Prediction of Student's Academic Performance using Clustering

Prof. Prashant Sahai Saxena
Joint Director, School of Computer and Systems Sciences
Jaipur National University, Jaipur
sahai.prashant0@gmail.com

Prof. M. C. Govil
Head – Department of Computer Engineering
Malviya National Institute of Technology (MNIT), Jaipur
govilmc@yahoo.com

----------ABSTRACT----------

One of the significant facts in higher learning institution is the explosive growth of educational data. These data are increasing rapidly without any benefit to the management. The main objective of any higher educational institution is to improve the quality of managerial decisions and to impart quality education. Predicting successful and unsuccessful students at an early stage of the degree program help academia not only to concentrate more on the bright students but also to apply more efforts in developing remedial programs for the weaker ones in order to improve their progress while attempting to avoid student dropouts. The aim of this study is to apply the k-means clustering technique to analyze the relationships between students's behavioral and their success.

Keywords: DM, Student Academic Performance, Clustering, k-means.

I. INTRODUCTION

To identify potential dropouts of the institute's graduate program is a complex process mostly due to the fact that students coming from different backgrounds have certain characteristics as well as perceptions and apprehensions of the environment of the university. Students' failure to integrate and acquire good marks are considered to be one of the main factors but many researchers have also suggested that there are various other factors that may affect students' progress at the university level.

Predicting successful and unsuccessful students at an early stage of the degree program help academia not only to concentrate more on the bright students but also to apply more efforts in developing remedial programs for the weaker ones in order to improve their progress while attempting to avoid student dropouts. Performance evaluation is one of the basis to monitor the progression of student performance in higher education. With traditional grouping of students based on their average scores, it is difficult to obtain a comprehensive view of the state of the students' performance and simultaneously discover important details from their time to time performance. With the help of data mining techniques, such as clustering, it is possible to discover the key characteristics from the students' performance and possibly use those characteristics for future prediction.

This paper analyzes the clustering analysis in data mining that analyzes the use of k-means clustering algorithm in improving student's academic performance in higher education and presents k-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in higher educational institution.

Cluster analysis could be divided into hierarchical clustering and non-hierarchical clustering techniques. Examples of hierarchical techniques are single linkage, complete linkage, average linkage, median, and Ward. Non-hierarchical techniques include k-means, adaptive k-means, k-medoids, and fuzzy clustering. To determine which algorithm is good is a function of the type of data available and the particular purpose of analysis. In more objective way, the stability of clusters can be investigated in simulation studies [4]. The problem of selecting the "best" algorithm/parameter setting is a difficult one. A good clustering algorithm ideally should produce groups with distinct non-overlapping boundaries, although a perfect separation can not typically be achieved in practice. Figure of merit measures (indices) such as the silhouette width [4] or the homogeneity index [5] can be used to evaluate the quality of separation obtained using a clustering algorithm. The concept of stability of a clustering algorithm was considered in [3]. The idea behind this validation approach is that an algorithm should be rewarded for consistency. In this paper, traditional k- means clustering algorithm[6] and Euclidean distance measure of similarity was chosen to be used in the analysis of the students' scores.

II. METHODOLOGY

A. Development of k-mean clustering algorithm

This study uses an extraction method known as principal component analysis to predict cluster analysis. Principal component analysis carries out the reduction of data by deriving similarly few tools from relatively several

measured variables based on how the estimated variables load on the components. Then the individual records location can be investigated on basis of every score of record on components that are retained. If n components are retained they refer n-dimensional space in which every record can be located. This analysis uses a technique of data clustering termed K-means clustering which is applied to examine academic performance of students. K-means is one of the easiest algorithms of unsupervised learning used for clustering. K-means separates observations (i.e. "n") into clusters (i.e. "k") in which every observation belong to cluster with closest mean. This algorithm targets at reducing an objective function. This study conducts principle component analysis by considering 16 variables. The variables included in higher education research were subjected principle component analysis to find out the validity of the variables. The variables used in this study are Gender, Category, Grade division in X class, Grade division in XII class, Grade Division in Graduation, Admission type, Medium of Teaching till qualifying exam, Living location of student, Family annual income status, Father's qualification, Mother's qualification, Father's occupation, Mother's occupation, Programme, Semester and Section.

Given a dataset of n data points $x1, x2, \ldots, x_n$ such that each data point is in $\mathbf{R}^d$, the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points $\{m_j\}$ (j=1, 2, …, k) in $\mathbf{R}^d$ such that

$$\frac{1}{n}\sum_{i=1}^{n} [\min_i d^2(x_i, m_j)] \quad (1)$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j. The points $\{m_j\}$ (j=1, 2, …,k) are known as cluster centroids. The problem in Eq.(1) is to find k cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized.

The k-means algorithm provides an easy method to implement approximate solution to Eq.(1). The reasons for the popularity of k-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data.

The k-means algorithm can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function in Eq.(1). The k-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids. The problem of finding the global minimum is NP-complete. The k-means algorithm updates cluster centroids till local minimum is found. Fig.1 shows the generalized pseudocodes of k-means algorithm; and traditional k-means algorithm is presented in fig. 2 respectively.

Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational time complexity of the algorithm is $O(nkl)$, where n is the total number of objects in the dataset, k is the required number of clusters we identified and l is the number of iterations, $k \leq n$, $l \leq n$ [6].

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters
- Take first k instances or
- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters
Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

Figure 1: Generalised Pseudocode of Traditional k-means

```
1  MSE = largenumber;
2  Select initial cluster centroids {mj}j K = 1;
3  Do
4  OldMSE = MSE ;
5  MSE1 = 0;
6  For j = 1 to k
7  mj = 0; nj = 0;
8  end for
9  For i = 1 to n
10 For j = 1 to k
11 Compute squared Euclidean distance d2(xi, mj);
12 end for
13 Find the closest centroid mj to xi;
14 mj = mj + xi; nj = nj+1;
15 MSE1=MSE1+ d2(xi, mj);
16 end for
17 For j = 1 to k
18  nj = max(nj, 1); mj = mj/nj;
19 end for
20 MSE=MSE1;
while (MSE<OldMSE)
```

Figure 2: Traditional k-means algorithm [6]

III : METHODOLOGY

Through extensive search of the literature and discussion with experts on students' academic performance, a number of factors that are considered to have influence on the performance of a student were identified. The primary data is collected from a self financed university, based at Jaipur (India). These influencing factors were categorized as input variables.

The variables used in this study are Gender, Category, Grade division in X class, Grade division in XII class, Grade Division in Graduation, Admission type, Medium of Teaching till qualifying exam, Living location of student, Family annual income status, Father's qualification, Mother's qualification, Father's occupation, Mother's occupation, Programme, Semester and Section. The below table shows the data set variables used for K-means clustering analysis:

Initial Cluster Centers		
	Cluster	
	1	2
GENDER	2	2
CATEGORY	4	1
"Student Type in X Private / Regular"	1	1
"School Type/Class X (Govt. / Private)"	2	1
"Syllabus X (ICSE / CBSE / State)	2	2
GRADE/ DIVISION IN CLASS Xth	2	3
"School Type Class XII (Govt. / Private)"	2	1
"Syllabus XII (ICSE / CBSE / State)"	3	3
"Student Type in XII Private / Regular"	2	2
Grade/ division in class XIIth	2	1
Grade/division in graduation (only for MCA)	1	1
Admission type	1	1
Medium of teaching till qualifying exam	1	1
Living location of student	1	3
Family annual income status	2	2
Father's qualification	5	1
Mother's qualification	7	1
Father's occupation	4	1
Mother's occupation	2	1
Programme	2	2
Sem.	1	1
Section	1	2

Table 1: Initial Centers of Clusters

From the above table each and every variable is described. The first variable considered in principle component analysis using K means clustering was gender. This study considered male and female respondents under the variable gender for conducting clustering analysis. The second variable considered was the category. The category variable consists of three sub variables one is X class in private schools and in government schools, second sub variable X class in government and private schools and third sub variable is X class syllabus in CBSE, state board and in ICSE. Following the gender variable the third variable used in this analysis was grade division in X class. This variable also consists of three sub variables. The first sub variable is XII class in private and government schools. Another sub variable is XII class syllabus in state board, ICSE and CBSE. The third sub variable is XII students in regular and private session. Similar to the grade division in X class the fourth variable is grade division in XII class. The subsequent variable that is the fifth variable used in this analysis was grade division in graduation. This variable considers only MCA students for K-means clustering analysis. The sixth variable is the type of admission. The type of admission considered for this study was degree seeking and non degree seeking students. Following the sixth variable the seventh variable used for the study was the medium of teaching till qualifying exam. The medium of teaching considered in analysis was English, Hindi and others. The eighth variable used in the study was students' location. Following the location of student the next 8 variables were regarding the personal background information of students. The ninth variable was annual income of family. Following the annual income status of family the tenth and eleventh variable used in this analysis was father's qualification and mother's qualification respectively. Following the qualification the twelfth and thirteenth variables used in this analysis was father's occupation and mother's occupation. The last 14^{th}, 15^{th} and 16^{th} variables used for the analysis were programme, semester and section respectively. All the above mentioned variables were used for cluster analysis to examine the academic performance of students.

After describing each and every variable the K-means algorithm is applied on data set. The number of clusters was determined as an essential parameter. Varied number of clusters was attempted and successful partitioning was accomplished with 2 clusters. The result generated through iterations is shown in the below table:

Iteration	Change in Cluster Centers	
	1	2
1	2.657	3.192
2	.319	.526
3	.317	.405
4	.091	.104
5	.093	.112
6	.075	.088
7	.000	.000

Table 2: Iteration History

The K-means algorithm is performed in this analysis by following five steps. The first step in K-means algorithm was to accept the cluster numbers to combine data and the variables to cluster as values of input. The second step in K-means algorithm was to start the initial K-clusters that is considering k examples or considering random sampling of k elements. The third step in K-means algorithm was to estimate every cluster's arithmetic values which are comprised in the variables. The fourth step of algorithm would be to allocate every record in the variable to one of the starting clusters using K-means (i.e. every record must be allocated to closest cluster using a distance measure). The last step in the algorithm is to re-allocate every record in the variable to common clusters and re-evaluates cluster's arithmetic means of the variable using K-means. Before the convergence of K-means algorithm, centroid and distance estimations are performed while clusters are several times the positive integer referred as the number of K-means iterations. The accurate value of iteration differs relying on initial starting centroid even on similar set of data. From the above table it can be understood that the iteration history reveals number of iterations those were

enough until centers of cluster did not alter substantially. From the table the convergence was achieved due to no or small change in cluster centers. For any center the maximum absolute coordinate change is .000. The current iteration is 7 and the minimum distance between initial centers is 8.944. From the above table in the first iteration the overall distance of first cluster was 2.657 whereas in second cluster it was 3.192. There is 0.535 distance between the two clusters. Similarly in the second iteration the first cluster distance was .319 and the second cluster was .526. There are 0.2 differences in the second iteration. Following the second iteration the distance in the first cluster was .317 whereas in second cluster was .405 in the third iteration. The distance between the two clusters was only 0.1 difference. In the fourth iteration the first cluster distance was .091 whereas in second cluster was .104. The difference between the two clusters was .00. In fifth iteration the first cluster distance was .093 and second cluster was .112. The difference between the two clusters was same difference (.00) as in fourth iteration. Following the fifth iteration in sixth iteration the distance of first cluster was .075 and second cluster distance was .088. The difference between the two clusters was .013. Thus in the last iteration both the clusters distance and difference is .000.

After performing the iterations the F tests are used for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The below figure shows the ANOVA test table:

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Gender	.800	1	.136	92	5.896	.017
Category	10.842	1	2.071	92	5.234	.024
"Student Type in X Private / Regular"	.000	1	.000	92	.	.
"School Type/Class X (Govt. / Private)"	.105	1	.198	92	.529	.469
"Syllabus X (ICSE / CBSE / State)	.800	1	.136	92	5.896	.017
Grade/Division in class Xth	.168	1	.388	92	.432	.512
"School Type Class XII (Govt. / Private)"	.155	1	.231	92	.673	.414
"Syllabus XII (ICSE / CBSE / State)"	.351	1	.155	92	2.265	.136
"Student Type in XII Private / Regular"	.036	1	.021	92	1.718	.193
Grade/Division in class XIIth	.096	1	.254	92	.379	.540
Grade/Division in graduation (only for MCA)	.445	1	.147	92	3.040	.085
Admission Type	.000	1	.000	92	.	.
Medium of teaching till qualifying exam	.022	1	.051	92	.425	.516
Living location of student	11.730	1	1.362	92	8.615	.004
Family annual income status	.171	1	.085	92	2.015	.159
Father's qualification	11.164	1	.994	92	11.232	.001
Mother's qualification	305.441	1	.870	92	351.185	.000
Father's occupation	11.565	1	1.232	92	9.385	.003
Mother's occupation	.978	1	.141	92	6.949	.010
Programme	.009	1	.011	92	.842	.361
Sem.	.000	1	.000	92	.	.
Section	.233	1	.251	92	.927	.338

Table 3: ANOVA Test

From the above table the observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal. The ANOVA refers to the analysis of variance and it uses an F test to contrast the group's means. An F distribution is common to distribution of chi-square. The ANOVA F test predicts if there is any rapport between 2 variables. From the analysis the ANOVA F tests represents which variables contribute the highest to the solution of cluster. The variables with biggest errors of mean square offer the least support in distinguishing between clusters. Thus, according to the value presented in ANOVA table, the assets have huge impact in combining the clusters and net profit the least.

The following table illustrates the performance index of the MCA students:

Division	Cut off Percentage	Number of students
First division	More than 60%	77
Second division	Less than 60%	17

Table 4: Performance Index

From the above table, it is clearly understood that, MCA students belonging to the first division have more than 60% of cut-off percentage. At the same time, MCA students belonging to the second division have less than 60% of cut-off percentage. So, it is clearly identified that, MCA students belonging to the first division perform better than the MCA students belonging to the second division.

The overall performance of the MCA students is calculated by applying the deterministic model in the equation: $\frac{1}{N\left\{\sum_{j=1}^{n}\left\{\frac{1}{n}\Sigma_{i=1}^{1} xi\right\}\right\}}$.

Apart from these, the group assessment in each cluster size is calculated by summing the average of individual scores in each cluster.

Where

N represents the total number of students in the cluster and n represents the dimension of the data.

A cluster analysis was conducted by using Weka software. The following table illustrates the cluster size and overall performance of the students when K=2

Cluster#	Cluster size	Overall Performance
1	50	67.74
2	44	83.32

Table 5: K=2

From the above table, it is clearly observed that, for k=2, in cluster 1, the cluster size is50 and so the overall performance is calculated as 67.74%. The cluster size for cluster2 was 44 and it has its overall performance of 83.32%. It is clearly understood that, the students in both the cluster 1 and cluster 2 have an overall performance more than 60% and they belong to the first division. So it is clearly understood that, MCA students have a cut-off with good performance. Following figure illustrates the graph of overall performance vs. cluster size (# of students) when k = 2.

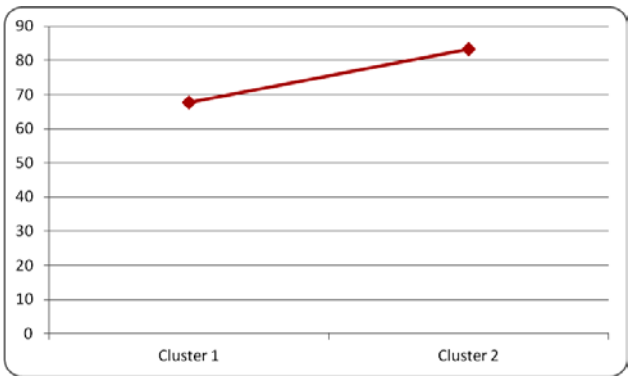

Figure 3: Overall Performance vs. Cluster Size (# of students) k = 2

The following table illustrates the cluster size and overall performance of the students when k=3

Cluster#	Cluster size	Overall Performance
1	23	74.53
2	30	83.97
3	41	68.5

Table 6: K=3

From the above table, it is clearly observed that for k = 3, the cluster size in cluster 1 was23, cluster 2 was 30 and cluster 3 was 41 respectively. By conducting the cluster analysis the overall performance was identified as 74.53%, 83.97% and 68.5% respectively for the clusters. It is clearly understood that, the students in the entire cluster 1, 2 and 3 have an overall performance more than 60% and they belong to the first division. So it is clearly understood that, MCA students have a cut-off with good performance. Following figure illustrates the graph of overall performance vs. cluster size (# of students) when k = 3.

Figure 4: Overall Performance vs. Cluster Size (# of students) k = 3

The following table illustrates the cluster size and overall performance of the students when k=4

Cluster#	Cluster size	Overall Performance
1	43	68.75
2	14	86.44
3	20	73.68
4	17	10.03

Table 7: K=4

From the above table, it is clearly observed that for k = 4, the cluster size for cluster 1, 2, 3and 4 were 43, 14, 20 and 17 respectively. By conducting the cluster analysis the overall performance was identified as 68.75%, 86.44%, 73.68% and 10.03% respectively. It is clearly understood that, the students in cluster 4 which means only 17 have an overall performance 10.03% (which is less than 60%) whom belonging to the second division. Apart from these, the students in cluster 1, 2 and 3 have an overall performance more than 60% and they belong to the first division. So it is clearly understood that, most of the MCA students have cut-off with good performance. Following figure illustrates the graph of overall performance vs. cluster size (# of students) when k = 4.

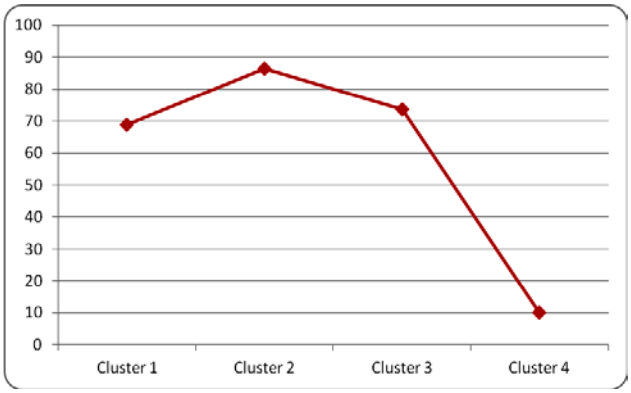

Figure 5: Overall Performance vs. Cluster Size (# of students) k = 4

The following table illustrates the cluster size and overall performance of the students when k=5

Cluster#	Cluster size	Overall Performance
1	14	84.36
2	16	50.45
3	31	79.69
4	18	80.85
5	15	75.07

Table 8: k=5

From the above table, it is clearly observed that for k = 5, the cluster size for cluster 1 was 14, cluster 2 was 16, cluster 3 was 31, cluster 4 was 18 and cluster 5 was 15. By conducting the cluster analysis the overall performance was identified as 84.36%, 50.45%, 79.69%, 80.85% and 75.07% for cluster 1, 2, 3, 4 and 5 respectively. It is clearly understood that, the students in cluster 2 which means only 16 have an overall performance which is less than 60% that belongs to the second division. Apart from these, the students in cluster 1, 3, 4 and 5 have an overall performance more than 60% and they belong to the first division. So it is clearly understood that, most of the MCA students have cut-off with good performance. Following figure illustrates the graph of overall performance vs. cluster size (# of students) when k = 5.

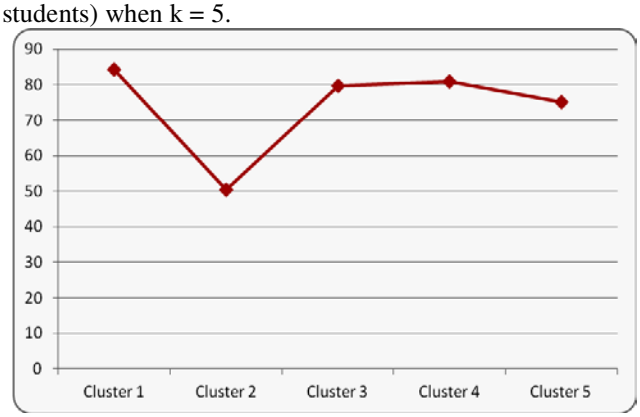

Figure 6: Overall Performance vs. Cluster Size (# of students) k = 5

Thus from this study, it is clearly understood that, while identifying the students overall performance, K-means clustering algorithm provides more effectiveness and also provides accurate results in expected time.

IV. CONCLUSION

In this study, data based on some selected input variables collected through questionnaire method. Some of most influencing factors were identified and taken to predict the performance in semester end examinations. By using k-means Clustering algorithm we can acquire effectiveness on supervising the development of students' academic performance in higher educational institutions to offer exact outcomes in a small time period. The obtained results reveal that 'type of school' is not influence student performance and on the other hand, parent's occupation plays a major ole in predicting performance.

REFERENCES

[1] S. Sujit Sansgiry, M. Bhosle, and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006.

[2] Susmita Datta and Somnath Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," Bioinformatics, vol. 19, pp.459–466, 2003.

[3] Rousseeuw P. J, "A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational Appl Math, vol 20, pp. 53– 65, 1987.

[4] Sharmir R. and Sharan R., "Algorithmic approaches to clustering gene expression data," In current Topics in Computational Molecular Biology MIT Press; pp. 53-65, 2002.

[5] Mucha H. J., "Adaptive cluster analysis, classification and multivarite graphics,"Weirstrass Institute for Applied Analysis and Stochastics, 1992.

[6] Oyelade, O. J., Oladipupo, O. O. and Obagbuwa, I. C., "Applications of k-Means Clustering algorithms for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, No. 1,