PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE BY K-MEANS CLUSTERING

Yann Ling Goh*

Department of Mathematical and Actuarial Sciences
Universiti Tunku Abdul Rahman, Malaysia
gohyl@utar.edu.my

Yeh Huann Goh

Department of Mechanical Engineering
Kolej Universiti Tunku Abdul Rahman, Malaysia
gohyh@tarc.edu.my

Chun-Chieh Yip

Department of Civil Engineering
Universiti Tunku Abdul Rahman, Malaysia
yipcc@utar.edu.my

Chen Hunt Ting

Department of Mechanical and Material Engineering
Universiti Tunku Abdul Rahman, Malaysia
tingch@utar.edu.my

Raymond Ling Leh Bin

Department of Accountancy
Universiti Tunku Abdul Rahman, Malaysia
linglb@utar.edu.my

Kah Pin Chen

Department of Mechanical and Material Engineering
Universiti Tunku Abdul Rahman, Malaysia
chenkp@utar.edu.my

***Corrospounding author's Email: gohyl@utar.edu.my**

Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee

<http://connectingasia.org/scientific-committee/>

© 2020 Published by Readers Insight Publisher,

lat 306 Savoy Residencia, Block 3 F11/1, 44000 Islamabad. Pakistan,

editor@readersinsight.net

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Abstract

Schooling system should offer finest teaching and learning opportunities to reach the educational requirements and ensuring achievement for every student. All teachers monitor their students' progress throughout the year, includes formative assessment, questioning, providing feedback, etc. This practice helps teachers continually assess students' academic performance and evaluate the effectiveness of their teaching. In this paper, method of k-means clustering with deterministic model is applied to analyze the student's overall performance. The results is important for educators to identify students who are at risk academically and areas where teaching strategies may need adjustment to better meet these students' needs.

Research Highlights

Typically, k-means algorithm is applied to numeric and continuous data [1]. It is commonly used in wireless sensor networks, pattern recognition, document classification, rideshare data analysis, diagnostic systems, etc. The advantages of k-means clustering are that it is relatively simple to perform, assemble stable and tight clusters, converges after some iterations and low computational cost [2].

K-means clustering is used to categorize and sort a given data set by specify the number of clusters k at the beginning. There are k centroids for k clusters. The algorithm is used to find observational groups that have not been specifically chosen in the data. The algorithm aims to obtain better data arrangement in order to work out an appropriate decision. The results is that similar characteristics data will be grouped within the same cluster, but the clusters themselves are disparate. The target is to secure data points as homogeneous as possible in the same cluster and data points as heterogeneous as possible in the opposed cluster [3].

Research Objectives

The aim of clustering in the paper is to partition students into homogenous groups according to their academic achievements, as measured by assessment scores [4]. These applications can help both the instructors and student to improve the quality education. The teacher can analyze different causes of low academic achievement and introducing effective teaching-learning methods. The new teaching strategy may motivate students to study and progress in their academic performance [5-6].

Methodology

The initial set of centroids are selected randomly, which are used as the initial points for every cluster. Then, individual data point is assigned to be a member of a group by iterative calculations. The K-means algorithm attempts to gather objects which have similar attributes into same group while dissimilar attributes into different groups. It assigns data points to a group in which each data point is in close proximity with the cluster's centroid. The methodology intends to have familiar features objects gather within clusters, so that homogeneity are found within the same cluster.

The k-means algorithm can be summarized as below:

1. Choose a value k to classify the data into k clusters.
2. Select cluster centre randomly for each cluster.
3. Assign data point to a cluster which the data point have the shortest distance with that cluster centre.
4. Determine the new cluster centre for each cluster.
5. The process repeats from step 2 until there is no data points change group.

The purpose of the algorithm is to minimize the sum of squares of distances between data points and the corresponding cluster centre. The objective function is

$$\sum_{i=1}^m \sum_{j=1}^n |d_j^{(i)} - c_i|^2 \quad (1)$$

where

m = total number of students in a cluster

n = total number of subjects taken by each student

$|d_j^{(i)} - c_i|^2$ denotes sum of squares of distances between the cluster centre c_i and a data point $d_j^{(i)}$ while $\sum_{i=1}^m \sum_{j=1}^n |d_j^{(i)} - c_i|^2$ denotes total distance of all the data points from their corresponding cluster centres.

Results

We applied equation (1) on the data set (student's scores in academic year 2019) of a college. The number of students involved in analysis is 106 and dimensions (total number of subjects) are 8. The Annual Average Mark (AAM) achieved by student across all subjects attempted in an academic year is calculated. The AAM is out of 100.

We apply the deterministic model in equation (2) & (3) to find the overall performance. The assessment in every cluster size is evaluated by finding the average mark in each cluster.

$$AAM = \frac{\text{sum of (credit value * marks of each subject)}}{\text{sum of credit values}} \quad (2)$$

$$\text{average mark} = \frac{1}{N} \sum_{j=1}^N AAM \quad (3)$$

where

N = total number of students in a cluster

The grades are being associated with various percentage intervals to compare performance of the students. The results derived are the students having above 70 are in A category, marks between 65-69, 60-64, 50-59 are in A-, B, C category respectively. Marks less than 50 are in F category.

For $k = 6$, cluster size 12 (cluster 1) had been listed as having overall performance 70.66% while cluster size 20 (cluster 2) had been listed as having overall performance 64.03%. Cluster size 14 (cluster 3) has overall performance 73.66% while cluster size 19 (cluster 4) has overall performance 58.22%. Cluster size 24 (cluster 5) and 17 (cluster 6) has the overall performance 66.37% and 63% respectively. This performance analysis indicates that 26 students are in A category (24.53%), 24 students are in A- category (22.64%), 37 students are in B category (34.91%) and 19 students are in C category (17.92%).

Figure 1 shows that the highest scoring subject is Physics (88.43%) for cluster 1, English (80.12%) for cluster 2, Physics (86.30%) for cluster 3, English (80.12%) for cluster 4, English (81.08%) for cluster 5 and Physics (77.3%) for cluster 6. The lowest scoring subject for cluster 1 is History (51.67%). Additional Mathematics is the lowest scoring subject for cluster 2, 3, 4 and 5. The scores are 46.06%, 65.21%, 28.54% and 40.05% respectively. For cluster 6, Mathematics (49.97%) is recorded as the lowest scoring subject.

Figure 1: Students' Overall Performance for Each Subject for Cluster $k = 6$

Findings

K-means clustering algorithm and deterministic model are used to evaluate the academic performance of a college students. This methodology will assist academic planners in measuring students' academic performance and assessing students' progression whether students are meeting course requirements. This methodology provides an indication of overall academic performance and steps that need to be taken to improve students' academic performance in next academic year.

REFERENCES

- [1] Jafar OM. A New Hybrid Hard-Fuzzy (K-MFCM) Data Clustering Method for Finding Cluster Centroid. *International Journal of Advanced Research in Computer Science*. 2018 Mar 1;9(2):626.
- [2] Chauhan M, Gupta V. Comparative Study of Techniques Used in Prediction of Student Performance. *World Scientific News*. 2018;113:185-93.
- [3] Hamoud AK. Classifying Students' Answers using Clustering Algorithms Based On Principle Component Analysis. *Journal of Theoretical & Applied Information Technology*. 2018 Apr 15;96(7).

- [4] Uddin H, Nafis MT. Students Academic Performance using Partitioning Clustering Algorithms. *International Journal of Advanced Research in Computer Science*. 2017 May 1;8(5).
- [5] Dengen N, Budiman E, Wati M, Hairah U. Student Academic Evaluation using Naïve Bayes Classifier Algorithm. In 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT) 2018 Nov 6 (pp. 104-107). IEEE.
- [6] Yadav RS, Singh VP. Modeling Academic Performance Evaluation using Soft Computing Techniques: A Fuzzy Logic Approach. *International Journal on Computer Science and Engineering*. 2011 Feb;3(2):676-86.