PREDICTION OF STUDENT ACADEMIC PERFORMANCE USING CLUSTERING

Dr. G. Rajitha Devi
Asst. Prof. in computer science, Government Degree college, Hayathnagar

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 programs for the weaker ones in order to improve their progress while attempting to a void student dropouts. The aim of this study is to apply the k-means clustering technique to analyze the relationships between students behavioral and their success and to develop the model of student performance predictors. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variables predictors.

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

INTRODUCTION

To identify potential drop outs 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 bases to monitor the progression of student performance in higher Institution of learning. Base on this critical issue, grouping of students into different categories according to their performance has become a complicated task. 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 methods, such as clustering algorithm, 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 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 cannot 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 to 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.

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 componentanalysisbyconsidering16variables.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 data set of *n* data points $x_1,x_2,\ldots,x_n$ such the teach data point is in $\mathbf{R}^{\mathbf{d}}$, the problem of finding the minimum variance clustering of the dataset into *k* clusters is that off in ding *k* points

$\{m_j\}(j=1,2,\ldots,k)$in $\mathbf{R}^{d}$such that

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

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j.The points$\{m_j\}(j=1,2,\ldots,k)$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 agradient descent procedure, which begins at starting cluster cencroids , 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 off in ding the global minimum Is NP-complete. The *k*-means algorithm updates cluster centroid still local minimum is found.Fig.1showsthe generalized pseudo codes of *k*-means algorithm; and traditional k-means algorithm ispresentedinfig.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 centroid seven 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]. Step1: Accept the number of clusters to group data In to and the data set to cluster as input values

Step2: Initialize the first K
clusters
- Take first k instances or
- Take Random sampling of k elements

Step3: Calculate the arithmetic means of each cluster formed in the dataset.

Step4: K- means assign seach record in the data set to only one of the initial clusters
- Each record is assigned to the nearest cluster using a measure of distance (e.gEuclideandistance).

Step5: K- means re- assign search record in the dataset to the most similar cluster and re-calculates
The arithmetic mean of all the cluster sin the dataset.

Fig: Generalized Pseudo code of Traditional k- means

```
MSE= large number;
Select initial cluster centroids {mj}j
K= 1;
Do
    Old MSE=MSE;
    MSE1=0;
    For j=1tok
        mj=0;nj=0;
    end for
    For i= 1ton
        For j=1tok
            Compute squared Euclidean distance $d^2(x_i,m_j)$;
        end for
        Find the closest centroid mjtoxi;
        mj=mj+xi;nj=nj+1;
        MSE1=MSE1+d2(xi,mj);
    end for
    For j=1tok
        nj=max(nj,1);mj=mj/nj;
    end for
    MSE=MSE1;
while(MSE<Old MSE)
```

Fig.2:Traditional *k*-meansalgorithm[6]

RESULTS

We applied the model on the data set (academic result of one semester)for MCA Programme of a University in Jaipur(India). The result generated is shownintables2, 3, and4, respectively.Intable2,for k= 3; in cluster 1, the cluster size is 25 and the overallperformanceis62.22. Also, the cluster sizes and the overall performances for cluster numbers 2 and 3 are 16, 27 and 45.73, and53.03, respectfully. Similar analyses also hold for tables 3 and 4. The graphs regenerated infigures3, 4and5, respectively, where the overall performances plotted against the cluster size.

Table5showsthedimension of the dataset(Student's scores)in the form N by M matrices, where N is the rows (# of students)and M is the column(#of courses)offered by each student. The overall performance Is evaluated by applying deterministic model inEq.2[7]where the group assessmentinea choftheclustersizeis evaluated by summing the average of the individual scores in each cluster.

$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{n}x_s\right)\right)$$

Where

$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{n}x_s\right)\right)$$

N= the total number of students in a cluster and n= the dimension of the data

Table1: PerformanceIndex

Table1: PerformanceIndex

70 and above	Excellent
60-69	Very Good
50-59	Good
45-49	Very Fair
40-45	Fair
Below45	Poor

In Figure3, the overall performance or cluster size

25 is 62.22% while the overall performance for clustersize15 is 45.73%andclustersize 29 has the overallperformanceof53.03%. This analysis showed that, 25 out of 79 students had a "Very Good" performance (62.22%), while 15out of 79 students had performance in the region of very"Fair"performance (45.73%)and the remaining 29studentshada"Good" performance (53.03%)as depicted in the performance indexintable1.Figure4 shows the trends in performance analysis as follows; overall performance for cluster size 24 is 50.08% while the overall performance for cluster size 16 is 65.00%. Cluster size 30 has the overall performance of58.89%, while cluster size 09 is 43.65%. The trends in this analysis indicated that, 24 students fall in the Region of" Good" performance index intable1above(50.08%),while16students has performance in the region of" Very Good" performance (65.00%).30 students has a "Good" performance (58.89%)and 9studentshadperformance of "Fair" result(43.65%). In figure 5, the overall performance for cluster size 19 is 49.85%, while the overall performance for clustersize17is60.97%. Clustersize9hastheoverall performance of43.65%, while the clustersize14 has overall performance of 64.93%andclustersize20has overall performance of55.79%. This performance analysisindicatedthat,19studentscrossedover to "Good" performance region (49.85%), while 17 students had "Very Good" performance results (60.97%).9 students fall in the region of" Fair" performance index(43.65%),14studentswereinthe region of" Very Good" performance (64.93%)and the remaining 20studentshad"Good" performance (55.79%).

Table2:K= 3

Cluster#	ClusterSize	Overall
1	25	62.22
2	15	45.73
3	29	53.03

Table2:K= 4

Cluster#	ClusterSize	Overall Performance
1	24	50.08
2	16	65.00
3	30	58.89
4	9	43.65

Table2:K= 4

Cluster#	ClusterSize	Overall Performance
1	19	49.85
2	17	60.97
3	9	43.65
4	14	64.93
5	20	55.79

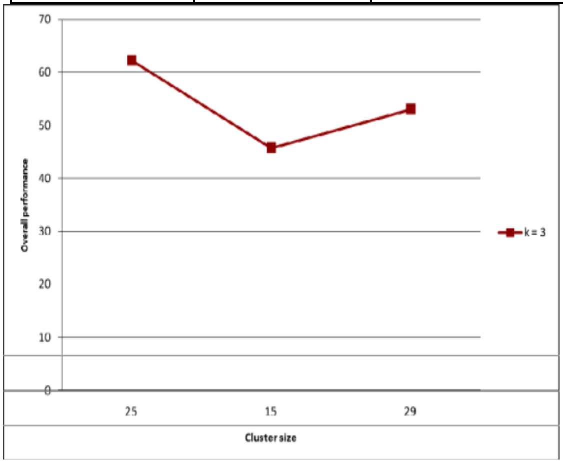


Fig.3: Overall Performance versus cluster size (# of students)k=3

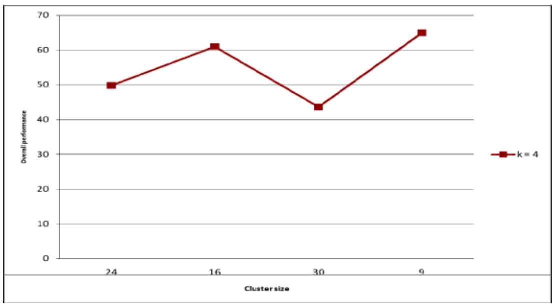
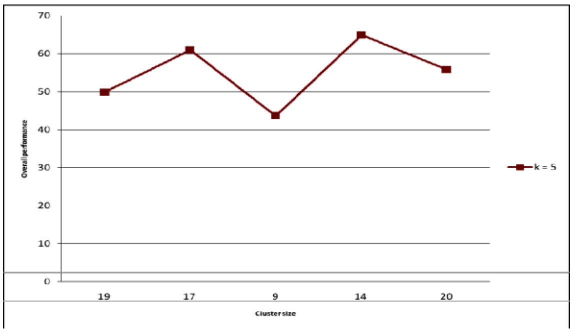


Fig.4: Overall Performance versus cluster size (# of students)k=4



DISCUSSION AND CONCLUSION

In this paper, a simple and qualitative methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. We demonstrated our technique using k-means clustering algorithm[6]and combined with the deterministic model in [7] on a data set of private school results with n in e courses offered for that semester f or each student for total number of 79 students, and produces the numerical interpretation of the results for the performance evaluation. This model improved on some of the limitations of the existing methods, such as model developed by [7]and[8].These models applied fuzzy model to predict students 'academic performance on two dataset only(English Language and Mathematics)of Secondary Schools results. Also there search work by [9] only provides Data Mining frame work for Students' academic performance. The research by[10] used rough Set theory as a classification approach to analyze student data where the Rosetta toolkit was used to evaluate the student data to describe different dependencies between the attributes and the student status where the discovered pattern s are explained in plain English.

Therefore, this clustering algorithm serves as a good benchmark to monitor the progression of students' performance in higher institution .It also enhances the decision making by academic planners to monitor the candidates' performance semester by semester by Improving on the future academic results in the sub sequence academic session.

REFERENCES

[1] S. SujitSansgiry, M. Bhosle, andK. Sail, "Factors that affect academic performance among pharmacy students,"American Journal of PharmaceuticalEducation, 2006.

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

[3] RousseeuwP.J,"Agraphicalaidtothe interpretation and validation of cluster analysis," Journal of Computational ApplMath,vol20,pp.53–65,1987.

[4] SharmirR.andSharanR.,"Algorithmic approaches to clustering gene expression data," Incurrent Topics in Computational Molecular BiologyMITPress;pp.53-65,2002.

[5] MuchaH.J.,"Adaptive cluster analysis, classification and multi varite graphics, Weirstrass Institute for Applied Analysis andStochastics,1992

[6] FahimA.M., SalemA.M.,TorkeyF.A.and Ramadan M. A., "An efficient enhanced k- means clustering algorithm," Journal of ZhejiangUniversityScienceA.,pp.1626–1633, 2006

[7] J.O.Omolehin, J.O.Oyelade,O.O.Ojeniyiand K Rauf, "Application of Fuzzy logic in Decision making on students' academic performance," Bulletin of Pure and Applied Sciences,vol.24E(2),pp.281-187,2005.

[8] J . O. Omoleh in, A. O. Enikuomehin, R. G JimohandK.Rauf,"Profile of conjugate gradient method algorithm on the performance appraisal for a fuzzy system, "African Journal of Mathematics and Computer Science Research,"vol.2(3),pp.030-037,2009

[9] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique," European Journal of Scientific Research, vol.34(4), 2009.

[10] Varapron P. et al., "Using Rough Set theory for Automatic Data Analysis,"29 Congress on Science and Technology of Thailand, 2003.

[11] Oyelade, O. J., Oladipupo, O. O. and Obagbuwa .C.,"Applications of k-Means Clustering algorithms for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security Vol.7,No. 1,2010.