IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Application of clustering algorithm for analysis of Student Academic Performance

A Seetharam Nagesh[1*], Ch V S Satyamurty[2]

[1*]IT Department, CVR College of Engineering, Hyderabad, India
[2]IT Department, CVR College of Engineering, Hyderabad, India

[*]*Corresponding Author: nageshf25@gmail.com, Tel.: +919491242085*

Available online at: www.ijcseonline.org

Received: 30/Dec/2017, Revised: 06/Jan/2018, Accepted: 22/Jan/2018, Published: 31/Jan/2018

Abstract— The analysis of the Student academic performance in educational institutions is a crucial task to make managerial decisions and to impart quality education. The data pertaining to the educational institutions is increasing rapidly. Mining these large volumes of the data will help the management to make academia decisions. Predicting the academic performance of the student at an early stage of their course will help the academia to identify the merit students and also to put more efforts in developing remedial programs for the weaker students to improve their performance. In this paper, we applied k-means clustering algorithm for analysing the students result data and predicting the students' performance.

Keywords— Academic Performance, Data Mining, Student's result data, clustering

I. INTRODUCTION

The data mining algorithms have proved their usefulness in various application domains such as Credit Card frauds, Sports, Health Care, Banking, Education and Insurance. The data mining techniques are used by researchers in the Educational domain [1] and it is known as Educational data mining. The Data mining techniques are applied to extract the hidden patterns from the educational data. The hidden patterns are helpful to the educational institutions to make the decisions. The decisions taken by academia are useful to improve the performance of the weaker students. The classification or clustering techniques are used by most of the researchers to know the implicit patterns from the educational data.

The data mining techniques are broadly classified as supervisory and unsupervisory techniques. The supervisory techniques are helpful for the data already contain known class labels. On the other hand, unsupervisory techniques are also called as clustering techniques [1-2]. The clustering techniques gained interest by researchers. Clustering is broadly used in many applications such as market research, pattern recognition, data analysis, image processing, academic performance and intrusion detection.

The k-means is the old and partition based clustering algorithm. The application of K-means helps to partition the data set into clusters using Euclidean distance as measure

(A) Different Types of Cluster Methods

Clustering is the process of making a group of abstract objects into classes of similar objects. Clustering is the process of grouping the data into clusters such that objects within the same cluster are similar to each other, and dissimilar to every object not in the same cluster. Clustering is an unsupervisory learning. Clustering algorithms can be categorized into partition-based, hierarchical-based, density-based, grid-based and model-based algorithms.

The partition based clustering decomposes the set of data objects into clusters where the number of the resulting clusters is predefined by the user. The partition-based algorithms are K-Means and K-medoids. The Hierarchical clustering method creates a hierarchical decomposition of the given set of data objects. Hierarchical clustering can be either agglomerative (bottom-up) or divisive (top-down). *Density-based clustering* forms the clusters of densely gathered objects separated by sparse regions. Density-based algorithms forms the non-spherical clusters. Grid-based clustering partitions the data space into finite number of cells to form the grid structure on which all the operation for clustering re performed. Grid based clustering is used for spatial data analysis.

Section II consists of Related Work, Section III consists of the Methodology, Section IV consists of Results and Discussion and Section V deals with conclusions.

II. RELATED WORK

Md.Hedayetul Islam Shovon [1] presented a paper on prediction of student academic performance by applying Kmeans clustering algorithm. The student's evaluation factor like class quizzes, mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class advisor before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students.

Monika Goyal and Rajan Vohra (2012) [2] applied data mining techniques to improve the efficiency of higher education institution. If data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. This is an approach to examine the effect of using data mining techniques in higher education.

Ayesha et al. [3], used k-means clustering algorithm as a data mining technique to predict students' learning activities in a students' database including class quizzes, mid and final exam and assignments. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students. These correlated information will be conveyed to the class teacher before the conduction of final exam. This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

M.N. Quadri et al [4] predicted student's academic performance using the CGPA grade system where the data set comprised the student's gender, attendance, his parents educational details, his financial background and so on

Baradwaj and Pal [3], applied the classification as data mining technique to evaluate student performance, they used decision tree method for classification. The goal of their study is to extract knowledge that describes students 'performance in end semester examination. They used students 'data from the student 'previous database including Attendance, Class test, Seminar and Assignment marks. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

Oyelade, O. J [6] presented a method of using K-means clustering algorithm for the prediction of

Students' Academic Performance. The ability to monitor the progress of students' academic performance is a critical issue to the academic community of higher learning. This paper is aims to present a systematic review on different clustering techniques applied for educational data mining to predict academic performance of students and its implications

Sajadin Sembiring [7] et al applied the kernel method as data mining techniques to analyze the relationships between students behavior and their success, and to develop the model of student performance predictors .This is done by using Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering techniques. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variable predictors.

III. METHODOLOGY

(A) Implementation of K-Means Clustering Algorithm

The K-Means is the unsupervised algorithm used for clustering the data objects. The K-Means clustering algorithm partitions the 'n' data objects into 'k' clusters (groups), in which each data object belongs to the cluster with the nearest mean. The data objects in each group are highly cohesive and the object in other group are disjoint. The K-means algorithm creates 'k' distinct group of elements using the sum of squares. The input parameter for the algorithm is number of centroids. Then the distance between each element with each centroid is calculated. The distances calculated for a data element with each centroid is compared and assigned the data element to nearest centroid. In this way, all the data elements are assigned to one of the centroid. Initially, K clusters are formed by assigning the data elements to respective centroid. Then recalculate the centroid of assigned data elements in each cluster. Once again with the new centroids calculate the distance between each data element with the new centroids and reassign the data element to centroid which is near. This process continues until the no data element assigns to any new centroid, that means the centroids of 'n-1' iteration is equal to centroid of 'n' iteration.

The distance measure in K-means clustering is Euclidean distance. Suppose the elements are $X=(x_1,x_2,x_3\ldots)$ and $Y=(y_1, y_2,y_3,\ldots)$.

$$D(X, Y) = \underline{(x_1-y_1)^2 + (x_2-y_2)^2} + \ldots.$$

$$\sqrt{\sum_{i=1}^{n} (xi - yi)^2} \qquad (1)$$

Using the eq.1, the distance between each data element and the centroid is calculated. And the data element is assigned to the centroid with the minimum distance. The centroid is the mean of all data points in that group. Each centroid with set of the data elements is called as cluster.

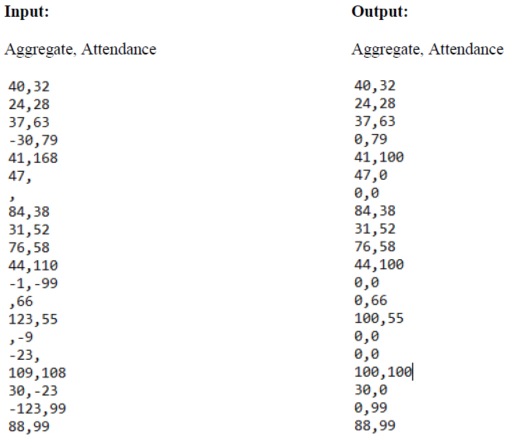
k-means Algorithm:

1. Accept the number of clusters to group data into and the dataset to cluster as input values
2. Initialize the first K clusters
   a. Take first k instances or
   b. Take Random sampling of k elements
3. Calculate the arithmetic means of each cluster formed in the dataset.
4. K-means assigns each record in the dataset to only one of the initial clusters
   a. Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

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

IV. RESULTS AND DISCUSSION

The Students' academic performance is given considerable importance by various committees like NBA, NAAC inspecting the colleges, especially technological institutions. The dataset used for the experiment was obtained from the Information Department of the engineering college. The attributes aggregate and attendance were considered for experiment purpose. In the first step the data set is normalized for missing values and erroneous values by using min-max normalization. The sample resultant dataset after normalization is shown in fig.1



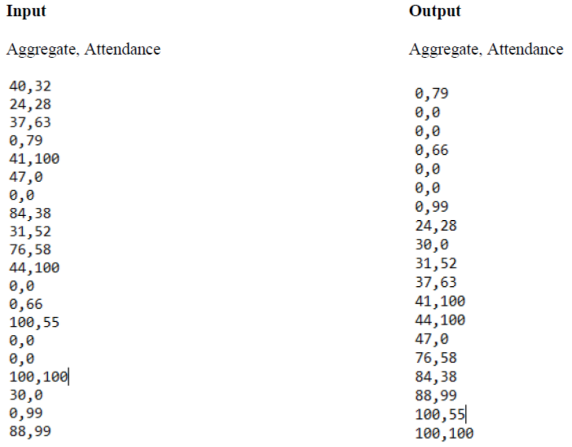
```
Input:                          Output:

Aggregate, Attendance           Aggregate, Attendance

40,32                           40,32
24,28                           24,28
37,63                           37,63
-30,79                          0,79
41,168                          41,100
47,                             47,0
,                               0,0
84,38                           84,38
31,52                           31,52
76,58                           76,58
44,110                          44,100
-1,-99                          0,0
,66                             0,66
123,55                          100,55
,-9                             0,0
-23,                            0,0
109,108                         100,100
30,-23                          30,0
-123,99                         0,99
88,99                           88,99
```

Figure 1:Input and Output values of Normalization

Before applying the k-means clustering algorithm on the normalized dataset, the dataset is sorted on the aggregate values. The sample dataset before and after sorting are shown in fig.2



```
Input                           Output

Aggregate, Attendance           Aggregate, Attendance

40,32                           0,79
24,28                           0,0
37,63                           0,0
0,79                            0,66
41,100                          0,0
47,0                            0,0
0,0                             0,99
84,38                           24,28
31,52                           30,0
76,58                           31,52
44,100                          37,63
0,0                             41,100
0,66                            44,100
100,55                          47,0
0,0                             76,58
0,0                             84,38
100,100                         88,99
30,0                            100,55
0,99                            100,100
88,99
```

Figure 2:Input and Output after Sorting

Now, by applying the k-means clustering algorithm on the sored data, the student data is grouped into three classes "Poor", "Average" and "Good". The result is shown in the fig.3, where green color indicates the cluster of good, red indicates the cluster of average and blue indicates the cluster of the poor.

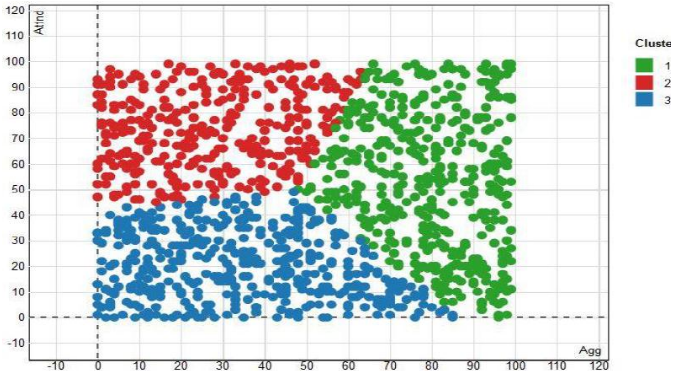


Figure 3:Final Output after Clustering

V. V. CONCLUSIONS

In this paper, we analysed the student data and predicted the percentage of students whose academic performance is poor, average and good by making use of k-means clustering algorithm. The 50% of the data clustered as good shown in green. If management taken appropriate measures to improve the academic performance of the students from average and poor clustered. This simple analysis work shows that the proper data mining application on student's performance can be efficiently used for hidden knowledge / information retrieval from the vast data, which can be used for the process of decision making by the management of an educational institution. We hope that the information generated after the implementation of data mining and data clustering technique may be helpful for instructor as well as for management. For future work, we hope to refine our technique in order to get more valuable and accurate outputs, useful for instructors to improve the students learning outcomes.

REFERENCES

[1] H. Jiawei. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, San Francisco California, Morgan Kaufmann Publishers, 2012.

[2] A.K. Jain, Data clustering: "50 years beyond K-means, *Pattern Recognition Letters", Elsevier,* vol.31, pp.651-666, 2010.

[3] Md. Hedayetul Islam Shovon, "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm*", International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2(7), July 2012.

[4] Monika Goyal, Rajan Vohra, “Applications of Data Mining in Higher Education”, International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012

[5] Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) ‗Data Mining Model for Higher Education System‘, European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[6] M. N. Quadri1 and Dr. N.V. Kalyanka- Drop Out Feature of Student Data for Academic Performance Using Decision Tree, Global Journal of Computer Science and Technology Vol. 10 Issue 2 (Ver 1.0), April 2010.

[7] Baradwaj, B. and Pal, S. (2011) ‗Mining Educational Data to Analyze Student s‘ Performance‘, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[8] Oyelade, O. J, “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance”, *(IJCSIS) International Journal of Computer Science and Information*
Security, Vol.7, 2010.

[9] Sajadin Sembiring, M. Zarlis, Dedy Hartama,Ramliana S, Elvi Wani. Prediction of Student Academic Performance by An Application of Data Mining Techniques, International Conference on Management and Artificial Intelligence, Bali, Indonesia,IPEDR vol.6 ,pp.110-114,2011.

[10] P. Ajith, M.S.S.Sai, B. Tejaswi, Evaluation of Student Performance: An Outlier Detection Perspective, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-2, January, 2013.

[11] D. Kabakchieva, “Analyzing University Data for Determining Student Profiles and Predicting Performance”, *Cybernetics and Information Technologies*, Vol.1(3), March 2013.