The Latent of Student Learning Analytic with K-mean Clustering for Student Behaviour Classification

Andi Besse Firdausiah Mansur[1]*, Norazah Yusof [2]

[1][2]*Faculty of Computing and Information Technology Rabigh, King Abdulaziz University*
Jeddah,Kingdom of Saudi Arabia

[1]abmansur@kau.edu.sa

[2]nyusof@kau.edu.sa

Abstract

Since the booming of "big data" or "data analytic" topics, it has drawn attention toward several research areas such as: student behavior classification, video surveillance, automatic navigation and etc. This paper present k-mean clustering technique to monitor and assess the student performance and behavior as well as give improvement toward e-learning system in the future. Data set of student performance along with teacher attributes are collected then analyzed, it was filtered into 6 attributes of teacher that may potentially affect the student performance. Afterwards, k-mean clustering applied into the filtered data set to generate particular cluster number. The result reveal that Teacher1 statistically hold the highest density (0.27) and teachers with good speech/lectures tend to have strong correlation with another factor such as: commitment of teacher on preparing lecture material and time management utilization. If this synergy between teacher and student running flawlessly, it will be great achievement for e-learning system to the society.

I. INTRODUCTION

The student engagement in online discussion or forum plays important roles toward the high quality education system in the future. Some researchers have design the social learning analytic to monitor student discussion while doing webinar or online lecture [1]. They develop a framework that can convert the discussion in e-learning system into a kind of information that exposes student's behavior information. MOOCs (Massive Open Online Courses) enable students to distribute and pathway toward their awareness. Authors presenting big data analysis technique toward education collection to reveal how the communication, unspecified association, and unseen designs among between students that utilized MOOCs [2]. The analyzed information will be used to study pattern learning behavior of student that can bring a feedback toward the courses and teacher in the forthcoming.

Social Network Mining is a method which typically utilized to investigate the actions of community within group or population. Because it's adept to symbolize communal connection among populace and study the flow of information within network using flow betweens approach [3][4]. Social Network Analysis (SNA) is a technique to determine grade of member, principal player or to recognize the formation of communal relations [5]. Additionally, a number of researcher used avatar to enhance the user interaction when bump up student engagement during learning process [6]. Moreover, collaborative network such as wiki has a ton of features such as shared page editing, upload, give comment or tag other person to enrich societal communication among them. It also can be augmented to analyze people attitude within social learning network [7][8]. Generally, Social learning network analysis (SLNA) is studying e-learning with social network technique. The main idea is to monitor the behavior of student within collaborative system and take an innovative method to augment the teaching and education process [9][10][11]. Even though automatic text analysis has been studied, most of previous research mainly paying attention on the content analysis itself not the flow of information within community. The relation inside community through social

---

* Corresponding author

network such as Youtube video has been studied as well to do integration between social network and e-learning system [12][13].

The connection of education and schoolwork policy contribute to student success rate. On the other hand, character of each student also embraces the failure rate of student [14][15][16][17]. More researchers have investigated the student learning behavior and produce cluster according to the student performance in the classroom. It also explored the social website for shared group work in the social media [18][19] [20] [21] [22] [23].

II. METHODS

This paper reveals the behaviour of student by using k-mean clustering technique as well as hierarchical clustering. We have analysed the dataset from Gunduz, G. & Fokoue, E that consist of student performance evaluation during taking the courses [24]. The data consist of instructor, class, repeat (if student failed in the previous session). Attendance as well as difficulty level of courses also being monitored. These data are intended to collect the correlation value between student failure with instructor preparation and their way of teaching. There are 28 questions are distributed then collected after obtaining answer from student as described in Table I.

TABLE I
QUESTION MODEL FOR STUDENTS [24]

Number	Question
Q1	"The semester course content, teaching method and evaluation system were provided at the start"
Q2	"The course aims and objectives were clearly stated at the beginning of the period."
Q3	"The course was worth the amount of credit assigned to it."
Q4	"The course was taught according to the syllabus announced on the first day of class."
Q5	"The class discussions, homework assignments, applications and studies were satisfactory."
Q6	"The textbook and other courses resources were sufficient and up to date."
Q7	"The course allowed field work, applications, laboratory, discussion and other studies."
Q8	"The quizzes, assignments, projects and exams contributed to helping the learning."
Q9	"I greatly enjoyed the class and was eager to actively participate during the lectures."
Q10	"My initial expectations about the course were met at the end of the period or year."
Q11	"The course was relevant and beneficial to my professional development."
Q12	"The course helped me look at life and the world with a new perspective."
Q13	"The Instructor's knowledge was relevant and up to date."
Q14	"The Instructor came prepared for classes."
Q15	"The Instructor taught in accordance with the announced lesson plan."
Q16	"The Instructor was committed to the course and was understandable."
Q17	"The Instructor arrived on time for classes."
Q18	"The Instructor has a smooth and easy to follow delivery/speech.
Q19	"The Instructor made effective use of class hours."
Q20	"The Instructor explained the course and was eager to be helpful to students."
Q21	"The Instructor demonstrated a positive approach to students."
Q22	"The Instructor was open and respectful of the views of students about the course."
Q23	"The Instructor encouraged participation in thecourse."
Q24	"The Instructor gave relevant homework assignments/projects, and helped/guided students."
Q25	"The Instructor responded to questions about the course inside and outsideof thecourse."
Q26	"The Instructor'sevaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives."
Q27	"The Instructor provided solutions to exams and discussed them with students."
Q28	"The Instructor treated all students in a right and objective manner."

Each response toward the questions are recorded then will be used for extra investigation. The data mining approach that used for the research is k-mean clustering. It is part of data analytic method considered as unsupervised clustering. In k-mean, it started with K-centroids initialization for every cluster then assigned the centroid position into suitable location because different location may generate different result. Therefore the best choice for k-mean is to put the centroid far away from each other. Then, it continued by taking data set closed to the arranged centroids until there are no points left then the primary stage is finished and the premature cluster is completed. Then k as new centroid for midpoint in the generated cluster will be recalculated. Furthermore, a relationship has been established among identical data series points and the closest new centroid then a loop will be established. During the process we may observe that the k centroid will adjust their position accordingly until it cannot be changed (centroid position is fixed). The general steps of k-mean clustering are describe in the following procedure. Equation 1 shows the general form of k-mean algorithm, while complete algorithm is described in Figure 1.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Where $'\|x_i - v_j\|'$ is the Euclidian distance between x_i and v_j

'c_i' is the amount of data points in i^{th} cluster

'c' is the amount of cluster centres

```
                 K-means Algorithm
   1. Let X={x₁, x₂, x₃,,...., xₙ} be the number of
      data points
   2. V={v₁, v₂, v₃,,...., v_c} assigned as centres
      of clusters
   3. Choose 'c' as cluster centers randomly
   4. Measure the distance between each data
      points and the centre of cluster
   5. Allocate the position for data point
      closed to the centre that has minimum
      distance toward the cluster centre
   6. Recomputed the new cluster using:
```

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

```
      Where, 'cᵢ',represents the number of data
      points in the iᵗʰ cluster.

   7. Recompute the distance among every data
      point and new obtained cluster centres'
   8. If there are no data points that was
      reallocated then stop, or else go over
      step 3
```

Figure 1. K-mean clustering algorithm

For the current analysis we are focusing on question 15 to 20 that closely affect the lecture process in the classroom. Mainly, there are three instructors that being monitored toward their method on delivering lectures which is closely related to student performance, refer to Table II.

TABLE II
STUDENT EVALUATION DATASET

Instructor	Course repetition	Q15	Q16	Q17	Q18	Q19	Q20
Teacher1	1	3	3	3	3	3	3
Teacher1	1	3	3	3	3	3	3
Teacher1	1	5	5	5	5	5	5
Teacher1	1	3	3	3	3	3	3
Teacher1	1	1	1	1	1	1	1
Teacher1	2	4	4	4	4	4	4
Teacher1	1	4	4	4	4	4	4
Teacher1	1	5	5	5	5	5	5
Teacher1	1	4	4	4	4	4	4
Teacher1	1	4	4	4	4	4	4
Teacher1	1	3	3	3	3	3	3
Teacher1	1	3	3	3	3	3	3
Teacher1	1	4	4	4	4	4	4
Teacher1	1	1	1	1	1	1	1
Teacher1	1	1	1	2	2	2	3
Teacher1	2	4	4	5	5	5	4
Teacher1	1	1	1	1	1	1	1
Teacher1	1	4	4	4	4	4	4
Teacher1	1	1	1	1	1	1	1
....	..	...	...	...	...	...	...

TABLE III
K-MEAN CLUSTERING RESULT

Instructor	Cluster	Silhouette
Teacher1	Cluster1	0.592
Teacher1	Cluster1	0.696
Teacher1	Cluster1	0.512
Teacher1	Cluster1	0.605
Teacher1	Cluster1	0.657
Teacher1	Cluster1	0.696
Teacher2	Cluster1	0.577
Teacher2	Cluster1	0.696
Teacher3	Cluster1	0.696
Teacher3	Cluster1	0.517
Teacher2	Cluster2	0.673
Teacher2	Cluster2	0.649
Teacher2	Cluster2	0.649
Teacher2	Cluster2	0.673
Teacher2	Cluster2	0.647
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.69
Teacher3	Cluster2	0.69
Teacher3	Cluster2	0.69
Teacher3	Cluster2	0.562
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.514
Teacher3	Cluster2	0.649
Teacher3	Cluster2	0.589
Teacher1	Cluster3	0.61
Teacher1	Cluster3	0.709
Teacher1	Cluster3	0.61
Teacher2	Cluster3	0.709
Teacher2	Cluster3	0.597
Teacher2	Cluster3	0.709
Teacher2	Cluster3	0.709
Teacher3	Cluster3	0.709
Teacher3	Cluster3	0.709
Teacher3	Cluster3	0.61

III. RESULTS

The data set that presented in table II has been analyzed using k-mean algorithm which is limited into three clusters. Table III show the result of k-mean clustering with its silhouette. This table only show selected data due to the limited space, the whole three hundreds records of data are not able to be presented here.

As shown in Figure 2, there are three main clusters that can be recognized by the colours: red, green and blue. Each instructor they have its cluster. The data are grouped together closed to their centroids that dynamically changed until certain value is achieved.

The distribution graph of k-mean clustering as shown in Figure 3 illustrate that Teacher1 has the highest density followed by Teacher3 and Teacher2. This mean the student with number of repeat (failure) with Teacher1 is high. This value can be used to do further observation which involved Teachers and student personality to acquire good partnership among them to obtain great lecture and study experience.

IV. DISCUSSION

Furthermore, statistical analysis is generated, as we can observe from Table IV, Q18 has the greatest value of Gain ratio (0.0617), Gini(0.0612) and Anova(13.7543). Q18 related to Teacher should have good skill on delivering the course material to student, this mean; the personal factor of teacher on conducting smooth lecture during study has strong correlation with student performance during study. The second and third rankings are holding by Q16 which are teacher commitment and communication must be excellence then Q19: teacher should have good time management during lectures. The teacher commitment toward course and the efficient use of lectures time are contributed on supporting teacher on delivering good speech.
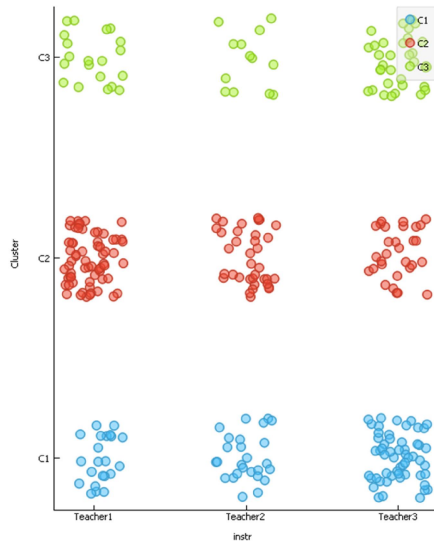
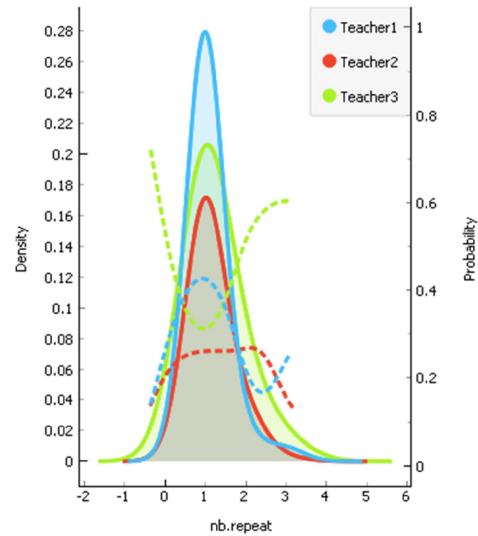Figure 2. The result of K-mean clustering of dataset: C1, C2 or C3.

Figure 3. Distribution graph of K-mean clustering

TABLE IV
DATASET RANGKIG AFTER K-MEAN CLUSTERING COMPUTATION

	Gain ratio	Gini	ANOVA
Q18	0.06173779751482743	0.061226330429944964	13.75435857744165
Q16	0.05391650441924183	0.05300712812883368	11.533221522643418
Q19	0.05173226767917495	0.05088367647677994	11.710503771885245
Q15	0.04641110665617076	0.04651872232710519	9.405336484380879
Q20	0.044923619986542845	0.0451223323560892	10.260378404147538
Q17	0.04322736172014927	0.042918241707087024	8.282915271677632
nb.repeat	0.0353527523925063	0.010967667483660182	2.915881629366609

Q15 has correlation with teaching plan, where teacher should follow the structure of the syllabus. This factor has contributed to cluster construction even though the ranked is 4th position. This is due to teaching plan is flexible and the arrangement is based on the educator knowledge. If the syllabus designed by other tutor most probably it will require an update or synchronization if continued by different teacher. Q20 related to tutor explanation toward the courses and their eagerness to help or assist student during the lectures time. This element has 5th ranked in the K-means dataset ranking cluster, the course description usually given in the beginning and during the learning process , instructor should keep track the student knowledge regarding the course and assist them whenever they are out of track. Q17 has the lowest rank in the Table IV, even though the learning time starting is important, the quality of learning or interaction during the study is more significant toward student learning success.

V. CONCLUSIONS

Student learning analytic has strong potential to reveal not only student behaviour but teacher personality as well. With the growth of fruitful data mining technique, user may classify certain data based on their characteristic. More data we have (big data) it will reveal more behaviour. Based on the conducted experiment teacher that capable on conducting smooth teaching with efficient communication will establish strong correlation with other factors such as: commitment of teacher on preparing lecture material and time management utilization. If this synergy between teacher and student running flawlessly, it will be great achievement for e-learning system to the mankind. The future works of this research are the diversity of data and size of data, the bigger dataset will reveal more pattern or behaviour. The big data with deep learning technique may be helpful and it also will become great steps on student analytic accomplishment.

REFERENCES

[1] B. Chen, Y.-H. Chang, F. Ouyang and W. Zhou, "Fostering student engagement in online discussion through social learning analytics," *The Internet and Higher Education 37 (2018) 21–30*, vol. 37, no. 2018, pp. 21-30, 2018.

[2] Y. Yang, H. Wu and J. Cao, "SmartLearn: Predicting Learning Performance and Discovering Smart Learning Strategies in Flipped Classroom," in *International Conference on Orange Technologies*, Melbourne, 2016.

[3] A. Mansur, N. Yusof and A. Basori, "Comprehensive analysis of Student's Academic Failure Classification through Role-Sphere Influence and Flow betwenness centrality," *Procedia Computer Science,* vol. 116, pp. 509-515, 2017. https://doi.org/10.1016/j.procs.2017.10.031.

[4] J. Ji, C. Zhou, Z. Wang and H. Yang, "Maximizing the Community Coverage of Influence through a Social Network," *AISS (Advances in Information Sciences and Service Sciences),* vol. 3, no. 9, pp. 339-346, 2011.

[5] B. Zhou and C. Wu, "Semantic Model for Social Networking Federation," *AISS (Advances in Information Sciences and Service Sciences),* vol. 3, no. 11, pp. 213-223, 2011.

[6] A. Basori, A. Tenriawaru and A. Mansur, "Intelligent Avatar on E-learning Using Facial Expression and Haptic," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 9, no. 1, pp. 115-124, 2011. http://dx.doi.org/10.12928/telkomnika.v9i1.676.

[7] F. Ortiza and R. Fraile, "Social Network Featuring Entertainment, Culture and Technology in Spanish Universities: The Infocampus Project," *Open Information Systems Journal,* vol. 3, no. 1, pp. 48-53, 2009.

[8] S. J. Kepp and H. Schorr, "Analyzing Collaborative Learning Activities in Wikis Using Social Network Analysis," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, Boston, MA, USA, 2009.

[9] H.-L. Twu, "A Predictive Study of Wiki Interaction: Can Attitude toward Wiki Predict Wiki Interaction in High-Context Cultures Groups?," *Journal of Educational Technology Development and Exchange (JETDE),* vol. 3, no. 1, 2010. 10.18785/jetde.0301.05.

[10] P. De Meo, F. Messina, D. Rosaci and G. M. Sarné, "Combining trust and skills evaluation to form e-Learning classes in online social networks," *Information Sciences,* vol. 405, pp. 107-122, September 2017. https://doi.org/10.1016/j.ins.2017.04.002.

[11] R. Hubsche, "Assigning Students to Groups Using General and Context-Specific Criteria," *IEEE Transactions on Learning Technologies,* vol. 3, no. 3, pp. 178-189, 2010. https://doi.org/10.1109/TLT.2010.17.

[12] A. Rad and M. Benyoucef, "Similarity and ties in social networks: a study of the youtube social network," *Journal of Information Systems Applied Research,* vol. 7, no. 4, 2014.

[13] H. Roreger and T. C. Schmidt, "Socialize online learning: Why we should integrate learning content management with Online Social Networks," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, Lugano, Switzerland, 2012. https://doi.org/10.1109/PerComW.2012.6197601.

[14] M. Erdt, A. Fernández and C. Rensing, "Evaluating Recommender Systems for Technology Enhanced Learning: A Quantitative Survey," *IEEE Transactions on Learning Technologies,* vol. 8, no. 4, pp. 326 - 344, 2015. https://doi.org/10.1109/TLT.2015.2438867.

[15] P. A. Grabowicz, L. M. Aiello, V. M. Eguiluz and A. Jaimes, "Distinguishing topical and social groups based on common identity and bond theory," in *Proceedings of the sixth ACM international conference on Web search and data mining*, Rome, Italy, 2013.

[16] C. M. Kokkinos, A. Kargiotidis and A. Markos, "The relationship between learning and study strategies and big five personality traits among junior university student teachers," *Learning and Individual Differences,* vol. 43, pp. 39-47, 2015. https://doi.org/10.1016/j.lindif.2015.08.031.

[17] M. Komarraju, S. J. Karau, R. R. Schmeck and A. Avdic, "The Big Five personality traits, learning styles, and academic achievement," *Personality and Individual Differences,* vol. 51, no. 4, pp. 472-477, 2011. https://doi.org/10.1016/j.paid.2011.04.019.

[18] M. Heller and G. Marchant, "Facilitating Self-Regulated Learning Skills and Achievement With a Strategic Content Learning Approach," *Community College Journal of Research and Practice,* vol. 39, no. 9, pp. 808-818, 2015. https://doi.org/10.1080/10668926.2014.908752.

[19] Y. Park, J. Yu and I. Jo, "Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute," *The Internet and Higher Education,* vol. 29, pp. 1-11, 2016.

[20] L. Halverson, C. Graham, K. Spring, J. Drysdale and C. Henrie, "Athematic analysis of the most highly cited scholarship in the first decade of blended learning research," *The Internet and Higher Education,* vol. 20, no. 1, 2014..

[21] W. Shiau, Y. Dwivedi and H. Yang, "Co-citation and cluster analyses of extant literature on social networks," *International Journal of Information Management,* vol. 37, no. 1, 2017.

[22] S. Chai and M. Kim, "A socio-technical approach to knowledge contribution behavior: An empirical investigation of social networking sites users," *International Journal of Information Management,* vol. 32, no. 1, 2012.

[23] S. K. Chu, C. M. Capio, J. C. van Aalst and E. W. Cheng, "Evaluating the use of a social media tool for collaborative group writing of secondary school students in Hong Kong," *Computers & Education,* vol. 110, pp. 170-180, 2017.

[24] E. Fokoue and N. Gunduz, "UCI Machine Learning Repository : Turkiye Student Evaluation Data Set," 2013.