Model Development in Predicting Academic Performance of Students Based on Self-Efficacy Using K-Means Clustering

Mohammad Zahid Mohammad Sabri, Nazatul Aini Abd Majid and Siti Aishah Hanawi
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia.

Email: nazatulaini@ukm.edu.my

Abstract. Many prediction models have been developed using data mining tools in order to predict the performance of the students at the early stage. The academic performance of higher education students commonly was predicted based on their results in the end of the previous semester or during the semester like test score or mid-term exam. However, there is lack of models that emphasize the use of data related to student’s behaviour for predicting the academic performance. Therefore, the aim of this study is to investigate the use of self-efficacy behaviour data to predict the academic performance of students using principal component analysis (PCA) and k-means clustering (KMC). This study focuses on the first part of the prediction which is model development. In the model development phase, the number of variables were reduced from 20 into two by using PCA. The scores for the first two principal components were clustered by using KMC. The results show that the scores can be clustered into three main clusters to differentiate the student’s self-efficacy behaviour. Next research will investigate the underlying causes of the clusters in order to predict the risky students.

1. Introduction
One of the essential tasks in learning process is monitoring and evaluation of students’ performance. For many reasons, identifying students that are risky to perform poorly or fail has been very important to institutions, lecturers, tutors and support staff [1]. Therefore, the uses of predictive modelling can benefit the teachers to identify students at risk in ongoing courses [2]. Therefore, it is crucial to predict the students’ performance accurately because it will be very useful in finding weak students [3] and [4]. Identifying students who are likely to fail is very important in order to prevent it from happening. This information may then give lecturers and tutors an opportunity to make timely supportive interventions designed to increase the student’s likelihood of success. Knowledge transfer also will be optimized when teachers or lecturers can customize their teaching methods based on students’ type or behaviours. Students’ personality also has been found as a factor to differentiate between them [5]. Since students’ academic performance is a major issue that is often a question for academic teachers whether at the primary, secondary or tertiary level, data mining has the potential to predict student’s academic performance not only using previous results but also other scopes such as self-efficacy and study habits.

Developments in the field of data mining allow researchers to mine data in the field of education to improve the quality and process of education itself [6]. Data mining is one of the well-known methods and has been widely used in education [7]. In data mining, cluster analysis is a common statistical data analysis used to gather a set of similar attributes into one cluster and other similar attributes into another
cluster [8]. Several researchers had been using clustering technique for analysis and modelling in various fields such as automotive [9], healthcare [10], transportation [11], text [12], marketplace stocks [13] and security [14]. K-means clustering (KMC) is one of the data mining tools that can be used to predict output by using the attributes in the data set. KMC has been used widely because it can be applied to the huge database [9]. [5] stated that students can be grouped together using KMC and the traits of each group can be investigated further. Basically, the clustering in KMC is performed by gathering nearby objects into numbers of centroids with the coordinates of each centroid [15]. KMC was selected as a data mining method in this study because it was popular with its simplicity of implementation, ease to use, scalability, convergence speed and can be adapted to sparse data [16]. KMC also can help to boost the capability of PCA to partition the data into a number of clusters precisely. However, lack of clustering models that emphasize the use of data related to student’s behaviours for predicting the academic performance in higher education. Therefore, the aim of this research is to investigate the use of self-efficacy behaviours data to predict the academic performance of students using PCA and KMC.

2. Methodology

One of the important phases in predicting students’ performance was model development. In this phase, data were examined and selected to build the prediction model. This is to determine the students’ performance based on their self-efficacy behaviours. In this paper, selected steps in a typical procedure for knowledge discovery and databases were applied in the model development phase. These steps were adapted from [17] and are shown in Figure 1.

![Figure 1: Overall steps of the model development phase for predicting students’ performance](image)

Steps involved in the model development phase were as follows:

Step 1: In data collection, data were collected by using questionnaire that includes questions regarding students’ background and self-efficacy items adapted from [18]. The data set contains 153 respondents which were the first-year university students with information technology background.
Step 2: In data pre-processing step, data were pre-processed by imputation, data integration and data standardization. This step is to eliminate abnormal data and clean the data from outliers.

Step 3: In data reduction, PCA was applied as this technique reduces the variation dimension of the self-efficacy data set [19]. Data were reduced and projected to two principal component scores. The output from this step is principal component loading vectors.

Step 4: In the clustering step, the scores from the first two principal components were clustered using k-means clustering. This algorithm aims at minimizing an objective function known as squared error function given by Equation 1:

\[ J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left\| x_i - v_j \right\|^2 \]

where,

\[ \left\| x_i - v_j \right\| \] is the Euclidean distance between \( x_i \) and \( v_j \).

\( c_i \) is the number of data points in \( i^{th} \) cluster.

\( c \) is the number of centroids.

Let \( X = \{ X_1, X_2, X_3, ..., X_n \} \) be the set of data points and \( V = \{ V_1, V_2, ..., V_c \} \) be the set of centers.

1. Randomly select \( c \) centroids.
2. Calculate the distance between each data point and centroids.
3. Assign the data point to the centroid whose distance from the centroid is minimum of all the centroids.
4. Recalculate the new centroids using:

\[ v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i \]

where, \( c_i \) represents the number of data points in \( i^{th} \) cluster.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3

The output of the algorithm is a set of \( c \) centroids: \( V_1, V_2, ..., V_c \). Euclidean distance was used as a distance measure between sample and nearest centroid because it was good and commonly used when applying k-means clustering.

Step 5: In interpreting patterns step, the patterns of students’ performance were then recognized by investigating the related grade for each cluster.

The steps in the model development phase as shown in Figure 1 were applied to the student’s self-efficacy dataset for 153 samples with 20 variables related to self-efficacy behavior. The results of applying the five steps will be discussed in the section as follow.

3. Result and Discussion

First year students’ self-efficacy data were collected as in Step 1 and pre-processed. After the data were cleaned as in Step 2, principal components of the data were extracted using PCA as for Step 3. Based on the scree plot as shown in Figure 2, the first two principal components were selected for the next step. Based on the principal component loading vectors, the scores for principal components are shown in Figure 3. KMC was applied as in Step 4 to divide the scores to the appropriate clusters. Figure 4 shows the process of calculating the centroids using k-means algorithm. It shows the calculation of centroid stopped after seven iterations.

In Figure 5, the scores were clustered into three clusters based on the calculated three centroids. In this model development, it can be concluded that if new data is introduced to the model, the cluster of the new data can be determined by using principal loading vectors from step 3 and centroids from step 4. As the clusters have been plotted, student’s preference can be identified as in Step 5. The observations in the clusters were compared with the students’ grade. From there, the clusters can be defined by the similarity of characteristics in each cluster. Cluster 2 was basically for students with good grades that
consist of 65 students with grade B and above. Meanwhile, Cluster 1 consists 52 students with grade B and above although there was an E grade student in this cluster. On the other hand, Cluster 3 only shows 23 students with grade B and above but the number of excellent students in Cluster 1 is higher than Cluster 3.

From the results of the clusters characteristics, the clusters were not really accurate in differentiating the students’ self-efficacy for their academic performance. The accuracy of the cluster might be better by using bigger dimension of variables such as collecting self-efficacy data throughout the term or semester instead of only one-time survey. This type of data also can make use of Multiway PCA during model development. Additional type of variables such as holistic dimension also can give a different detection. Combination of self-efficacy variables and students learning behaviors in lectures or tutorials might also give different and interesting result in the future. The development of more complicated and advanced model can help in increasing the accuracy. Furthermore, besides predicting the student’s outcome, it is also important to know the causes that affect their performance [20].

4. Conclusions
The use of self-efficacy behaviour data to predict the academic performance of students using PCA and KMC was investigated in this study. By using PCA, the data set have been reduced from 20 variables for self-efficacy items into two principal component variables. As this study focused on the first part of the prediction which is model development, the principal component loading vectors and the three centroids for the clusters have been identified.

Further investigation is needed to identify the underlying patterns for each cluster and to predict future data sets by transforming the new data using the principal component loading vectors and then identifying the cluster using the centroids. This study shows that a prediction model for academic performance may not depend on one perspective alone like self-efficacy. More research should be done to look at the new dimension of data in predicting the academic performance of students holistically.
References

[1] Wakelam E, Jefferies A, Davey N and Sun Y 2019 Br. J. Educ. Technol. 51(2) pp 347 – 70
[2] MonllaóOlívé D, Huynh D Q, Reynolds M, Dougiamas M and Wiese D 2020 J. Comput. High. Educ. 32(1) pp 9–26
[3] Azmi M S B M and Paris I H B M 2011 Int. Conf. Commun. Softw. Netw. pp 24–27
[4] Tomasevic N, Gvozdenovic N and Vranes S 2020 Comput. Educ. 143 pp 103676
[5] Mohd Ariff N, Abu Bakar M A and Zamzuri Z H 2020 J. Fundam Appl Sci 16(3) 328–33
[6] Asif R, Merceron A, Ali S A and Haider N G 2017 Comput. Educ. 113 pp 177–94
[7] Romero C and Ventura S 2010 Transactions on Systems, Man, and Cybernetics, Part C; Applications and Reviews 40(6) pp 601–18
[8] Swathi B V and Govardhan A 2009 Int. J. Comput. Sci. Commun. Technol. 2(1) pp 268 – 72
[9] Yunoh M F M, Abdullah S, Saad M H M, Nopiah Z M and Nuawi M Z 2017 J. Braz. Soc. Mech. Sci. Eng. 39 (3) pp 757–64
[10] Mahmoudi M R, Baleanu D, Mansor Z, Tuan B A and Pho K H 2020 Chaos, Solitons and Fractals 140 pp 1–9
[11] Shima Y, Kadir R A, Fathelalem A and Sulaiman R 2020 J. Theor. Appl. Inf. Technol. 10 (19) pp 3211–21
[12] Ahmed M A, Baharin H and Nohuddin P N E 2020 Int. J. Adv. Comput. Sci. Appl. 11 (8) pp 248–54.
[13] Nazifah S, Abidin Z, Jaaman S H and Ismail M 2020 Symmetry 12 (7) pp1148
[14] Nur N, Nik A, Norul S, Sheikh H, Zaharah S and Rahman A 2019 Test Eng. Manag. 81 pp 4789–99
[15] Park H S and Jun C H 2009 Expert Syst. Appl. 36 (2) pp 3336–41
[16] Oyelade O J, Oladipupo O O and Obagbuwa I C 2010 Int. J. Inf. Comput. Secur. 7 pp 292–95
[17] Aini Abd Majid N, Young B R, Taylor M P and Chen J J J 2012 Conference on Data Mining and Optimization pp 43-46
[18] Sachitra V and Bandara U 2017 International Journal of Educational and Pedagogical Sciences 11 pp 2363–68
[19] Xue L S, Majid N A A and Sundararajan E A 2020 TEM Journal 9 (1) pp 93–100
[20] Sokkhey P and Okazaki T 2019 Int. Tech. Conf. Circuits/Syst. Comput. Commun. pp 1–4.

Acknowledgments
This research is supported by the GUP-2020-090 grant.