RESEARCH ARTICLE OPEN ACCESS

Efficient K-Means Clustering Algorithm for Predicting of Students' Academic Performance

Ei Ei Phyo [1], Ei Ei Myat [2]
Department of Information Technology
Technological University (Thanlyin)
Myanmar

ABSTRACT
The academic leaders encourage in promoting the education of students who can lead the nation one day. Therefore, monitoring the progress of student's academic performance is an important issue to the academic community. To achieve this issue, there is a system that consists of two parts. Firstly, the student's results are analysed based on cluster analysis and second the standard statistical algorithm is used to arrange their results data according to the level of their performance. In this paper, efficient k-means clustering algorithm is implemented to analyse the students' data. The efficient k-means clustering can be used to monitor the progress of the students' performance and to make the decisions by the teachers/academic planner in each semester for the improving of the future academic results of the students.

Keywords :— clustering, k-means clustering, initial centroid, min-max normalization, gain ratio.

I. INTRODUCTION

Every university set commonly used indicators of academic performance is Grade Point Average (GPA). But grade system is also used some of the public schools/university. This grade system and GPA scales may vary significant. But it is important to determine specifically how grade and GPA are calculated to measure academic performance. It is difficult to obtain a comprehensive view of the state of the students' performance and simultaneously discover important details from their performance with grouping of students based on their average scores [1]. In this paper presents k-means clustering as a simple and efficient tool to monitor the progression of students' performance in higher institution.

Data mining tools and techniques are used to generate an effective result which was earlier difficult and time consuming. Data mining is widely used in various areas like financial data analysis, retail and telecommunication industry, biological data analysis, fraud detection, spatial data analysis and other scientific applications.

Clustering is a technique of data mining in which similar objects are grouped into clusters. Clustering techniques are widely used in various domains like information retrieval, image processing, etc. [2] [3]. There are two types of approaches in clustering: hierarchical and partitioning. In hierarchical clustering, the clusters are combined based on their proximity or how close they are. This combination is prevented when further process leads to undesirable clusters. In partition clustering approach, one dataset is separated into definite number of small sets in a single iteration [4].

The most widely used clustering algorithm is the K-means algorithm. This algorithm is used in many practical applications. It works by selecting the initial number of clusters and initial centroids [5] [6]. Moreover initial centroids are chosen randomly due to which clusters produced vary from one run to another. Also various data points exist on which K- means takes super polynomial time [7] [8]. Different researchers have put forward various methods to improve the efficiency and time of K-means algorithm. K-means uses the concept of Euclidean distance to calculate the centroids of the clusters. This method is less effective when new data sets are added and have no effect on the measured distance between various data objects. The computational complexity of k means algorithm is also very high [2] [9]. Also, K-means is unable to handle noisy data and missing values. Data pre-processing techniques are often applied to the datasets to make them cleaner, consistent and noise free.

Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences [10].

A feature weight algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets [11].

II. K-MEANS CLUSTERING ALGORITHM

The basic idea of K-means algorithm is to classify the dataset D into k different clusters where D is the dataset of n data; k is the number of desired clusters. The algorithm consists of two basic phases [12]. The first phase is to select the initial centroids for each cluster randomly. The second and final phase is to take each point in dataset and assign it to the nearest centroids [12]. To measure the distance between points Euclidean Distance method is used. When a new point is assigned to a cluster the cluster mean is immediately updated by calculating the average of all the points in that cluster [10].

After all the points are included in some clusters the early grouping is done. Now each data object is assigned to a cluster based on closeness with cluster centre where closeness is measured by Euclidean distance. This process of assigning a data points to a cluster and updating cluster centroids continues until the convergence criteria is met or the centroids don't differ between two consecutive iterations. Once, a situation is met where centroids don't move any more the algorithm ends. The k-means clustering algorithm is given below.

Step 1: Begin with a decision on the value of k = number of clusters.

Step 2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:
1. Take the first k training sample as single-element clusters
2. Assign each of the remaining (N-k) training samples to the cluster with the nearest centroid.
After that each assignment, recomputed the centroid of the gaining cluster.

Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

III. METHODOLOGY

A. *Min-Max Normalization*

Min-max normalization [11] performs a linear transformation on the original data. Min-max is a technique that helps to normalize a dataset. It will scale the dataset between the 0 and 1. Suppose that min_A and max_A are the minimum and maximum values of an attribute, *A*. Min-max normalization maps a value, *v*, of *A* to *v'* in the range [new_min_A, new_max_A] by computing

$$v' = \frac{v - minA}{maxA - minA}\ (new\_maxA - new\_minA) + new\_minA$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for *A*.

B. *Gain Ratio*

Information gain applied to attributes that can take on a large number of distinct values might learn the training set too well. The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. The gain ratio [13] is defined as

$$\text{GainRatio (A)} = \frac{Gain(A)}{SplitInfo(A)}$$

The attribute with the maximum gain ratio is selected as the splitting attribute. The split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large at least as great as the average gain over all tests examined.

C. *Proposed Methodology*

The proposed efficient k-means clustering is upgraded the origin k-means clustering to reduce the computational complexity. In the proposed efficient k-means clustering method, the normalization and feature weight are applied. Firstly, the methodology employs normalized dataset by using min-max normalization to improve the efficiency of clustering algorithm. After that gain ratio method compute feature weights for each attributes of the data to minimize the error rate. It the centroids are then posted to the traditional clustering algorithms for being executed in the way it normally does. The results of the proposed work are validated against number of iterations and accuracy obtained and compared with the randomly selected initial centroids.

Step 1: Accept the dataset to cluster as input values

Step 2: Perform a linear transformation on the original dataset using mix-max normalization

Step 3: Compute the feature weight for each attribute and update the dataset.

Step 4: Initialize the first K cluster

Step 5: Calculate centroid point of each cluster formed in the dataset.

Step 6: Assign each record in the dataset for only one of the initial cluster using a measure Euclidean distance.

Step 7: Repeat step 4 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

IV. RESULTS

We tested the dataset (academic result of one semester) of a Technological University (Thanlyin). The performance index of the grade system is shown in Table 1.

Table 1. Performance Index

81 and above	Grade A
61 – 80	Grade B
41 - 60	Grade C
21 - 40	Grade D
0 - 20	Grade E

In Table 2, all marks of 7 courses of 66 students in Technological University (Thanlyin).

Table 2. Statistics of the Data used

Student's Scores	Number of Students	Total number of courses
Data	66	7

For K = 3, the result is generated as shown in Table 3. In cluster 1, the cluster size is 29 and the overall performance is 39.57. The cluster numbers 2 and 3 are 18, 19 and the overall performance are 56.43 and 69.93, respectfully.

Table 3. K= 3

Cluster#	Number of Students	Overall Performance
1	29	39.57
2	18	56.43
3	19	69.93

In Table 4, there are 4 clusters. The cluster sizes are 1, 2, 3 and 4 are 17, 16, 16, 17 and the overall performance are 32.57, 54.14, 67.79 and 66.43, respectively.

Table 4. K= 4

Cluster#	Number of Students	Overall Performance
1	17	32.57
2	16	54.14
3	16	67.79
4	17	66.43

For K = 5, the result is generated as shown in Table 5. The cluster sizes are 1, 2, 3, 4 and 5 are 11, 13, 17, 10, 15 and the overall performance are 30.00, 59.00, 52.71, 67.36 and 70.50, respectively.

Table 4. K= 5

Cluster#	Number of Students	Overall Performance
1	11	30.00
2	13	59.00
3	17	52.71
4	10	67.36
5	15	77.50

In Figure 1, the cluster 1 analysis showed that, 29 out of 66 students had a "Grade D" performance (39.57%). The cluster 2 analyses showed that, 18 out of 66 students had performance in the region of very "Grade C" performance (56.43%) and the remaining 19 students had a "Grade B" performance (69.9%).

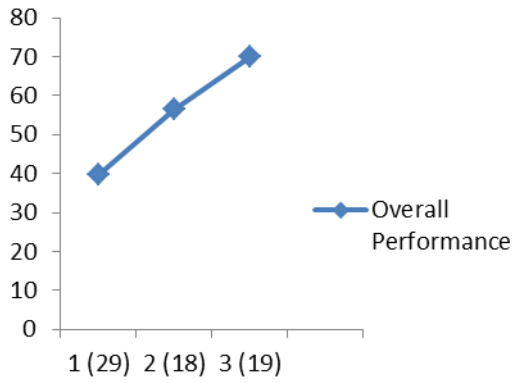

Figure 1. Overall performance of students k = 3

Figure 2 shows the trends in this analysis indicated that, 17 students fall in the region of "Grade D" performance index in table 1 above (32.57%), while 16 students has performance in the region of "Grade C" performance (54.14%). 16 and 17 students has a "Grade B" performance (67.79%) and (66.43%).

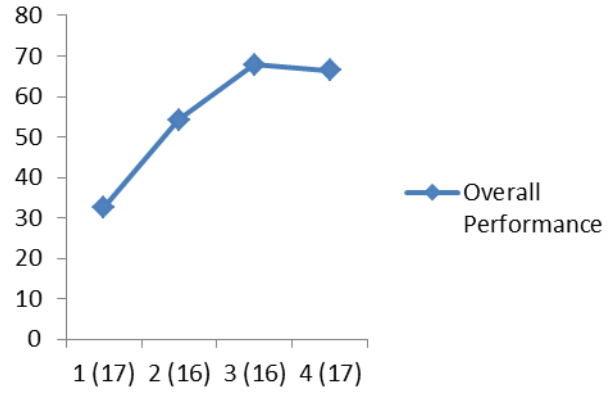

Figure 2. Overall performance of students k = 4

In Figure 3, The overall performance analysis indicated that, 11 students crossed over to "Grade B" performance region (39%), while 13 students had "Grade C" performance results (59%). 17 students fall in the region of "Grade C" performance index (52.71%), 10 students were in the region of "Grade B" performance (67.36%) and the remaining 15 students had "Grade B" performance (77.50%).

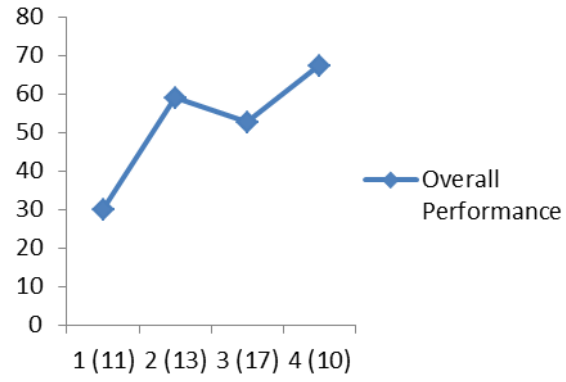

Figure 3. Overall performance of students k = 5

V. CONCLUSIONS

In this paper, the methodology to compare k-means clustering and efficient k-means clustering. The proposed algorithm using k-means clustering and combine with the min-max normalization and gain ratio. A data set of results with seven courses offered for that semester for each student for total number of 66 students, and produces the numerical interpretation of the results for the performance evaluation. The efficient k-means algorithm serves as a good mark to monitor the progression of students' performance in higher institution. The decision making by academic planners to monitor the candidates' performance semester by semester by improving on the future academic results in the subsequence academic session.

REFERENCES

[1] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C, "Application of k-Means Clustering Algorithm for prediction of Students' Academic Performance", IJCSIS International Journal of Computer Science and Information Security, Vol. 7, No 1, 2010, pp. 292-295.

[2] Md Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", 7th International Conference on Electrical and Computer Engineering, 2012, pp. 647-650.

[3] Madhu Yedla, Srinivasa Rao Pathakota, TM Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies (IJCSIT),Vol.1 (2), 2010, pp. 121-125.

[4] Margaret H. Dunham, "Data Mining Introductory and Advanced Concepts", Pearson Education, 2006.

[5] Vaishali R. Patel, Rupa G. Mehta, "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, 2011, pp. 331-336.

[6] McQueen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., Vol. 1, 1967, pp. 281–297.

[7] Zhang Chen, Xia Shixiong, "K-means Clustering Algorithm with improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining, 2009, pp. 790-792.

[8] David Arthur & Sergei Vassilvitskii, "How Slow is the k means Method?", Proceedings of the 22nd Symposium on Computational Geometry (SoCG), 2006, pp. 144-153.

[9] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress On Engineering 2009 Vol I, WCE 2009, pp. 308-312.

[10] Margaret H. Dunham, Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006

[11] Navdeep Kaur and Krishan Kumar, "Normalization Based K-means Data Analysis Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 6, Issue, June 2016, pp. 455-457

[12] A.K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters 31 (2010) 651–666.

[13] Ester Martin, Kriegel Hans-Peter introduced the Idea of "Clustering for Mining in Large Spatial Database".