Tracking At-risk Student Groups from Teaching and Learning Activities in Engineering Education

Christopher Chung Lim KWAN[1][0000-0002-4085-9210]

[1] The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR

ceclkwan@polyu.edu.hk

Abstract. Tracking student groups, in particular, at-risk student group is a challenging but meaningful work in a large class of an engineering mathematics course, enabling instructors to ascertain how well students are learning and when they need interventions of their studies during the delivery of teaching and learning activities. In the paper, two unsupervised learning algorithms, hierarchical clustering and k-means clustering, are used and compared with the use of LMS data such as the level of achievements in online class activities, assignments, a mini-project and a mid-term test for tracking at-risk student groups at the end of weeks 3, 5, 7, 9 and 11 in a 13-week semester of an academic year. Notwithstanding the higher accuracy of both clustering, the k-means clustering significantly outperforms the hierarchical clustering in terms of the precision, recall and f-measure at the end of week 11. It is found that the k-means clustering can be employed to track at-risk students with the recall of 0.640 and the f-measure of 0.533 for the initial intervention of their studies by the end of week 7.

Keywords: At-risk student, Hierarchical clustering, K-means clustering, Precision, Recall, F-measure.

1 Introduction

Traditionally, educational data, generally generated from results of many assessment tasks like assignments, tests, laboratory reports and examinations, are used to grade student performance at the end of a subject or a course, informing students of how well they have learned for the progression of studies and graduation. These assessment results are further analyzed by course instructors to measure the achievement of the subject intended learning outcomes for quality assurance and accreditation purposes [2, 3, 11]. On the other hand, assessment can be regarded as formative feedback to students, providing them with frequent responses and precise information on how well they are on track during learning, and timely interventions of their studies if at-risk student group can be identified and tracked as early as possible during the delivery of teaching and learning activities [6, 7]. This is a particularly challenging work for lecturing in large classes [9].

With advances in artificial intelligence, it is possible to identify at-risk students in class and to predict students' success in a course [4, 5, 7, 10]. Marbouti et al. [8] built three logistic regression-based models to identify at-risk students in a large first-year

engineering course at weeks 2, 4 and 9 in a semester. These models are highly predictive in identifying at-risk students. However, these models like other supervised learning models cannot be trained and tested in the absence of observed data or output variable such as students' final grade, addressing the value of creating unsupervised learning models like hierarchical clustering and k-means clustering for tracking and identifying at-risk student groups.

2 The Context of the Study

The dataset of an engineering mathematics course offered in a 13-week semester of an academic year is used for the present study and extracted from Blackboard LMS for hierarchical clustering and k-means clustering. In total, there are a total of 240 students participating in class activities and various assessment tasks, designed on the basis of the subject curriculum and the subject intended learning outcomes.

Identifying at-risk students with the aid of artificial intelligence is the focus of the study. The present study thus aims at addressing the following research questions:

1. What is the performance of hierarchical clustering and k-means clustering for tracking at-risk student groups in terms of the accuracy, precision, recall and f-measure?
2. Which clustering can be employed to track at-risk students for timely intervention of their studies by the end of week 7 with certain degrees of recall and f-measure?

For the dataset, there are 16 input variables such as 2 assignments, a mini project, a mid-term test, and 12 online class activities held in each week of the semester. The online class activities are done in face-to-face (F2F) sessions for recording the number of multiple-choice questions correctly attempted as well as students' attendance. The score of the online class activities is not counted in the calculation of the coursework assessment as these activities are designed for enhancing student engagement in class and checking their understanding of the topics, concepts, and theorems. The input variables used for hierarchical clustering and k-means clustering are summarized in Table 1.

The output variable is the final examination score which is always an unknown variable before the end of the 13-week course and is intended not to be used for clustering. As the final examination score is made available at the end of the semester, this variable is simply used for evaluating the performance of hierarchical clustering and k-means clustering respectively at the end of weeks 3, 5, 6, 7, 9 and 11 in terms of the accuracy, precision, recall and f-measure. A binary variable (i.e. 0 or 1) which indicates whether the student is at-risk or not is also defined. An integer "1" can be assigned to the binary variable which represents an at-risk student who either fails in the final examination or is absent from the final examination. Conversely, an integer "0" is assigned to a not-at-risk student passing the final examination.

3 Methodology

Initially, three input variables such as 1st, 2nd and 3rd online class activities are used for hierarchical clustering and k-means clustering respectively at the end of week 3. At the end of week 5, 1st–5th online class activities and assignment 1 are selected as input variables for clustering. Because of an in-class mid-term test held in week 7, clustering is also carried out for finding different groups of similar characteristics like at-risk student groups by the end of week 7. In this connection, nine input variables such as 1st-7th online class activities, assignment 1 and mid-term test are selected. Furthermore, twelve input variables such as 1st-9th online class activities, assignment 1, mid-term test and mini-project are chosen for clustering by the end of week 9. At the end of week 11, fifteen input variables except the 12th online class activity are used for clustering as shown in Table 1.

Table 1. Input variables used for hierarchical clustering and k-means clustering.

Input Variable	Completed by week	Type	Point
Assignment 1	5	Numeric	0 - 15
Mid-term test	7	Numeric	0 - 50
Mini-project	8	Numeric	0 - 20
Assignment 2	11	Numeric	0 - 15
1st Online class activity	1	Integer	0 - 3
2nd Online class activity	2	Integer	0 - 8
3rd Online class activity	3	Integer	0 - 4
4th Online class activity	4	Integer	0 - 6
5th Online class activity	5	Integer	0 - 2
6th Online class activity	6	Integer	0 - 3
7th Online class activity	7	Integer	0 - 6
8th Online class activity	8	Integer	0 - 2
9th Online class activity	9	Integer	0 - 3
10th Online class activity	10	Integer	0 - 3
11th Online class activity	11	Integer	0 - 1
12th Online class activity	12	Integer	0 - 1

The goal of clustering is to categorize the data into similar groups. The distance between two data points are generally defined by “Euclidean distance”, where k is the number of independent variables.

$$d_{ij} = \sqrt{(x_{i1} - x_{j1)}{}^2 + (x_{i2} - x_{j2)}{}^2 + \cdots + (x_{ik} - x_{jk)}{}^2} \quad (1)$$

As distance is highly influenced by scale of variables, it is customary to normalize the data first. Both hierarchical clustering and k-means clustering are then used and

compared with the use of LMS data such as the level of achievements in online class activities, assignments, a mini-project and a mid-term test for tracking at-risk student groups at the end of weeks 3, 5, 7, 9 and 11 in a 13-week semester of an academic year.

3.1 Hierarchical Clustering

This hierarchical clustering is a bottom-up approach to construct a cluster dendrogram. The algorithm of hierarchical clustering is addressed as follows:

1. Assign a cluster to each data point initially such that 'n' clusters for 'n' data points;
2. Combine two nearest clusters by calculating the distance and the centroid;
3. Repeat to proceed the step 2 until all data points are in one cluster, then stop the iteration.

3.2 K-means Clustering

This method is also one of the simplest unsupervised learning algorithms [1]. The algorithm of k-means clustering is used for categorizing groups of similar characteristics. Firstly, the number of 'k' cluster centers is specified and initialized randomly. Then, the distances between each data point and cluster centers are calculated by using Euclidean distance formula. Secondly, assignment of the data points to that cluster center whose distance from the cluster center is minimum as compared to all the cluster centers is made. In other words, the minimum-distance classifier can be used to separate the above data into k clusters, where a data x_t is in cluster i if $\|x_t - m_i\|$ is the minimum of all k distances. That is,

$$b_i^t = \begin{cases} 1 & if\ \|x_t - m_i\| = \min_k \|x_t - m_k\| \\ 0 & otherwise \end{cases} \tag{2}$$

The algorithm aims to minimize an objective function which is defined as

$$E\left(\{m_i\}_{i=1}^k | X\right) = \sum_t \sum_i b_i^t \|x_t - m_i\|^2 \tag{3}$$

Thus, taking its derivative with respect to m_i and setting it to zero yield

$$m_i = \frac{\sum_t b_i^t x_t}{\sum_t b_i^t} \tag{4}$$

The new cluster center can thus be updated by using the assigned data points and the equation (4). Thirdly, the distances between each data point and new cluster centers are recalculated by using the equation (2). Therefore, this is an iterative procedure. If there is no reassignment of the data points, then the iteration is stopped. Otherwise, the second step is repeated for assigning the data points.

4 Result

The mean scores of five student clusters were determined from hierarchical clustering and k-means clustering respectively at the end of weeks 3, 5, 7, 9 and 11 respectively. In particular, it is found that Cluster 3 of hierarchical clustering is tracked and identified to be the potential at-risk student group based on the mean scores of input variables up to the end of week 7 as shown in Table 2. The mean scores of 1st, 2nd, 4th, 5th, and 7th online class activities are not shown in Table 2 for simplicity. The number of students in this cluster is 26. Students were not actively engaged in the online class activities as a result of the second lowest mean score among five groups. Their performances on both Assignment 1 and the mid-term test were also unsatisfactory as their mean scores were the lowest among the clusters. In particular, the mean scores of Assignment 1 and the mid-term test were 9.94 out of 15 and 19.06 out of 50 respectively. They thus scored on average 29.00 out of 65 for the completed coursework comprising Assignment 1 and the mid-term test. It is also found that the final examination score which is the output variable not to be used for hierarchical clustering was also the lowest among five groups. As identified to be the at-risk student group, 46.2% of students (i.e. 12 students) in this group can be correctly identified as at-risk students (i.e. true positive), representing a precision of 0.462 of the present clustering. However, 53.8% of students (i.e. 14 students) who are not-at-risk students can be misclassified (i.e. false positive).

Table 2. Hierarchical clustering of student groups at the end of week 7 in a 13-week course.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Group size	55	41	26	96	22
3th Online class activity	1.95	0.00	0.77	0.73	0.64
Assignment 1	13.21	13.45	9.94	12.63	13.16
6th Online class activity	2.38	0.00	0.04	0.97	1.09
Mid-term test	34.09	35.18	19.06	28.91	33.52
Coursework's score (wk.7)	47.30	48.63	29.00	41.53	46.68
Final examination	53.24	44.73	35.65	41.44	49.64
At-risk student %	3.6	17.1	46.2	27.1	13.6

Among these five clusters, students of Cluster 2 did not participate in any online class activity at all but they achieved the best performance on both Assignment 1 and the mid-term test. They obtained the highest mean score of the completed coursework up to week 7 but they only achieved the third highest mean score in the final examination. As Cluster 2 is identified to be the not-at-risk student group, 82.9% of students (i.e. 34 students) belonging to this cluster can be correctly classified as not-at-risk students (i.e. true negative). However, 17.1% of students (i.e. 7 students) who are really at-risk students can be misclassified (i.e. false negative).

Students belonging to Cluster 1 not only actively participated in online class activities, but also performed well on both Assignment 1 and the mid-term test. The final

examination score was the highest among other groups. As Cluster1 is not to be identified as the at-risk student group, 3.6% of students (i.e. 2 students) assigned to this group cannot be correctly tracked and classified as at-risk students for early intervention (i.e. false negative) but 96.4% of students (i.e. 53 students) can be correctly identified as not-at-risk students in this group (i.e. true negative).

Clusters 4 and 5 of hierarchical clustering are not identified to be groups of at-risk students because students of Clusters 4 and 5 ranked the second lowest mean score and the third highest mean score of the completed coursework up to the end of week 7 respectively. They were engaged in the online class activities as well. Overall, 72.9% and 86.4% of students belonging to Clusters 4 and 5 respectively (i.e. 70 and 19 students) can be correctly classified as not-at-risk students (i.e. true negative). However, 27.1% and 13.6% of students assigned to clusters 4 and 5 (i.e. 26 and 3 students) can be misclassified respectively (i.e. false negative).

Clusters 1 and 3 of k-means clustering are tracked to be the potential at-risk student group based on the mean scores of input variables up to the end of week 7 as shown in Table 3. The mean scores of 1st, 2nd, 4th, 5th and 7th online class activities are not shown in Table 3 for simplicity. Students belonging to Clusters 1 and 3 were not much engaged in the online activities among other clusters. They obtained the second lowest and the lowest mean score of the completed coursework up to week 7 respectively. Even though the final examination score which is the output variable is not used for k-means clustering as well, they ranked the lowest and the second lowest mean score in the final examination respectively. As tracked to be at-risk student groups, 48.3% and 33.3% of students (i.e. 28 and 4 students) in these two groups can be correctly classified as at-risk students (i.e. true positive), representing an overall precision of 0.457 of the present clustering. However, 51.7% and 66.7% of students (i.e. 30 and 8 students) who are not-at-risk students can be misclassified respectively (i.e. false positive).

Table 3. K-means clustering of student groups at the end of week 7 in a 13-week course.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Group size	58	57	12	44	69
3th Online class activity	0.47	1.12	0.33	1.84	0.25
Assignment 1	12.63	13.03	6.00	13.00	13.32
6th Online class activity	0.69	2.16	0.33	1.64	0.14
Mid-term test	22.03	31.18	20.29	35.17	35.94
Coursework's score (wk.7)	34.66	44.20	26.29	48.17	49.26
Final examination	33.22	49.02	37.83	50.45	48.75
At-risk student %	48.3	8.8	33.3	6.8	14.5

Clusters 2, 4 and 5 of k-means clustering are identified to be not-at-risk student groups because students of Clusters 2 and 4 actively participated in online class activities and did the coursework well. Clusters 2 and 4 ranked the third highest and the second highest mean score of the completed coursework by the end of week 7 respectively. Students of Cluster 5 showed the least participation in the online class activities

but achieved the highest mean score of the completed coursework. As a result, 91.2%, 93.2% and 85.5% of students belonging to Clusters 2, 4 and 5 respectively (i.e. 52, 41 and 59 students) can be correctly identified as not-at-risk students (i.e. true negative). Conversely, 8.8%, 6.8% and 14.5% of students assigned to these three clusters (i.e. 5, 3 and 10 students) can be misclassified respectively (i.e. false negative).

Clusters 4 and 5 of hierarchical clustering are identified to be the potential at-risk student groups based on the mean scores of input variables by the end of week 11 as shown in Table 4. The mean scores of 1st, 2nd, 4th, 5th, 7th – 11th online class activities are not shown in Table 4 for simplicity. Students assigned to Clusters 4 and 5 did not actively participate in the online class activities. They also obtained the second lowest and the lowest mean score of the completed coursework comprising two assignments, the mid-term test and the mini-project up to week 11 respectively. In fact, students of Cluster 4 did not submit Assignment 2 in week 11; some of them withdrew from their studies due to difficulties in handling tremendous workloads from studying 7 courses in a semester. Students of Clusters 4 and 5 ranked the lowest and the second lowest mean score in the final examination respectively. As detected to be at-risk student groups, 90.5% and 45.5% of students (i.e. 19 and 5 students) in these two groups can be correctly identified as at-risk students (i.e. true positive), representing an overall precision of 0.765 of the present clustering. However, 9.5% and 54.5% of students (i.e. 2 and 6 students) who are not-at-risk students can be misclassified respectively (i.e. false positive).

Table 4. Hierarchical clustering of student groups at the end of week 11 in a 13-week course.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Group size	79	51	78	21	11
3th Online class activity	1.57	0.00	0.77	0.33	0.18
Assignment 1	12.96	13.26	12.63	13.48	6.36
6th Online class activity	2.06	0.02	0.87	0.81	0.00
Mid-term test	33.19	30.84	29.97	28.52	17.64
Mini-project	17.59	15.98	17.05	14.57	5.27
Assignment 2	12.96	12.63	13.04	0.00	8.41
Coursework's score (wk.11)	76.70	72.73	72.69	56.57	37.68
Final examination	53.28	47.45	46.09	7.86	33.64
At-risk student %	2.5	19.6	17.9	90.5	45.5

Clusters 1, 2 and 3 of hierarchical clustering are tracked to be not-at-risk student groups because students of Clusters 1 and 3 were actively engaged in online class activities and did the coursework well. Clusters 1 and 3 ranked the highest and the third highest mean score of the completed coursework by the end of week 11 respectively. Students of Cluster 2 had the least participation in the online class activities but achieved the second highest mean score of the completed coursework. As a result, 97.5%, 80.4% and 82.1% of students belonging to Clusters 1, 2 and 3 respectively (i.e.

77, 41 and 64 students) can be correctly classified as not-at-risk students (i.e. true negative). Nevertheless, 2.5%, 19.6% and 17.9% of students assigned to these three clusters (i.e. 2, 10 and 14 students) can still be misclassified respectively (i.e. false negative).

Clusters 2 and 3 of k-means clustering are tracked to be the potential at-risk student group based on the mean scores of input variables up to the end of week 11 as shown in Table 5. Students belonging to Clusters 2 and 3 were not much engaged in the online activities. They obtained the lowest and the second lowest mean score of the completed coursework up to week 11 respectively. They ranked the second lowest and the lowest mean score in the final examination respectively, despite the fact that the final examination was not included in the clustering. As tracked to be at-risk student groups, 50% and 91% of students (i.e. 6 and 20 students) in these two groups can be correctly classified as at-risk students (i.e. true positive), corresponding to an overall precision of 0.765 of the present clustering. Conversely, 50% and 9% of students (i.e. 6 and 2 students) who are not-at-risk students can be misclassified respectively (i.e. false positive).

Table 5. K-means clustering of student groups at the end of week 11 in a 13-week course.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Group size	71	12	22	77	58
3th Online class activity	0.77	0.17	0.32	1.65	0.03
Assignment 1	12.57	6.83	13.45	13.01	13.21
6th Online class activity	0.97	0.00	0.77	2.10	0.02
Mid-term test	29.63	17.29	28.07	33.05	31.93
Mini-project	17.11	5.50	14.77	17.55	16.24
Assignment 2	13.00	8.33	0.27	12.95	12.97
Coursework's score (wk.11)	72.32	37.96	56.57	76.55	74.34
Final examination	46.28	33.33	8.86	52.69	48.64
At-risk student %	16.9	50.0	91.0	2.6	17.3

Clusters 1, 4 and 5 of k-means clustering are classified to be not-at-risk student groups because students of Clusters 1 and 4 were actively engaged in online class activities and did the coursework well. Clusters 1 and 4 ranked the third highest and the highest mean score of the completed coursework by the end of week 11 respectively. Students of Cluster 5 showed the least participation in the online class activities but achieved the second highest mean score of the completed coursework. Overall, 83.1%, 97.4% and 82.7% of students belonging to Clusters 1, 4 and 5 respectively (i.e. 59, 75 and 48 students) can be correctly classified as not-at-risk students (i.e. true negative). However, 16.9%, 2.6% and 17.3% of students belonging to these three clusters (i.e. 12, 2 and 10 students) can be misclassified respectively (i.e. false negative).

Accuracy, precision, recall and f-measure of a model are defined and calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F - measure = 2 \cdot \frac{Prescision \cdot Recall}{Precision + Recall} \quad (8)$$

where TP: true positive; TN: true negative; FP: false positive; FN: false negative

The performance of the present models is further evaluated in terms of accuracy, precision, recall (i.e. sensitivity), and f-measure as shown in Table 6. Despite the high accuracy of both models, it is found that the k-means clustering has achieved the higher recall of 0.640 and the f-measure of 0.533 by the end of week 7. Furthermore, it has achieved the higher precision of 0.765, the recall of 0.520, and the f-measure of 0.619 by the end of week 11.

Table 6. Accuracy, precision, recall and f-measure of hierarchical clustering and k-means clustering

	Week 7		Week 11	
	Hierarchical Clustering	K-means Clustering	Hierarchical Clustering	K-means Clustering
Accuracy	0.783	0.767	0.858	0.867
Precision	0.462	0.457	0.750	0.765
Recall	0.240	0.640	0.480	0.520
F-measure	0.316	0.533	0.585	0.619

5 Conclusion and Future Works

It is concluded that the k-means clustering significantly outperforms the hierarchical clustering in terms of the precision, recall and f-measure at the end of week 11. It is found that the k-means clustering can be employed to track at-risk students with the recall of 0.64 and the f-measure of 0.533 for the initial intervention of their studies once the results of the 1st – 7th online class activities, assignment 1, and the mid-term test are made available at the end of week 7.

To further confirm that the differences between clusters of the five-cluster solution are distinctive and significant, F statistics from one-way ANOVAs will be calculated to examine whether there are statistically significant differences between the five clusters on each of the clustering variables such as assignments, mid-term test and online class activities, and each of two non-clustering variables such as coursework's score and final examination. The independent variable is cluster membership, and the dependent variables are the clustering variables and two non-clustering variables. The results will show that there are significant differences between clusters on most of these variables with the p-value being below 0.05. The significant F statistics provide an evidence that each of the five clusters is distinctive.

References

1. Alpaydin, E.: Introduction to Machine Learning. Second Edition. MIT Press. (2010).
2. Biggs, J.: Teaching for quality learning at university, 2nd Edition. Society for Research into Higher education & Open University Press (2003).
3. Hong Kong Institution of Engineers: Professional Accreditation Handbook (Engineering Degrees). Accreditation Board. 1-35 (2013).
4. Kwan, C.L.C.: Identifying At-risk Students from Course-specific Predictive Analytics. In: 27th International Conference on Computers in Education, pp. 356-360 (2019).
5. Lackey, L. W., Lackey, W. J., Grady, H. M., & Davis, M. T.: Efficacy of using a single, non-technical variable to predict the academic success of freshmen engineering students. Journal of Engineering Education, 92(1), 41-48 (2003).
6. Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S.J.H.: Applying learning analytics for the early prediction of students' academic performance in blended learning. Journal of Educational Technology & Society, 21(2), 220-232 (2018).
7. Macfadyen, L. P., & Dawson, S.: Mining LMS data to develop an “early warning system” for educators: A proof of concept. Computers & Education, 54(2), 588-599 (2010).
8. Marbouti, F., Diefes-Dux, H. A., & Strobel, J.: Building course-specific regression-based models to identify at-risk students. In: The american society for engineering educators annual conference. Seattle, WA (2015).
9. Mulryan-Kyne, C.: Teaching large classes at college and university level: Challenges and opportunities. Teaching in Higher Education, 15(2), 175-185 (2010).
10. Olani, A.: Predicting first year university students' academic success. Electronic Journal of Research in Educational Psychology, **7**(3), 1053-1072 (2009).
11. Sazhin, S.: Teaching mathematics to engineering students. International Journal of Engineering Education 14, 145-152 (1998).