Student Data Analysis using Clustering Approach

Prasanna S. Karmarkar
Department of Computer Science,
Shivaji University,
Kolhapur,Maharashtra,India

Kavita S. Oza
Department of Computer Science,
Shivaji University,
Kolhapur,Maharashtra,India

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 a particular subject. Proposed worked developed a logistic regression analysis, as the study was carried out as a contribution to explain the success/failure in Exams. As a final step, the responses of both statistical analysis were studied

1. INTRODUCTION

The amount of data maintained in an electronic format has seen a dramatic increase in recent times[1]. The amount of information doubles every 20 months, and the number of databases is increasing at an even greater rate [2]. The search to determine significant relationships among variables in the data has become a slow and subjective process. As a possible solution to this problem, the concept of Knowledge Discovery in Databases – KDD has emerged [3]. The process of the formation of significant models and assessment within KDD is referred to as data mining [4]. Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful [5]. Cluster analysis is a technique used in data mining. Cluster analysis involves the process of grouping objects with similar characteristics [6], and each group is referred to as a cluster. Cluster analysis is used in various fields, such as marketing, image processing,geographical

information systems, biology, and genetics. In this study, university students were grouped according to their characteristics,

forming clusters. The clustering process was carried out using a Kmeans algorithm.

Cluster analysis is a multivariate analysis where individuals with similar characteristics are determined and classified (grouped) accordingly [7]. Through cluster analysis, dense and sparse region can be determined in the distribution, and different distribution patterns may be achieved. The concepts of similarities and differences are used in cluster analysis. Different measures are used in determining similarities and differences.[8, 9,10,11,12,13]

2. PROPOSED WORK

A database was created after an inquiry to Shivaji University students, which was developed with the purpose of identifying the factors that could socially and pedagogically frame the results in Exams. The data was collected from University ,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 Exams. As a final step, the responses of both statistical analysis were studied. In this study, university students were grouped according to their characteristics.

3. CLUSTER ANALYSIS APPROACH

The questions that arise when we try to frame socially and pedagogically the results in Computer Science 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.

Variables in **Cluster1**: mother qualifications; father qualifications; student's results in Computer Science as classified by the teacher; student's results in the exam of Computer Science; time spent studying. This result enhances the influence of parents' qualifications in the results in Computer Science and the time spent on studying, although the last one is less influent. It shows the influence of parents' qualifications in their students results in Computer Science 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 Computer Science; special aid of Computer Science teachers; too many hours attending College classes. Variables in this cluster reflect the family conversations about Computer Science classes.

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

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 Computer Science. 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 Computer Science .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.

4. LOGISTIC REGRESSION ANALYSIS APPROACH

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

Table 1

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

Logistic Regression Analysis (Backward Wald)

Table 2.

	B	SE	Wald	DF	sig	Exp(B)
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	354	2,215

Mother's qualification	----	----	11,28	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 Computer Science (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 Computer Science 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.

5. 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 analysis solution

Table 3

	Component				
	1	2	3	4	5
Results in Computer Science as classified by the Teacher	747	113	197	-200	-197
Result in Computer Science 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 Computer Science classes by the teacher	130	-622	084	366	-322
Doesn’t study, usually	-105	401	652	-147	247
Lack of attention/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 Computer Science	-040	437	165	--265	501

6. CONCLUSION

This study utilises data mining in the field of education. Cluster analysis were used as data mining techniques. The steps of the data mining process were carried out and explained in detail. The area of application was education, different from the usual data mining studies. The use of the data mining technique in education may provide us with more varied and significant findings, and may lead to the increase in the quality of education.

7. REFERENCES

[1] Vahaplar, A.,İnceoğlu, M., “Veri Madenciliği ve Elektronik Ticaret”, Türkiye’de İnternet Konferansları, Harbiye İstanbul, 1-3 Kasım 2001.

[2] Erdoğan, Ş. Z., “Veri Madenciliği ve Veri Madenciliğinde Kullanılan , 2004 K-Means Algoritmasının Öğrenci Veri Tabanında Uygulanması”, Yüksek Lisans Tezi, İstanbul Üniversitesi.

[3] Akpınar, H 2000., “Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği”, İ.Ü. İşletme Fakültesi Dergisi, Sayı:1 (1-22), Nisan.

[4] Thearling, K., 01 December 2003 “An Introduction to Data Mining”,http://thearling.com/text/dmwhite/dmwhite.htm.

[5] Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., Uthurusamy, R., 1994 “Advances in data mining and knowledge discovery”, MIT Pres, USA.

[6] Han, J., Kamber, W., 2001 “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, USA, 5-10,.

[7] Menteş, G. T., 2000, “Faktör ve Kümeleme Analizi Yardımıyla Bankacılık Ürün ve Hizmetlerinin Araştırılması Üzerine Bir Uygulama”, Doktora Tezi, İstanbul Üniversitesi.

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

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

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

[11] Jobson, J.D., 1991; Applied Multivariate Data Analysis, Vol.II, Springer. [12]Milligan, G. and Cooper M., 1985; An examination of procedures for determining the number of clusters in a data set. Psycometrika, 50, 159-179.

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