APPLICATION OF CLUSTER ANALYSIS AS A TOOL TO ANALYSE DISTANCE EDUCATION STUDENTS

Anurag Saxena[1], Pankaj Khare[2] and Suresh Garg[3]
Indira Gandhi National Open University, New Delhi, India

Abstract

The educational databases often have hidden knowledge about the students, their academic behavior, their study skills and their performance in an academic program. This explicit knowledge (Koulopoulos and Frappaolo, 1999) can be used to facilitate learning more effectively and efficiently by the educational institutions. There are many studies that tried to analyse students' characteristics and then draw conclusions about the students. These studies were mainly based on either nominal and interval data and the characteristic were judged by the percentage of students possessing these characteristics. Almost negligible number of studies tried to analyse students globally with respect to all their characteristics. Question thus arises that can students be classified on the basis of knowledge delivered by the student database. One of the data mining techniques of structuring data is "cluster analysis". Clustering literally means, "to gather" or "draw together". In terms of data, clustering means dividing the data in such a way that similar data points comes together.

The study has collected data on various student characteristics and attempted to apply cluster analysis technique on the science graduate students of Indira Gandhi National Open University, India from Delhi region. The dataset has taken a sample of 75 students from the data repository of 1,307 students of which 693 students were active students (Khare, et.al., 2003).

The study tried to obtain various clusters of students that are "heterogeneous among" them but "homogenous within" themselves. From the results obtained study has tried to identify the reasons that though students are different in their performance in the TEE, there is some similarity between them, which binds them together in a particular cluster. The study has also tried to build up a measure to indicate the dissimilarity among the students.

Introduction

Literal meaning of clustering is to gather, to congregate or draw together. In terms of data management, clustering means dividing the data in such a way that similar data points comes together. The objective of clustering is form groups that are heterogeneous but homogeneous within. Clustering is thus a method to divide a database into clusters that can be used for classification purpose. However, classification segments the data into groups that are already defined. Clustering facilitates segmentation of the data into groups that are not previously defined.

The clientele of distance education is normally heterogeneous in nature (Chisthty, 1992, Giglotti, 1995). They come from diverse environments, different socio-economic streams, and varied age groups and sometimes without even having formal education (Pathni, 1985). The learning styles, learning environment and learning culture can also be a function of the heterogeneity and may affect the performance of students in an academic programme. The academic institutions should address the issues that mushroom-up from the heterogeneity and should try to build bridges to create 'balancing parameters' among identified group(s) of students. For the purpose, educational institutions are to be equipped with a tool to find out the dissimilarities among the learners.

Then a question comes, is there a method of defining the "dissimilarity" between learners? A student-centric attitude forces to have a re-look into the "learners" from learning environment, learning culture and learning style point of view. The basic premises of question-set that may arise are:

[1] Reader, School of Management Studies, Indira Gandhi National Open University, New Delhi.

[2] Programme Officer, Commonwealth Educational media Centre for Asia, New Delhi.

[3] Pro-Vice-chancellor & Professor of Physics, Indira Gandhi National Open University, New Delhi.

- Whether demography has a role to make the students ‘different’ from the normal?
- Whether the prior qualifications, place of residence and/or social status make these students ‘different’?
- Whether ‘different students’ perform differently in the evaluation components and become different?
- Does a combination of surrounding learning environment have any effects on learning style?

This study is thus intended to make an in-depth study on various student-parameters through cluster analysis to device a tool to formulate group(s) of students who are ‘different’. The main purpose of this study is to build a sound logic and deduce how the students can be classified into different heterogeneous groups that are homogeneous within and try to find reasons that make the group(s) of ‘different’.

Clustering

According to Berry and Linoff (2001), “Cluster Analysis is an important human activity. Early in Childhood, one learns to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes”. With the help of clustering one can segment the data into small similar regions and thus comment on the overall distribution patterns of the data. In Education, clustering can help institutions discover distinct groups in their student databases and characterise student groups based on their socio-economic-geographical characteristics.

Clustering is done on the basis of a similarity measure – attributes/variables to derive the clusters so that data points in one cluster are more similar to another (homogeneous) and data points in separate clusters are less similar or dissimilar to the data points of another cluster(s) (heterogeneous) (Anderberg, 1973).

Clustering methods are discussed in various text books (Hartigan, 1975, Jain and Dubes, 1988). These methods have various techniques and can be performed in many ways. In case of more than two variables, use of mathematical models and computer programme becomes necessary for the analysis of data (Sharma, 2002). There are a few methods that start by considering all records to be part of one big cluster and then splits them into two or more smaller clusters. On the other hand, there are methods that start with each record taken as a cluster, and iteratively combine to form clusters. The former methods are called *Divisive methods* and the latter *Agglomerative methods* (Romesburg, 1984, Kaufman and Rousseeuw, 1990). Another method is grouping of two closest objects as a single cluster and thus number of objects is reduced to *n* –1. Then next two objects are grouped and the process continues till all *n* objects are covered under single cluster. Here the clustering is done step-by-step and the method is known as “*hierarchical clustering*” (Romesburg, 1984).

For applying clustering techniques, data is arranged in two matrices, called “data matrix” and “dissimilarity matrix”. While data matrix is a representation of *n* objects (such as students) with *m* attributes (such as gender, program, region, age, social status etc.), dissimilarity matrix is a collection of distances (not the distances as propounded by Moore, 1973) between the pair of objects. Data matrix can be shown as follows :

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{k1} & \cdots & x_{kj} & \cdots & x_{km} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \qquad \ldots (1)$$

Dissimilarity matrix can be shown as follows :

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,1) & 0 & & \\ \vdots & \vdots & \vdots & & \vdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix} \qquad \ldots (2)$$

If the distances in the matrix are near to zero then the objects are highly similar or “near” to each other.

Calculation of Distances

According to Han and Kamber (2001), one could come across various types of variables while clustering the data. In this paper, three data variables are used, interval, binary and nominal. A brief description of these variables is given below.

Interval scaled variables **:** These variables are continuous measurement of a roughly linear scale, e.g., weather temperature and weight and height etc. One can find the distances between the objects. This is called the Euclidian distance (*d*) and is defined as

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \ldots.. + \left|x_{im} - x_{jm}\right|^2} \qquad \ldots (3)$$

where, *i* and *j* are two *m* dimensional data objects represented by $(x_{i1}, x_{i2}\ x_{i3} \ldots x_{im})$ $(x_{j1}, x_{j2}, x_{j3}, \ldots x_{jm})$.

Binary variables **:** These variables have only two states : 0 or 1, where 0 means that the variable is absent and 1 means it is present. While calculating the distance between these type of variable, the data can be arranged in the form of a *contingency table* (Table 1) :

Table 1 : Contingency table for binary variables

		Object *j*		
		1	0	sum
Object *i*	1	*q*	*r*	*q* + *r*
	0	*s*	*t*	*s* + *t*
	Sum	*q* + s	*r* + *t*	*m*

We will deal with the symmetric binary variables and the distance between them are calculated on the basis of simple matching coefficient, defined as :

$$d(i,j) = \frac{r+s}{q+r+s+t} \qquad \ldots (4)$$

where, *q* is the number of variables which are 1 for both the objects *i* and *j* , *r* is the number of variable that are 1 for object *i* and 0 for object *j*, *s* is the number of variables which are 0 for object *i* and 1 for object *j* and *t* is number of variables that are 0 for both the objects. In the present case, the binary variables take the values 1 and 2, e.g., male = 1 and female = 2.

Nominal variables **:** Nominal variables are a generalization of binary variables in a sense that these variables can take more than two states, e.g., social status code, where more than 2 values are assigned (Tribal, backward, general, etc.). In distance education database, these states are generally represented by integers 1, 2, …, etc. The distance between these objects is defined by the *simple matching* approach and is defined as follows :

$$d(i,j)=\frac{m-m'}{m} \qquad \ldots(5)$$

Where m' is the number of matches, i.e., number of variable that have objects *i* and *j* in the same state and *m* is the total number of variables. In present study, we have not followed this approach, as the number of states of these variables is not same. Some variables have taken unequal number of states. The *nearest neighbor* approach to calculate the distances (SPSS, 2001) has been taken in this paper.

Since distance education databases contain all these variables (interval, binary and nominal), and therefore, to find distance between mixed type variables, a comprehensive distance matrix was used. This is necessary as it is unlikely in real world situations that separate cluster analysis for separate variable type yield compatible results. One of such techniques (Kaufman and Rousseeuw, 1990) is to combine all different variables in to a single distance matrix bringing all meaningful variables onto a common scale of interval (0.0, 1.0). If there are *p* variables of mixed type, then distance can be defined as :

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}} \qquad \ldots(6)$$

where, the indicator $\delta_{ij}^{(f)}$ = 0, if x_{if} or x_{jf} is missing. That is, there is no measurement of variable *f* for the objects *i* or object *j* ; otherwise $\delta_{ij}^{(f)}$ =1. We have given another weight to the distances of these different types of variables, i.e. the number of variables of each type included in the analysis. This was considered necessary because we have taken unequal number of variables in the analysis.

Methodology

To exhibit how clustering methods help in defining the distance between learners, a dataset of the science graduates of Indira Gandhi National Open University was taken and was arranged on the basis of information on 19 parameters. Luan (2001) used similar parameters in his study on predicting persisting and non-persisting students of a community college. The parameters use in this study are classified as follows :

Binary Variables : Medium of study (English or Hindi), gender (male or female), marital status (married or unmarried), employment status (employed or unemployed), and availability of telephone at residence (yes or no);

Nominal Variables : Category (scheduled castes, scheduled tribes, other backward class, physically handicapped or general), area (urban, rural or Kashimiri migrant) and social status (ex-service man, war widow or normal citizen); and

Interval Variables (1): Average marks obtained in assignment (theory), term-end theory examination, practical continuous assessment and practical examinations and number of assignment, term-end exam, practical and practical continuous assessments done by a student in the present programme.

Interval Variable (2) : Age of student at the time of admission in the present programme, gap in the studies (from previous qualification to present registration) and percentage of marks obtained in the previous qualification.

As the interval variables were having two groups, first related with present programme and second with the previous qualification, they were grouped them in two categories.

These variables are taken separately and three hierarchical cluster analyses was conducted to find the dissimilarities between the learners and finally applied the concept given by Kaufman and Rousseeuw (1990) for getting a single dissimilarity matrix. To exhibit the concepts more clearly we have taken only 10% (75 cases) of the data into analysis. This was done in order to derive meaningful conclusions, which could otherwise also be verified by inspection.

Analysis

Binary Data

The first analysis was done on the binary data comprising of medium of study, gender, marital status, availability of telephone facility at home and employment status. As mentioned earlier, rescaled simple matching criteria was carried out to find the dissimilarities. The data was divided into 5 clusters on the basis of similarities (Table 2).

Table 2 : Clusters obtained on binary variables

Cluster	Total cases in the cluster	Case identity numbers	General Profile
1	27	1, 12, 13, 16, 19, 24, 25, 27, 28, 29, 30, 31, 38, 39, 41, 43, 44, 47, 48, 49, 52, 57, 60, 61, 68, 70, 71	Unmarried, unemployed males opted English medium for study and have telephone facility at home
2	29	2, 4, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 20, 21, 23, 26, 33, 36, 42, 45, 51, 54, 55, 56, 58, 63, 65, 67, 73	Unmarried, unemployed females opted English medium for study and have telephone facility at home
3	15	3, 5, 22, 32, 34, 35, 37, 40, 46, 50, 53, 62, 64, 66, 69	Unmarried, unemployed males opted English medium for study and do not have telephone facility at home
4	2	59, 72	Married, employed females opted English medium for study and have telephone facility at home
5	2	74, 75	Married, employed males opted English medium for study and have telephone facility at home

Nominal Data

The second analysis (Table 3) was done on the nominal data, i.e., social category with specific reference to army, region of residence and social status. Rescaled Chi-square analysis between sets of frequencies was done to find the dissimilarities between the cases. This measure is based on the Chi-square test of equality for two sets of frequencies.

Table 3 : Clusters obtained on nominal variables

Cluster	Total cases in the cluster	Case identity numbers	General Profile
1	53	1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 33, 35, 36, 40, 41, 42, 43, 44, 45, 46, 51, 54, 55, 59, 60, 61, 62, 63, 65, 67, 70, 71, 72, 73, 74, 75	Urban citizens without any reservations in the society and are not associated with armed forces.
2	3	5, 38, 47	Rural citizens categorized under other backward classes in the society but have no association with armed forces.
3	10	7, 26, 32, 48, 49, 50, 52, 53, 58, 69	Rural citizens without any reservations in the society and are not associated with armed forces.
4	1	20	Urban citizens categorized under other Scheduled Caste in the society and have no association with armed forces.
5	8	34, 37, 39, 56, 57, 64, 66, 68	Urban citizens categorized under other backward classes in the society and have no association with armed forces.

Interval Data

The third analysis was done on the interval data, grouped as Interval1 and Interval2. The rescaled squared Euclidean Distance was used to find the dissimilarities between the cases. The clusters obtained on both the groups are given at Table 4 and 5.

Table 4 : Clusters obtained on interval1 variables

Cluster	Total cases in the cluster	Case identity numbers	General Profile
1	17	1, 5, 13, 24, 28, 29, 31, 32, 34, 36, 40, 47, 49, 56, 66, 67, 69	Attempted assignments and term-end exam only
2	45	2 , 3 , 4 , 6 , 8, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 30, 33, 35, 39, 41, 42, 43, 45, 46, 50, 51, 52, 53, 54, 57, 61, 62, 63, 65, 68, 71, 72, 73, 74, 75	Attempted all components and earned non- zero credits there
3	5	7, 9, 37, 38, 48	Attempted practical exams and practical assignments only
4	8	27, 44,55,58,59,60,64,70	Attempted term-end exam only

Table 5 : Clusters obtained on interval2 variables

Cluster	Total cases in the cluster	Case identity numbers	General Profile
1	44	1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 56, 57, 58, 59, 62, 63, 64, 66, 70	Age at the time of admission in present programme – 17 to 19 years; gap in studies 1-4 yrs, percentage obtained in previous qualification – 33-80%
2	26	8, 9, 16, 17, 18, 19, 32, 33, 34, 35, 36, 37, 38, 49, 50, 51, 52, 53, 54, 55, 60, 61, 65, 68, 69, 71	Age at the time of admission in present programme – 19 to 25 years; gap in studies 2-8 yrs, percentage obtained in previous qualification – 33-80%
3	4	67, 72, 73, 74	Age at the time of admission in present programme – 25 to 27 years; gap in studies around 9 years, percentage obtained in previous qualification – 33%
4	1	75	Age at the time of admission in present programme – 42 years; gap in studies 9 yrs, percentage obtained in previous qualification – 79%

Discussion

We return to our basic premises of questions set that we set earlier in the study.

- Whether demography plays an important role in making these students different ?
- Whether the qualifications, area code and social status etc. makes them different ?
- Whether the performance in the evaluation components make them different.?, and
- Whether the pre-admission conditions have an impact on them ?

It is quite evident from the analysis that clustering techniques are quite successful in segregating the students into heterogeneous groups that are homogeneous within. We started on binary variables and table 1 gives us 5 clusters on the basis of varying values for the parameters like medium of study, ownership of telephone, gender of the student, marital status of the student and employment status of the student. Then we analyzed the nominal variables and Table 2 gives 5 clusters on varying values of category to which student belongs, geographical area from where the student has come and the societal status of the student. The third analysis was done on the interval data. This component is based on the performance of the student in various evaluation components and the credits earned thereof. We are able to get 4 clusters of students according to the variability in their performance. Finally, we have taken into account the time lag that student has spent before resuming his studies, percentage obtained in the previous qualification and age at the time of admission. This analysis is able to give us 4 clusters.

Hence, we now have 18 different clusters of the students. There ought to be overlaps in these eighteen clusters and that was what we are interested in. So we decided to do cross-tabulations in order to find a pattern.

Table 6 : Showing inter-relations between the variable type (for referring to general profile of students in clusters, see Tables 2–5)

Table 6.1

Count of Cases		Nominal Clusters					Grand Total
		1	2	3	4	5	
Binary Clusters	1	19	2	3		3	27
	2	24		3	1	1	29
	3	6	1	4		4	15
	4	2					2
	5	2					2
Grand Total		53	3	10	1	8	75

The Table 6.1shows that 43 students have come from cluster 1 of the nominal variables.

Table 6.2

Count of Cases		Interval1 Clusters				Grand Total
		1	2	3	4	
Binary Clusters	1	18	9			27
	2	17	10	2		29
	3	8	7			15
	4	1		1		2
	5			1	1	2
Grand Total		44	26	4	1	75

The Table 6.2 shows that 35 students have come from cluster 1 of the interval1 variables.

Table 6.3

Count of Cases		Interval2 Clusters				Grand Total
		1	2	3	4	
Binary Clusters	1	8	13	2	4	27
	2	3	22	2	2	29
	3	6	7	1	1	15
	4		1		1	2
	5		2			2
Grand Total		17	45	5	8	75

The Table 6.3 shows that 35 students have come from cluster 1 of the interval2 variables.

Table 6.4

Count of Cases		Nominal Clusters					Grand Total
		1	2	3	4	5	
Interval1 Clusters	1	32	2	4	1	5	44
	2	16	1	6		3	26
	3	4					4
	4	1					1
Grand Total		53	3	10	1	8	75

The Table 6.4 shows that 48 students from the first cluster of nominal variables belonging to either cluster1 or cluster 2 of interval1 variables.

Table 6.5

Count of Cases		Nominal Clusters					Grand Total
		1	2	3	4	5	
Interval2 Clusters	1	9	2	3		3	17
	2	37		4	1	3	45
	3	1	1	2		1	5
	4	6		1		1	8
Grand Total		53	3	10	1	8	75

The Table 6.5 shows that 46 students from the first cluster of count variable belong to either cluster1 or cluster 2 of interval2 variables.

Table 6.6

Count of Cases		Interval2 Clusters				Grand Total
		1	2	3	4	
Interval1 Clusters	1	11	25	2	6	44
	2	5	16	3	2	26
	3	1	3			4
	4		1			1
Grand Total		17	45	5	8	75

The table shows that 41 students from the second cluster of intervel2 variable belong to either cluster1 or cluster 2 of interval1 variables.

By doing this exercise we were able to define bigger categories or clusters of students sharing the similar characteristics. We then did a global analysis taking the entire variables at the same time and then visualizing the students who were same on the basis of all these parameters.

Table 7 : Global positioning of 18 clusters obtained on the basis of 19 variables

Count of Cases		Interval1 Clusters											Grand Total
		1				2			3			4	
Interval2	Nominal	B1	B2	B3	B4	B1	B2	B3	B2	B4	B5	B5	
1	1	6		1			1		1				9
	2	1		1									2
	3					1		2					3
	5		1	1				1					3
2	1	5	12	4		4	7	1	1	1	1	1	37
	3		1			1		2					4
	4		1										1
	5	2				1							3
3	1						1						1
	2					1							1
	3	1	1										2
	5							1					1
4	1	3			1	1	1						6
	3		1										1
	5			1									1
Grand Total		18	17	8	1	9	10	7	2	1	1	1	75

B1, B2, B3, B4 and B5 represent Binary Clusters.

In the absence of any major pattern, we focused our attention on the value 12 (highlighted). The values can be explained as the number of students who belong to nominal=1, interval2=2, binary=2 and interval1=1. That means they are the students who were "Unmarried, unemployed, urban females within age range of 17 to 20 years, continued their studies on regular basis and have done satisfactory in their qualifying examinations. They are privileged with basic communication device (telephone) at residence and are regular in their studies in the present programme (attempts all assessment components of present programme).”

The identified group of 12 students is placed on the basis of similar parameters under observations. We verified this homogeneity on the basis of a combined analysis given by the clustering technique. These students were having identification numbers – 2, 4, 6, 10, 11, 14, 15, 21, 23, 42, 45, 63.

Further analysis on the combined distance matrix obtained vis-à-vis the performance of these 12 students is given at Table 8.

Table 8 : Dissimilarity coefficient between 12 female students

Dissimilarity Matrix

Case	2	4	6	10	11	14	15	21	23	42	45	63
2	0.00											
4	0.03	0.00										
6	0.02	0.05	0.00									
10	0.08	0.04	0.11	0.00								
11	0.10	0.04	0.12	0.02	0.00							
14	0.02	0.04	0.01	0.12	0.12	0.00						
15	0.02	0.05	0.03	0.09	0.10	0.02	0.00					
21	0.09	0.03	0.10	0.04	0.01	0.10	0.10	0.00				
23	0.09	0.05	0.11	0.01	0.01	0.11	0.10	0.02	0.00			
42	0.01	0.05	0.01	0.09	0.11	0.02	0.03	0.10	0.10	0.00		
45	0.02	0.07	0.03	0.12	0.12	0.02	0.01	0.11	0.11	0.03	0.00	
63	0.07	0.03	0.09	0.02	0.01	0.09	0.07	0.02	0.01	0.08	0.10	0.00

The Table 8 gives a wonderful visualization of the homogeneity. All 12 students have very low coefficients of dissimilarity among them (0.00 to 0.12 only). This can serve as a tool for segregating students according to their dissimilarity coefficients.

The second premise that is to be tested is whether these homogeneous students perform similarly. This can be visualized from the Table 9, given below.

Table 9 : Performance of 12 ‘similar’ students

Case	No. of assig. attempted	Average marks in assign.	No. of term-end papers attempted	Average marks in term-end	No. of lab assess. attempted	Average marks in cont. assess.	No. of lab exams attempted	Average marks in lab exams	Overall percentage of marks
2	10	60	5	48	1	79	1	67	63.50
4	7	63	10	61	4	82	4	80	71.50
6	6	73	3	52	1	53	1	60	56.70
10	19	66	14	51	4	75	4	77	65.55
11	18	84	11	47	7	79	7	74	67.80
14	3	60	3	36	2	63	2	67	53.70
21	11	81	10	54	7	85	7	68	71.00
23	20	78	11	46	6	76	6	64	64.00
42	9	51	4	53	1	71	1	50	58.55
45	7	77	4	26	1	86	1	43	57.20
63	17	75	11	46	6	71	6	76	63.60

As can be seen from the performance of these 12 students, they have attempted all the four components of assessments, they have earned credits, and they performed satisfactory in examinations. But for exceptional few cases, their performance in the individual assessments is almost similar and comparable.

Another interesting study could address to the socially disadvantaged students, as reflected in cluster 5 of Table 3. There were 8 students who belong to other backward classes (OBC) residing in urban area. A proximity matrix was created (Table 10) to find the dissimilarities.

Table 10 : Dissimilarity coefficient among "disadvantaged" students

Proximity Matrix for Disadvantaged Students

Case	34	37	39	56	57	64	66	68
34	0.00							
37	0.18	0.00						
39	0.28	0.32	0.00					
56	0.20	0.34	0.30	0.00				
57	0.25	0.25	0.14	0.26	0.00			
64	0.21	0.19	0.52	0.29	0.27	0.00		
66	0.07	0.21	0.24	0.18	0.23	0.20	0.00	
68	0.24	0.23	0.30	0.29	0.07	0.23	0.28	0.00

This is a dissimilarity matrix

One can easily visualize the varied dissimilarity coefficient in the table. This implies that although these students are tagged together by virtue of their category and are entitled to get privileged treatment. However, in reality, they are heterogeneous. For example, case 39 was employed but case 64 was not, case 39 had attempted all components of assessments whereas case 64 has attempted only term-end examinations. Thus, case 39 and case 64 are parametrically totally different students. Another example is of case 34 and case 66. Both these students are employed, have telephone at home, missed the studies for some times and attempted only assignments and term-end examinations of present programme. These two students are more or less similar students.

This generates a point of discussion on the schemes giving privileges to a particular category of students who are homogeneous among themselves on one set of variables but heterogeneous otherwise. The social category, thus have little or no impact on the learning styles or learning behavious of the students clustered in the group. Thus, should educational institutions or employing institutions continue to provide privileges, like cut-off in entrance examinations or job reservations need to be ascertained on the basis of findings. The findings are to be reconfirmed or reassured on different set of data before entering in to a debate.

Conclusion

There are a few things that emerge out from the study. Firstly, that there are almost 18 categories which can be used to segregate the students in to homogeneous groups. These 18 categories came out from different permutations of the parameters (variables) considered for the analysis. Since the study had dealt with only 10% of the cases, it may not be wise to conceive a global comment on student characteristics and this was, in fact, not the aim as well. The efforts were for development of a tool to define the homogeneous groups and the study has been successful in doing so. It is evident that the cluster analysis is capable of quantifying the dissimilarity between the students.

Secondly, the aim of this study was to empower the decision support system with different sets of students for which the system has to think differently, e.g. one can refer to Cluster 3 of Table 4, where 5 students attempted only laboratory courses and did not bother abut theory courses. These students are more likely to burnout in due course of time. This situation is encountered at a time when we have taken only active students

(Pathaneni, 1995) in to account. The data of students who are already deemed as passive students or dropouts was ejected out right at the beginning. This set of students could not initiate the learning process at the juncture when elective courses were offered to them. They failed to submit their assignments or failed to attempt term-end examinations. Among the students who were starter of the learning process but were not able to keep up with the pace, the present study was successful in "identification of possible dropouts".

Thirdly, the study was able to successfully generate interest in studying the effects of individual parameters in overall performance. This is yet to be seen whether two students who differ on the basis of some parameters may become similar because of contribution of other parameters. It will also be interesting to note effect of "balancing parameters" which are playing an important role in bringing "homogeneity among heterogeneity".

The discussions here could go infinitely longer. The need is that decision support system picks up the group it want to address and generate mechanism to strengthen learning culture. We conclude by saying, "Computational DSS has finally emerged as essential tool for effective pedagogy."

References

Anderberg, M.R. (1973) *Cluster Analysis for Applications,* New York : Academic Press.

Chisthty, S.B.H. (1992) Achievement motivation, self-concept, personal preferences, student's morale and other ecological correlates in relation to intelligence, socio-economic status and performance of higher secondary tribal students of Rajasthan, *Indian Education Review,* 27 (4).

Giglotti, Card Chafel (1995) The relationship between self-concept of academic ability and academic performance of adult students (28 and older), *Dissertation Abstracts International,* 55(7), 1791#A.

Han Jiawei and Kamber M (2001) *Data Mining : Concepts and Techniques*, CA : Morgan Kaufman Publishers, 2001.

Hartigan J.A. (1975) *Clustering Algorithms*, New York: John Wiley and Sons, 1975.

Jain A.K. and Dubes R.C.(1988) *Algorithms for Clustering Data*, Englewood Cliffs, NJ; Prentice Hall, 1988.

Kaufman L and Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley &Sons, 1990

Khare, Pankaj, Saxena, Anurag and Garg, Suresh (2003) Knowledge discoveries on performance of IGNOU science graduates through data mining – II, Indian *Journal of Open Learning,* 12 (3), in press.

Koulopoulos, T.M. and Frappaolo, C. (1999) *Smart things to know about knowledge management*, Dover : Capstone Publishing Ltd.

Luan, Jing (2001) *Data mining as driven by knowledge management in higher education : persistence clustering and prediction*, Keynote for SPSS Public Conference, UCSF.

Moore, M.G. (1973) Towards a theory of independent learning and teaching, *Journal of Higher Education*, 44, pp. 661-679.

Pathaneni, S. (1995) An Assessment Technique for Motivation of Distance Learners : A Case Study, Paper presented at VIII Annual Conference of Asian Association of Open Universities, New Delhi.

Pathni, R.S. (1985) Psycho-social development stage (identity vs. role confusion), self evaluation (self-concept) and need (self-analysing) as predictor of academic achievement (actual and perceived), in Buch, M.B. (Ed.), *Fourth survey of research in education,* 1991, New Delhi : NCERT.

Romesburg, H.C. (1984) *Cluster analysis for researchers*, Belmont, CA : Lifetime Learning Publication.

Sharma, K.R. (2002) *Research Methodology*, New Delhi : National Publishing House.

SPSS for Windows, (2001) Release 11.0.0, Standard version, SPSS Inc.