Identifying Hidden Patterns in Students' Feedback through Cluster Analysis

Anwar Muhammad Abaidullah, Naseer Ahmed, and Edriss Ali

***Abstract*—The analysis and assessment of the students' feedback in improving the educational environment as well as enhancing students' learning experience is one of the critical issues for the higher education community. The conventional methods of analysis and assessment are not sufficient to explore the hidden information from the student feedback data repositories. In this paper we present the analysis of students' feedback data using k-means clustering algorithm for effective decision making by educational community responsible for monitoring and reviewing the effectiveness of educational programs and for improving the quality of teaching and learning experience for their students.**

***Index Terms*—Centroid, data mining, homogeneous groups, k-means.**

I. Introduction

The notion of feedback as a response to a sender's message originally comes from communication theorists whose early work laid the foundation for the understanding of feedback as an element of instruction [1]. The students' feedback is an indirect assessment measuring tool which is extensively being used as an evaluation of teaching in the field of higher education [2]. The students' feedback gathered in a structured way in an academic setting provides a genuine opportunity to individual student to express his/her opinion and raise issues for the consideration of academic and administrative officers of the higher educational institutions. This kind of feedback is not only beneficial for addressing students' concerns but also facilitates appropriate enhancement activities undertaken by the institution. A variety of formal and informal procedures based on qualitative and quantitative methods are commonly used with the aim of identifying a variety of issues concerning faculty, curriculum, teaching methodology and essential support services for resolving the identified issues and for enhancing the overall quality of academic programs and services provided by the institution.

The conventional methods used for this purpose consist of predefined queries and charts to analyze the data in the academic repositories. These conventional methods, however, are unable to explore some useful hidden information. Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets [3]. The amount of data stored in educational databases is increasing rapidly. Clustering technique is most widely used technique for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. The selection of data clustering tools and techniques mostly depends on the scope of the problem and the expected results from the analysis. Table I presents the summary of the clustering studies in the field of educational data mining.

TABLE I: Summary of the Clustering Studies in the Field of Educational Data Mining

Author(s) Name	Clustering Method	Choice Justified	Evaluation Measure	Triangulation Dataset
Durfee *et al*. [4]	SOM	No	-	Student's Dataset
Anaya and Boticario [5]	EM	No	-	Expert Opinion
Wang *et al*. [6]	ISODATA	Yes	-	-
Shih *et al*. [7]	Step-wise HMM	Yes	-	Students' Learning Outcome
Hubscher *et al*. [8]	Hierarchical clustering; K-means	Yes	-	-
Maull *et al*. [9]	K-means; EM	Yes	-	-
Lee [10]	PCA over SOM K-means	Yes	Within Cluster Variance	-
Dogan and Camurcu [11]	K-means	Yes	Within Cluster Variation	-
Perera *et al*. [12]	K-means	Yes	Same Results	Group Performance

In this paper we present k-means clustering algorithm as a simple and efficient tool to analyze and assess the students' feedback data for identifying good practices that could contribute towards the enhancement of educational environment for students.

The rest of the paper is organized as follows: In Section II, we present the clustering algorithm used and in Section III we present the data collection, results and analysis of the results. The conclusions of our work are given in Section IV.

II. Clustering Algorithm

The data clustering is an unsupervised and is a statistical data analysis technique to classify the same data into a homogeneous group and to operate on a large data-set to

Manuscript received February 23, 2014; revised April 22, 2014.

Anwar Muhammad Abaidullah and Edriss Ali are with the College of Engineering and Computing of Al Ghurair University, 37374 UAE (e-mail: {anwar, edirss}@agu.ac.ae).

Naseer Ahmed is with Al Ghurair University, 37374 UAE (e-mail: naseer@agu.ac.ae).

discover hidden pattern and relationship helps to make decision quickly and efficiently. There are two types of cluster analysis; hierarchical clustering and non-hierarchical clustering techniques. The single linkage, complete linkage, average linkage, median, and Ward are some of the types of hierarchical clustering techniques and non-hierarchical techniques include k-means, adaptive k-means, k-medoids, and fuzzy clustering. To determine which algorithm is good is a function of the type of data available and the particular purpose of analysis.

A. K-Means Algorithm

Originally known as Forgy's method [13], the K-means is one of the famous algorithms for data clustering and it has been used widely in several fields including data mining, statistical data analysis and other business applications. The K-means clustering algorithm builds clusters by RFM attributes (R: Recency, F: Frequency, M: Monetary).

The K-means algorithm suggested by [14] for describing an algorithm assigns each item to the cluster with the nearest centroid i.e. mean. The k-means clustering method produces exactly k different clusters of largest possible distinction and the best number of clusters k leading to the largest separation is not known as a priori and must be computed from the data.

B. Algorithmic Steps for K-Means Clustering

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3, ..., v_c\}$ be the set of centers.

1) Randomly select '*c*' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers.
4) Recalculate the new cluster center using the following formula.

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

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 *c*.

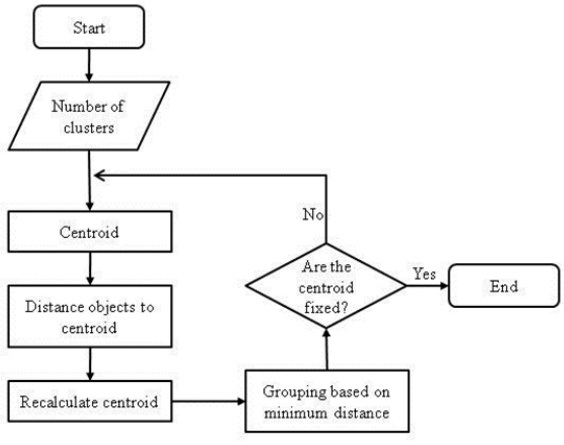

Fig. 1. Flow chart of k-means clustering.

The flow chart of the k-means algorithm is shown in Fig. 1.

III. Data Collection, Results and Analysis

This data used in this paper is available at UCI website [15] that contains a total 5820 evaluation scores provided by students from a University. There are thirty three (33) attributes/questions out of which twenty eight (28) are course specific questions. The data relating to Q-1 to Q-28 consists of Likert–type scale, i.e., the response values of these questions are of the form {1, 2, 3, 4, 5}. The non-course specific attributes are; instructor, course code having values from {1, 2, …, 13}, how many times student is taking this course, and attendance level values from {0, 1, 2, 3, 4} and level of difficulty of the course as perceived by the student with values taken from {1, 2, 3, 4, 5}. Twenty eight course specific questions used in this paper are as follows:

Q-1. The semester course content, teaching method and evaluation system were provided at the start.
Q-2. The course aims and objectives were clearly stated at the beginning of the period.
Q-3. The course was worth the amount of credit assigned to it.
Q-4. The course was taught according to the syllabus announced on the first day of class.
Q-5. The class discussions, homework assignments, applications and studies were satisfactory.
Q-6. The textbook and other courses resources were sufficient and up to date.
Q-7. The course allowed field work, applications, laboratory, discussion and other studies.
Q-8. The quizzes, assignments, projects and exams contributed to helping the learning.
Q-9. I greatly enjoyed the class and was eager to actively participate during the lectures.
Q-10. My initial expectations about the course were met at the end of the period or year.
Q-11. The course was relevant and beneficial to my professional development.
Q-12. The course helped me look at life and the world with a new perspective.
Q-13. The Instructor's knowledge was relevant and up to date.
Q-14. The Instructor came prepared for classes.
Q-15. The Instructor taught in accordance with the announced lesson plan.
Q-16. The Instructor was committed to the course and was understandable.
Q-17. The Instructor arrived on time for classes.
Q-18. The Instructor has a smooth and easy to follow delivery/speech.
Q-19. The Instructor made effective use of class hours.
Q-20. The Instructor explained the course and was eager to be helpful to students.
Q-21. The Instructor demonstrated a positive approach to students.
Q-22. The Instructor was open and respectful of the views of students about the course.
Q-23. The Instructor encouraged participation in the course.
Q-24. The Instructor gave relevant homework

assignments/projects, and helped/guided students.

Q-25. The Instructor responded to questions about the course inside and outside of the course.

Q-26. The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.

Q-27. The Instructor provided solutions to exams and discussed them with students.

Q-28. The Instructor treated all students in a right and objective manner.

A. Tools Used for Clustering

We used Weka [16] (Waikato Environment for Knowledge Analysis) software which is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. Due to the limitation of the space, we present cluster analysis details appearing in the Weka explorer window for only one question which is shown in Fig. 2.

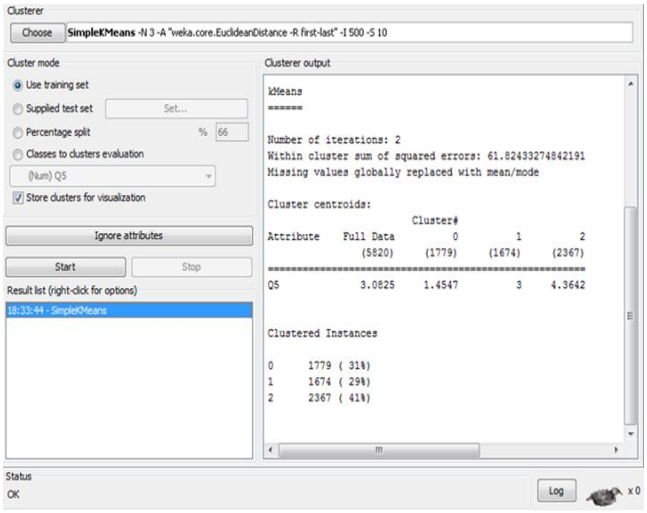

Fig. 2. Snapshot of Weka explorer for question 5.

B. Results and Discussion

The result of three (03) clusters generated using K-Means method with Euclidean distance for all questions is shown in Table I. Some of the clusters have overall percentage one less or more than hundred. This is due the sum of squared errors within cluster for example 61.82433274842191 in Question 5.

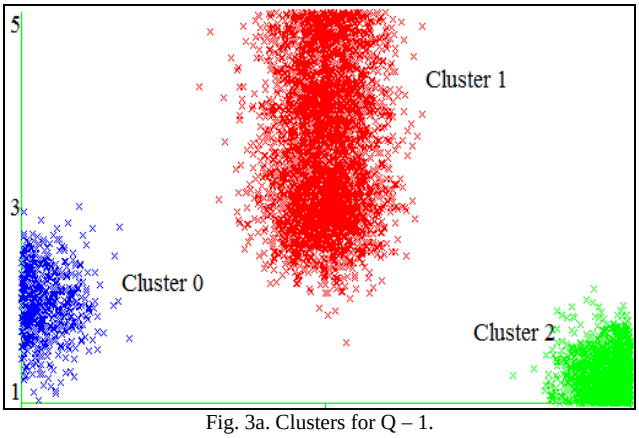

Fig. 3a. Clusters for Q – 1.

We present the cluster analysis of only two questions (Q-1 and Q-24) in visualization mode as observed in the Weka software.

The cluster analyses of these questions are shown in Fig. 3a and Fig. 3b. Fig. 3a presents the three clusters for Q-1 and Fig. 3b shows the pictorial view of three clusters for Q-24. The points outside the boundaries around the clusters are considered as outliers.

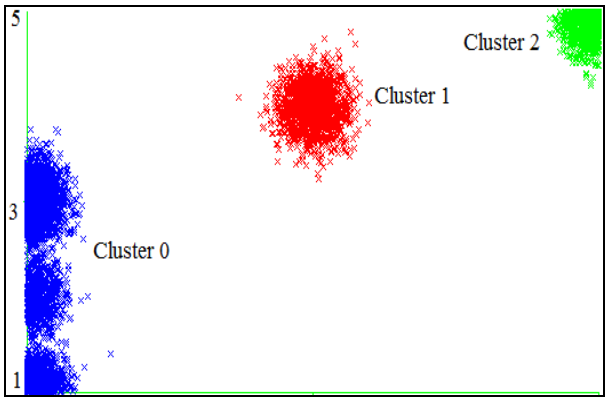

Fig. 3b. Clusters for Q – 24.

The cluster analysis of the first question shows that centroid of full data is 2.7835 which is an average rating for this question on Likert scale. The centroid of cluster 0 is 2 and 9% of the students fall in this cluster. The centroid of cluster 1 is 3.6927 and 63% of the students fall in this cluster which shows that majority of faculty members provided course content, teaching method and evaluation system at the start of the semester. However, 28% students in cluster 2 with centroid only 01 indicate an opinion that is contrary to the cluster 1. It is interesting to note that this opinion is not negligible.

The comparison of questions Q-2, Q-4 and Q-15 (highlighted borders in Table II) offers an interesting pattern by presenting similar clusters for these questions. For question 2, the centroid of cluster 1 is 3.0 and percentage is 27% whereas the centroid of cluster 2 is 4.3929 and its percentage is 37%. This analysis shows that 64% of the students are normally satisfied with the fact that course aims and objectives were clearly stated at the beginning of the period, out of which 37% fall in a group who rated it almost 4.4. Similarly, for Q-4, 30% of the students have the same opinion with centroid 3 of the cluster 1. The cluster 2 has centroid at 4.3537 with 44% students falling in this cluster. We can say that 74% of the students in total in cluster 1 and 2 are satisfied with Q-4, i.e., the course was taught according to the syllabus announced on the first day of class. A similar analysis could be drawn from the data of clusters of Q-15 where 49% of the students are satisfied with Q-15, collectively as the centroid values for cluster 0 and 2 are 4 and 5 respectively. In accordance with the analysis of question 15, 51% of the students are in cluster 1 having a centroid 2.2779 which means that the opinion of almost half of the students indicates that the instructor taught in accordance with the announced lesson plan. It is important to note that Q-2, Q-4 and Q-15 are interlinked and this analysis should be viewed in the same perspective. A similar analysis can be observed for the cluster analysis of Q-16 and Q21-Q-22.

TABLE II: STATISTICS OF CLUSTERS

Question No.	Overall		Cluster 0		Cluster 1		Cluster 2	
	Centroid	%age	Centroid	%age	Centroid	%age	Centroid	%age
1	2.7835	100	2	9	3.6927	63	1	28
2	2.9299	100	1.4162	36	3	27	4.3929	37
3	3.0739	100	1.4479	31	3	29	4.3608	40
4	3.1787	100	1.4441	26	3	30	4.3537	44
5	3.0825	101	1.4547	31	3	29	4.3642	41
6	3.1058	100	2.2241	59	4	26	5	15
7	3.1074	100	4	26	2.2229	59	5	15
8	3.0663	99	4	25	2.2112	60	5	14
9	3.0419	99	4.3836	38	3	29	1.4816	32
10	3.166	99	2.7194	42	4.3842	42	1	15
11	3.0907	100	3	30	4.3789	40	1.4552	30
12	3.1838	100	4	27	2.2283	56	5	17
13	3.0356	100	1.4386	32	3	29	4.3924	39
14	3.2428	100	4	29	2.2652	54	5	17
15	3.2909	100	4	31	2.2779	51	5	18
16	3.2873	100	4	30	2.2872	52	5	18
17	3.1696	101	3	29	4.3857	44	1.429	28
18	3.3985	100	4.399	53	2.7487	34	1	13
19	3.2225	101	1.4244	26	3	29	4.3837	46
20	3.2617	100	1.4216	24	3	29	4.3862	47
21	3.2854	100	4	29	2.2686	52	5	19
22	3.3074	99	4	29	2.2777	51	5	19
23	3.3175	100	4	30	2.2815	51	5	19
24	3.2019	100	2.2689	56	4	27	5	17
25	3.1668	101	3	30	4.3865	43	1.4514	28
26	3.3125	100	2.742	38	4.3853	49	1	13
27	3.2222	100	3	29	4.3807	45	1.4278	26
28	3.1548	99	1.4378	28	3	28	4.3781	43

IV. CONCLUSION

We have presented a model for using one of the data mining approaches i.e. clustering to enhance the learning experience of students that would ultimately improve the quality of educational environment of an educational institution. All these and alike hidden patterns could serve as an important feedback for instructors, curriculum planners, academic managers, and other stakeholders in making informed decisions for evaluating and restructuring curricula as well as teaching and assessment methodologies with a view to improve students' performance in their respective programs.

ACKNOWLEDGMENT

The authors wish to acknowledge the financial support provided by the Al Ghurair University and David Gil for providing the dataset available at UCI website.

REFERENCES

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana: University of Illinois Press, 1949.

[2] J. Richardson, "Instruments for obtaining student feedback: A review of the literature," *Assessment and Evaluation in Higher Education*, vol. 30, issue 4, pp. 387-415, 2005.

[3] T. Connolly, C. Begg, and A. Strachan, *Database Systems: A Practical Approach to Design, Implementation, and Management*, 3rd ed., Harlow: Addison-Wesley, 2001, pp. 687.

[4] A. Durfee, S. Schneberger, and D. L. Amoroso, "Evaluating students computer based learning using a visual data mining approach," *Journal of Informatics Education Research*, vol. 9, pp. 1-28, 2007.

[5] A. R. Anaya and J. G. Boticario, "A data mining approach to reveal representative collaboration indicators in open collaboration frameworks," in *Proc. the 2nd International Conference on Educational Data Mining*, 2009, pp. 210-219.

[6] W. Wang, J. Weng, J. Su, and S. Tseng, "Learning portfolio analysis and mining in SCORM compliant environment," in *Proc. the 34th ASEE/IEEE Frontiers in Education Conference*, vol. 1, 2004.

[7] B. Shih, K. R. Koedinger, and R. Scheines, "Unsupervised discovery of student learning tactics," in *Proc. the 3rd International Conference on Educational Data Mining*, 2010.

[8] R. Hubscher, S. Puntambekar, and A. H. Nye, "Domain specific interactive data mining," in *Proc. Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling*, 2007.

[9] K. E. Maull, M. G. Saldivar, and T. Sumner, "Online curriculum planning behavior of teachers," in *Proc. the 3rd International Conference on Educational Data Mining*, 2010, pp. 121-130.

[10] C. Lee, "Diagnostic, predictive and compositional modeling with data mining in integrated learning environments," *Computers & Education*, vol. 49, no. 3, pp. 562-580, 2005.

[11] B. Dogan and A. Y. Camurcu, "Visual clustering of multidimensional educational data from an intelligent tutoring system," *Computer Applications in Engineering Education*, vol. 18, issue 2, pp. 375-382, 2008.

[12] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaıane, "Clustering and sequential pattern mining of online collaborative learning data,"

IEEE Transactions on Knowledge and Data Engineering, vol. 21, issue 6, pp. 759-772, 2009.
[13] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency vs interpretability of classifications," *Biometrics*, vol. 21, pp. 768-769, 1965.
[14] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proc. 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 1967, vol. 1, pp. 281-297.
[15] G. Gunduz and E. Fokoue, *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science, 2013.
[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, issue 1, 2009.

M. Abaidullah Anwar is working as an associate professor and deputy dean of College of Engineering and Computing in Al Ghurair University, UAE. He received his doctorate of engineering with specialization in object-oriented databases from Kyushu Institute of Technology, Japan in 2001. Since 2001, he has been affiliated with renowned universities in GCC and Pakistan. The research interests of Anwar include data mining and data warehousing, database systems, query optimization and curriculum development.

Naseer Ahmed is working as a director in institutional effectiveness and planning at Al Ghurair University, UAE. His professional experience spans a number of academic assignments held at highly acclaimed institutions in South East Asia, South Asia, Canada, Saudi Arabia and UAE. Since last thirty years, he has been publishing his research work in a number of prestigious international journals. Currently his active areas of interests are quality enhancement, institutional effectiveness and accreditation, instructional methodology, and student assessment.

Edriss Ali is working as the acting dean of the College of Engineering and Computing at Al Ghurair University, UAE. He obtained his M.Sc. and PhD from University of Newcastle Upon Tyne and has 25 years of extensive teaching and research experience in electrical, electronics and computer engineering fields. Edriss carried out a number of consultancy work for UNDP and UNFPO and is a member of Engineering Society Sudan, IEEE, International Microwave Power Institute and AMPERE and worked as a visiting faculty in University of Liverpool, Newcastle Upon Tyne (UK) and Darmstadt Institute of Technology (Germany). His recent interests include curriculum development, student assessment, and program evaluations.