Students' Academic Performance Analysis by K-Means Clustering for Investigating Students' Health Conditions within Clusters

Yann Ling Goh*, Yeh Huann Goh †, Chun-Chieh Yip ‡,
Chen Hunt Ting §, Raymond Ling Leh Bin ‡, Kah Pin Chen §

Departments and Address

*Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia
† Department of Mechanical Engineering, Kolej Universiti Tunku Abdul Rahman, Jalan Genting Kelang, Setapak, 53300 Kuala Lumpur, Malaysia
‡ Department of Civil Engineering, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia
§ Department of Mechanical and Material Engineering, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia

*Corresponding author Email: *gohyl@utar.edu.my

Abstract

Schooling systems always offer finest teaching and learning opportunities to reach the educational requirements and ensuring achievement for every student. However, health affects students’ academic performance directly. All teachers monitor their students’ progress throughout the year, includes formative assessment, attendance rates, involvement in the organization, etc. This practice helps teachers continually assess the conditions of students and their academic performance. Data mining is a process to explore certain style and hidden correlation among massive volume of data. Data mining is applied in multiple disciplinary fields, for example, insurance, education, banking and bioinformatics. Data mining skills such as clustering, classification, regression and prediction are commonly used by educators to measure academic performance. In this paper, method of k-means clustering with deterministic model is applied to analyze the student’s overall performance. The students’ assessment scores are assigned to k clusters without prior knowledge of the scores. The result is important for educators to further investigate the effect of sickness of students within a cluster that may lead to poor academic performance.

Keywords: k-means clustering; deterministic model; academic performance


1 Introduction

Data mining techniques are widely used in statistics, financial banking, healthcare, artificial intelligence, marketing, etc. Examples of data mining techniques are patterns tracking, classification, association, outlier detection, clustering, regression and prediction. The techniques involve diagnose patterns from the data sets, classify various facets in the dataset into discernible families, verify anomalies facets or facets that are related mutually, form blocks of data based on features similarities, formulate relationship between target and predictors, recognize historical trends to scheme a forecast in the future. The benefits of data mining techniques are companies manage to make profitable refinement on products and business plans, organizations manage to capture knowledge based on a collection of information, assist company management team to come up with decision, etc. In addition, data mining is a lucrative and competent solution when compare with other statistical methods (Basri et al., 2019; Del Río et al., 2016; Harwati et al., 2015).

Data Mining has been applied extensively in retail stores. From consumer purchasing data, manager can identify consumers’ shopping behaviors and market trends in order to develop more effective marketing strategies and upgrade the quality of customer service. The new business plan may help in retaining customers and increase sales (Humamuddin et al., 2017). In telecommunication industry, data mining assists in detecting fraudulent patterns and analyzing complex association data to improve telecommunication services. For financial industry, data mining is used to analyze loan policy, credit card policy and spot money laundering. Furthermore, in biomedical research, data mining helps in analyzing numerous nucleotide sequences and genetic data patterns (Bansal et al., 2017).

The k-means clustering algorithm is a data mining which can be used for unsupervised machine learning. Typically, it is applied to numeric and continuous data. It is commonly used in wireless sensor networks, pattern recognition, document classification, rideshare data analysis and diagnostic systems (Jafar, 2018). The advantages of k-means clustering are that it is
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 result 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 (Hamoud, 2018).

The aim of clustering in the paper is to partition students into homogenous groups according to their academic achievements, as measured by assessment scores. These applications can help both the instructors and students to improve the quality education. The teachers can analyze different causes of low academic achievement, such as students’ demographic, lack concentration in studies due to sickness, stress, etc. This analysis is important because health conditions of the students are affecting their studies (Haviluddin et al., 2018; Yadav & Singh, 2011).

2 Literature Review

In a study for Levin (2017), one-stage clustering approach, multi-stage clustering approach were described. Examples of clustering in network application were also considered. Partitioned Word2Vec-LDA, which was refer as novel clustering model was proposed by Li et al. (2018) to handle clustering in narrow-domain short texts difficulties. A novel design was carried out and regular words were taken out from the sentences. The proposed model used abstract text to cluster and preliminary findings revealed that the solutions of clustering obtained by applying PW-LDA were much more precise and solid.

A study for Fedorov et al. (2017) introduced the procedure of distributed clustering tutoring designs of vocal sounds. The clustering procedure based on K-means algorithm includes initialization, computation of cluster centers, distance calculation, guideline of the termination circumstance and computation of the target function. Nagaraj et al. (2019) used fraternal K-median clustering algorithm as the preprocessing schemes. The research applied dropout regularization mechanism to improve the precision by letting network over fit decision of CNN.

Sotiropoulos et al. (2016) dealt with clustering issue problem, introduced an encrypting strategy that was centroid based, developed the clustering question and suggested genetic algorithm. The intra and inter cluster semantic spacing of the cluster arrangements were determined to confirm the suggested genetic algorithm of clustering. Swardepp & Pandya (2016) mentioned that there are 5 types of clustering methods: hierarchical clustering method, density based clustering method, grid based clustering method, model based clustering method and partition clustering method. These clustering methods solve some topics of clustering severally and they are singular in nature.

Chagas et al. (2019) indicated that non-disjoint clusters distinguish many clustering applications in real world. The research proposed substitution method of cluster editing to partition a graph. In this work, experiments conclusions showed that proposed hybrid heuristic achieved superior standard overlapping cluster rewriting results. Yu et al. (2018) mentioned that K-means algorithm provides an adequate mechanism for grouping similar data into the same cluster, sensitive to deviations and receptive to original cluster centers. The bi-layer k-means algorithm can overcome the problems of noisy data and outliers in the same cluster.

3 Methodology/Materials

The initial set of centroids is 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 appears within the same cluster.

The algorithm of k-means clustering is 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

$$\sum_{i=1}^{m} \sum_{j=1}^{n} |d_{ij}^{(0)} - c_{i}|^2$$

Where $m = \text{total number of students in a cluster}$
$n = \text{total number of subjects taken by each student}$

http://doi.org/10.36295/ASRO.2020.231332
4 Results and Findings

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 students are required to take eight subjects namely Mathematics, Biology, Malay Language, Additional Mathematics, Chemistry, English, History and Physics. The Annual Average Mark (AAM) achieved by student across all subjects attempted in an academic year is calculated. The AAM is out of 100.

To calculate AAM, we multiply the marks of each subject by the credit value of the subject, then add these totals together. These totals are then divided by the sum of all credit values. The overall performance in every cluster is evaluated by finding the average AAM in each cluster. We apply the deterministic model in equation (2) & (3) to find the overall performance. Table 1 shows the credit value of each subject assigned by the college management team.

\[ \text{AAM} = \frac{\text{sum of (credit value * marks of each subject)}}{\text{sum of credit values}} \]  
\[ \text{overall performance (\%)} = \frac{1}{N} \sum_{j=1}^{n} \text{AAM} \]  

Where \( N \) = total number of students in a cluster

The grades are being associated with various overall performance intervals to compare performance of the students. The results derived are the overall performance above 70% are in category of A, overall performance between 65%-69%, 60%-64%, 50%-59% are in A-, B, C category respectively. Overall performance less than 50% are in F category. Table 2 shows the grade of each overall performance interval.

<table>
<thead>
<tr>
<th>Table 1. Credit value of each subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
</tr>
<tr>
<td>Mathematics</td>
</tr>
<tr>
<td>Biology</td>
</tr>
<tr>
<td>Malay Language</td>
</tr>
<tr>
<td>Additional Mathematics</td>
</tr>
<tr>
<td>Chemistry</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>History</td>
</tr>
<tr>
<td>Physics</td>
</tr>
</tbody>
</table>

Table 3, 4 and 5 show the students’ overall performance when grouping students’ scores into four, five and six clusters. In table 3, the overall performance is 72.11% for cluster size 27 (cluster 1). The overall performance for cluster size 32 (cluster 2), 27 (cluster 3), 20 (cluster 4) are 65.48%, 63.64% and 58.45% respectively. The analysis shows that 27 students are in A category (25.47% of total students), 32 students are in A- category (30.19% of total students), 27 students are in B category (25.47% of total students) and the remaining 20 students are in C category (18.87% of total students).

In table 4, cluster size 24 (cluster 1) and cluster size 25 (cluster 5) had been listed as having overall performance 73.09% and 66.35% respectively. The overall performance for cluster size 16 (cluster 2), 20 (cluster 3), 21 (cluster 4) are 62.59%, 64.41% and 58.45% respectively. The trends in this analysis indicate that 24 students obtain grade A (22.64% of total students), 25 students obtain grade A- (23.58% of total students), 36 students obtain grade B (33.96% of total students) and 21 students obtain grade C (19.81% of total students).

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

<table>
<thead>
<tr>
<th>Table 3. Students’ overall performance when grouping students’ scores into four clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Table 4. Students’ overall performance when grouping students’ scores into five clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster size</th>
<th>Overall performance (%)</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>73.09</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>62.59</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>64.41</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>58.45</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>66.35</td>
<td>A-</td>
</tr>
</tbody>
</table>

Table 5. Students’ overall performance when grouping students’ scores into six clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster size</th>
<th>Overall performance (%)</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>70.66</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>64.03</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>73.66</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>58.22</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>66.37</td>
<td>A-</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>63</td>
<td>B</td>
</tr>
</tbody>
</table>

Figure 1-3 show the students’ overall performance for each subject when grouping students’ scores into four, five and six clusters. In figure 1, the lowest scoring subject for cluster 1 is History. The score is 64.05%. Cluster 2 shows Additional Mathematics having the lowest score at 42.65%. The subject with minimum score for cluster 3 is Chemistry while the subject with minimum score for cluster 4 is Additional Mathematics. The scores are 50.33% and 28.82% respectively. The highest scoring subject for cluster 1, 2 and 3 is Physics. The scores are 87.45%, 79.23% and 77.73% respectively. The highest scoring subject for cluster 4 is English at 80.59%.

In figure 2, the highest scoring subject for cluster 1 and 2 is Physics. The scores are 87.06% and 79.26% respectively. English is the highest scoring subject for cluster 3, 4 and 5. The scores are 81.47%, 79.44% and 79.58% respectively. For cluster 1 and 2, the lowest scoring subject is History. The scores are 66.76% and 46.69% respectively. Additional Mathematics is recorded as the lowest scoring subject in cluster 3, 4 and 5. The scores are 46.41%, 28.85%, and 42.21% respectively.

In figure 3, the highest scoring subject for cluster 1, 3 and 6 is Physics. The scores are 88.43%, 86.30% and 77.3% respectively. For cluster 2, 4 and 5, English is the highest scoring subject. The scores are 80.12%, 80.12% and 81.08% respectively. The lowest scoring subject for cluster 1 is History. The score is 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 with 49.97% is the lowest scoring subject.

Figure 1. Students’ overall performance for each subject when grouping students’ scores into four clusters
Figure 2. Students’ overall performance for each subject when grouping students’ scores into five clusters.

Figure 3. Students’ overall performance for each subject when grouping students’ scores into six clusters.
5 Conclusion

K-means clustering algorithm and deterministic model are used to evaluate the academic performance of a college’s 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 for further investigations on effect of students’ conditions within clusters to improve students’ academic performance in next academic year.

References


