachers; too many hours attending school classes. Variables in this cluster reflect the family tconversations about Mathematics classes.

Variables in **Cluster3**: sex; lack of attention/concentration in class; studying unwontedness; difficulty in interpretation; student has already failed one school year. Not surprisingly this variables classify the students in a similar way, since boys and girls typically have different results in these variables.

Dendogram for the variables (complete linkage of the furthest neighbor)



A Cluster Analysis applied to students, using a k-means algorithm and choosing the first 4 interpretable clusters, shows that 52.5% of the students in Cluster3 have excellent results in Mathematics. The percentage of failure is 32% in Cluster1, 25% in Cluster2, 0.06% in Cluster3 and 29% in Cluster4. Also, the centers of the 4 clusters for each variable, show that the parents' background education is the variable that absolutely distinguishes between the 4 clusters. Students in Cluster1 have parents with low education background, are the only ones that complain about spending too many hours attending to classes and, also, are the only ones who suggest that a small number of pupils in each class, and classes with more stimulating subjects different from those that are taught now, would improve students' results in Mathematics.Conclusions on students' clusters are supported by variable clusters and these analyses absolutely help each other. We can reach the goals we set to ourselves when analyzing a database in Education, through a Cluster Analysis, going deep in students' behaviour and in its relation with the variables in the database.

2. Logistic Regression Analysis approach

The dependent variable in this approach is failure/success in Mathematics and the codings of the categorical variables in the final model can be found in the following table.

The codings of the categorical variables in the final model

		Frequency			
Time spent studying (per week)	0h	45	1,000	,000	,000
	1h	119	,000	1,000	,000
	2h	46	,000	,000	1,000
	More than 2h	21	,000	,000	,000
Mother's qualification	none	4	1,000	,000	,000
	basic	130	,000	1,000	,000
	high school	50	,000	,000	1,000
	university	47	,000	,000	,000

Last step in the Logistic Regression Analysis (Backward Wald)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7(a)	Time spent studying			11,142	3	,011	
	Time spent studying (1)	1,953	,828	5,565	1	,018	7,050
	Time spent studying (2)	,813	,793	1,051	1	,305	2,255
	Time spent studying (3)	,795	,855	,864	1	,353	2,215
	Mother's qualification			11,288	3	,010	
	Mother's qualification (1)	1,532	1,338	1,313	1	,252	4,629
	Mother's qualification (2)	2,006	,637	9,924	1	,002	7,434
	Mother's qualification (3)	1,314	,728	3,254	1	,071	3,721
	Constant	-3,786	,972	15,170	1	,000	,023

Conclusions are very similar (we could expect that father's qualification would not appear in the final model) and we can estimate that, for example, the odds that the student fails increase by a factor of 7.05 when the student doesn't study compared with a student that studies more than 2 hours per week, since other variables are controlled. But we loose important information when considering only two categories for the variable representing students' results in Mathematics (binomial logistic regression), as we would loose information considering more categories (multinomial logistic regression). Choosing one category to compare with the others doesn't make sense if we note that the possible results in Mathematics are 1, 2, 3, 4, or 5, and that it is not always very clear the difference between some of them. Exploring the relations between independent variables in logistic regression is not an inviting strategy.

It is more reasonable to interfere by choosing the number of clusters, or in the clusters interpretation, than interfering upon which category of variable will be compared with all the other categories, or how to group variable categories or in anyway "forcing" the data to fit in a logistic regression analysis.

3. Factorial Analysis approach

As it can be seen in the following table, the 5 factor solution of Factorial Analysis applied to the same database, shows that the first factor (which total variance explained is the most important) led us to formulate an identical interpretation to the results we obtained in the previous statistical analyses.

Anyway, the results of Factorial Analysis are not so easy to interpret as are the results of a Cluster Analysis.

Factor anlysis solution

Component Matrix(a)

	Component				
	1	2	3	4	5
Results in Mathematics as classified by the school teacher	,747	,113	,197	-,200	-,197
Result in Mathematics examimation	,738	,043	,044	-,307	-,154
Sex	-,031	,524	-,392	,333	,211
Had already failed	-,638	,117	,010	-,144	,000
Lives with his parents	,168	-,314	,329	,457	,494
Mother's qualification	,758	-,095	-,046	,081	,131
Father's qualification	,741	-,054	-,004	,102	,138
Time spent studying	,328	,381	-,413	-,223	,140
Special Mathematics classes by the school teacher	,130	-,622	-,084	,366	-,322
Doesn't study, usualy	-,105	,401	,652	-,147	,247
Lack of atention/concentration	,033	,324	,607	,178	-,383
Too many hours of classes	-,135	-,310	,025	,041	-,296
Difficulty on interpretation	,146	,454	,014	,541	-,016
How to improve results in Mathematics	-,040	-,437	,165	-,265	,501

Extraction Method: Principal Component Analysis. (a) 5 components extracted.

REFERENCES (RÉFERENCES)

Agresti, 1981; measures of nominal-ordinal association. *Journal of the American Statistical Association*, **76**, 524-529.

Everitt, Brian; Landau, S.; and Leese, M, 2001; *Cluster Analysis*, 4th ed., Arnold, London.

Gnanadesikan, R., 1997, *Methods of Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons.

Jobson, J.D., 1991; *Applied Multivariate Data Analysis*, Vol.II, Springer.

Milligan, G. and Cooper M., 1985; An examination of procedures for determining the number of clusters in a data set. *Psycometrika*, **50**, 159-179.

Hosmer, David; and Stanley Lemeshow, 1989, *Applied Logistic Regression*, John Wiley and Sons, Inc.

Abstract

The work developed shows that Cluster Analysis appropriately answers the questions that arise when we try to frame socially and pedagogically the success/failure in Mathematics of 14-15 - year old school students.