Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

The K-Means Algorithm for Generating Sets of Items in Educational Assessment

Lala Septem Riza[1], Rendi Adistya Rosdiyana[1], Asep Wahyudin[1], Alejandro Rosales Pérez[2]

[1] Department of Computer Science Education, Universitas Pendidikan Indonesia, Jln. Setiabudhi 229, Bandung, Indonesia

[2] School of Engineering and Science, Tecnologico de Monterrey, Monterrey, Mexico

Correspondence: E-mail: lala.s.riza@upi.edu

ABSTRACTS

In a national-scale educational assessment system, such as the National Examination, the need for several sets of questions that have the same level of difficulty is very required to avoid cheating by students. Therefore, the objective, which is to make a set of questions with the same level of difficulty automatically, is done in this research. It used a machine learning approach, namely K-Means. To achieve this goal, several following procedures need to be implemented. Firstly, we need to create banks of questions to be assigned to students. Then, we build training data by determining the value of each question based on Bloom's Taxonomy, item characters/types, and other parameters. Then, with utilizing K-Means, several cluster centers are obtained to represent the uniformity of the questions in the cluster members. By using several heuristics criteria defined previously, several sets or packages of questions that have the same characteristics and difficulty levels are obtained. From the experiments conducted, the analysis with descriptive (i.e., mean, standard deviation, and data visualization) and inference (i.e., ANOVA) statistics of results are presented showing that questions of each sets have the same characteristics to ensure the fairness of examinations. Moreover, by using this system, the contents of the questions in the generated set do not need to be the same, the package of questions can be generated automatically quickly, and the level of the difficulties can be measured and guaranteed.

ARTICLE INFO

Article History:
Received 23 Feb 2020
Revised 20 Aug 2020
Accepted 15 Sep 2020
Available online 20 Jan 2021

Keywords:
Education assessment,
Machine learning,
Data analysis,
Educational evaluation,
Intelligent systems.

1. INTRODUCTION

Since to know the quality of education and the ability level of students is an important task to improve the whole of education system, educational evaluation should be done systematically (Scheerens and Glas, 2003; Tyler, 1942). Therefore, we should evaluate and analysis performances of education in all stages (i.e., input, teaching and learning processes, output, instruments, curriculum, etc.). A part of the educational evaluation, namely educational assessment, is used for obtaining information about the understanding levels of students to the materials that have been taught.

The goals of educational assessment are basically aimed to placement, formative, summative, diagnostic, and selective assessments. The first one is used to put a student into a certain level/class according to the prior knowledge/achievement so that we have the same ability on the class. To know a gap between students' knowledge and teachers' instructions is the formative assessment while the summative one is aimed to give a final course grade of each student (Harlen and James, 1997). Meanwhile information regarding students' difficulties during learning processes can be obtained through the diagnostic assessment, the test used for filtering or choosing some best participants is called by the selective assessment.

Moreover, many ways can be used to evaluate students' performance, for example: written tests (i.e., essay, multiple choices, etc.) and oral tests (e.g., interview and observations). In this research, we focus on the written tests. An issue, that could be happened, in the written test is how to build some sets of questions/items that provide the same characteristics and difficulties. The sets are necessary to avoid cheating among students. The usual way used to build sets of questions is such as by randomizing the order of questions and by modifying the options of answers. Modifying of questions is rare to be done because this task is not easy and spends a lot of time. So it can be seen that this issue should be solved by finding a strategy to generate sets of questions automatically.

Therefore, this research is aimed to generate some packages/sets of questions automatically. It should be noted that to ensure fairness all sets should contain questions with same characteristics and difficulties. To achieve this objective, we consider to utilize the K-Means algorithm (Bansal *et al.*, 2017). It is a classic unsupervised learning method included in Machine Learning (Mitchell *et al.*, 1990) to define cluster centers and their members that have the same characteristics. There are some implementations of K-Means showing its contributions in dealing with various problems. For example, K-Means was used to determine shuttlecock placement and stroke types in badminton (Riza *et al.*, 2018). Related on generating sets of questions as our previous research, a variant of K-Means was used (Riza *et al.*, 2017). Additionally, a method in Machine Learning, called the apriori association rule, was utilized to detect aspects of students' difficulty and its recommendations (Munir *et al.*, 2018).

2. MATERIALS AND METHODS

Figure 1 is the proposed method used in this research for generating sets of items by using K-Means. It was adopted from the previous research in (Riza *et al.*, 2017). Basically there are three stages as follows:

2.1. Data preparation.

This stage is aimed to generate data training, which is the data used for training the algorithm so that we obtain sufficient model for building sets of items. There are some processes in this stage, as follows:

a. Collecting questions/items: In the data preparation step, we firstly need to

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

collect items on a particular subject. To simplify in this research we just collect 638 questions from three following chapters: computer and networking, application layer, and transport layer, in five text books used many universities in the worlds entitled as follows:

- Computer Network by Tanembaum & Wetherall (2011).
- Computer Network: A System Approach by Peterson & Davie (2011).
- Computer networking: Principles, Protocols , and Practice by Bonaventure (2011).
- Internetworking With TCP/IP, Principles Protocols , and Architecture by Comer (2006).
- Computer Network: A Top Down Approach by Kurose & Ross (2013).

b. Defining features: It means that we define some characteristics or features on each question. Therefore, this task is useful to determine whether a question is the same as another one or not. In this research, we had defined 14 features as follows:

- C1: The first level of cognitive domain (i.e., remembering) that has a value between 0 and 1.
- C2: The second level of cognitive domain (i.e., understanding) that has a value between 0 and 1.
- C3: The third level of cognitive domain (i.e., applying) that has a value between 0 and 1.
- C4: The fourth level of cognitive domain (i.e., analysing) that has a value between 0 and 1.
- C5: The five level of cognitive domain (i.e., evaluating) that has a value between 0 and 1.
- C6: The six level of cognitive domain (i.e., creating) that has a value between 0 and 1.
- TS: Question types containing several options: essay, correct-incorrect, common multiple choice, variant multiple choice, and matching that have values 1, 2, 3, 4, and 5, respectively.
- BR: Picture problem that means whether the question contains a figure or not.
- SC: Story problem that means whether the question is a story or not.
- PMG: Programming question that means whether there is a code in the question or not.
- TK: Level of difficulty that has a value between 0 and 1 to represent from very easy to very difficult.
- BW: Expected time in minutes that is needed by student to answer the question.
- MTS: Mathematics problem or not.
- DSK: The question contains a complex analysis or not.

It can be seen that these aspects are hidden features embedded in each question. Moreover, C1 until C6 are six levels of cognitive domain defined by Bloom *et al*. (1956).

c. Calculating values of all features: After we define the features, the respective values can be determined by human experts for all questions. To be more objective, we can determined the average values from some experts.

d. Arranging the data training: Then all values of the features can be constructed to be a table, namely data training.

2.2. Clustering using K-Means

Basically, in this step we implement and execute the algorithm K-Means with supplying some input data, such as data training, maximum iteration, and number of cluster centers. Regarding the algorithm, the detailed explanation can be found in (Na *et al*., 2010). In short , it contains four steps as follows:

a. Initialization of cluster centers: It means that at the first step we need to choose cluster centers. It can be done randomly. It should be noted that the

numbers of cluster centers are defined by users.

b. Assignement step: After choosing the cluster centers, distances all data to cluster centers are calculated to determine the cluster member. So, the instances included in the same cluster mean they are closed each other.

c. Update step: Then we update the position/location of cluster centers by averaging all values of all members included in the cluster.

d. Repeat the processes: The same processes are repeated until maximum iteration or convergence.

The output of this step is cluster centers with their members representing questions that have the same/closed characteristics. It should be noted that cluster centers represent sets of items while items are their members.

3. Building sets of items: The last thing that should be done is to pick question according to the cluster centers and their members. For example, we need to generate three packages of items where each set contains five questions. Therefore, we just need to choose one cluster center randomly. Then, we pick three questions from the selected cluster center to be a member of three packages. So, now we have one question for each set that have the same characteristics. Then, we repeat these process until we obtain five questions for all sets. It should be noted in these processes we can put other criteria to ensure the quality of questions, such as the duplication of questions is not allowed and the proportion of the selected question from all chapter is considered. Finally, by passing this step we obtain sets of items that have similar characteristics of questions so that fairness can be ensured.

3. RESULTS AND DISCUSSION

After designing the proposed computational model as explained previously, we build a web based application as shown in **Figure 2**. It is the result page showing some packages of items generated by the system. In the system, we also provide other functionalities, such as creating a new project, creating and loading items along with metadata required to build data, and then other parameters (e.g., numbers of sets and K-Means parameters).

Moreover, we had run some simulations to validate the performance of the proposed model. By using the data training obtained from five textbooks as introduced before, for example, we need to build three sets containing 10 questions obtained from 638 questions in the textbooks. The result can be seen as follows:

1. The first set contains the following questions with IDs: {1.95; 3.186; 2.38; 3.128; 3.213; 1.92; 1.16; 1.65; 1.224; 2.24}.
2. The second set contains the following questions with IDs: {1.84; 1.115; 2.38; 2.112; 3.166; 2.47; 2.68; 1.168; 3.112}.
3. The last set contains the following questions with IDs: {1.58; 1.17; 1.65; 2.162; 3.19; 3.95; 1.132; 2.16; 2.2; 1.107}.

It should be noted that the ID of questions represents the chapter and question number. For example, a question with the ID 1.58 means that it is from the first chapter and the 58th question. According to the results, it can be seen that we obtain equal proportions of chapters. In other words, all questions represent all chapters for all sets. Moreover, we also analysis the results based on the values of the features of all questions. The average of values can be seen in **Table 1**.

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

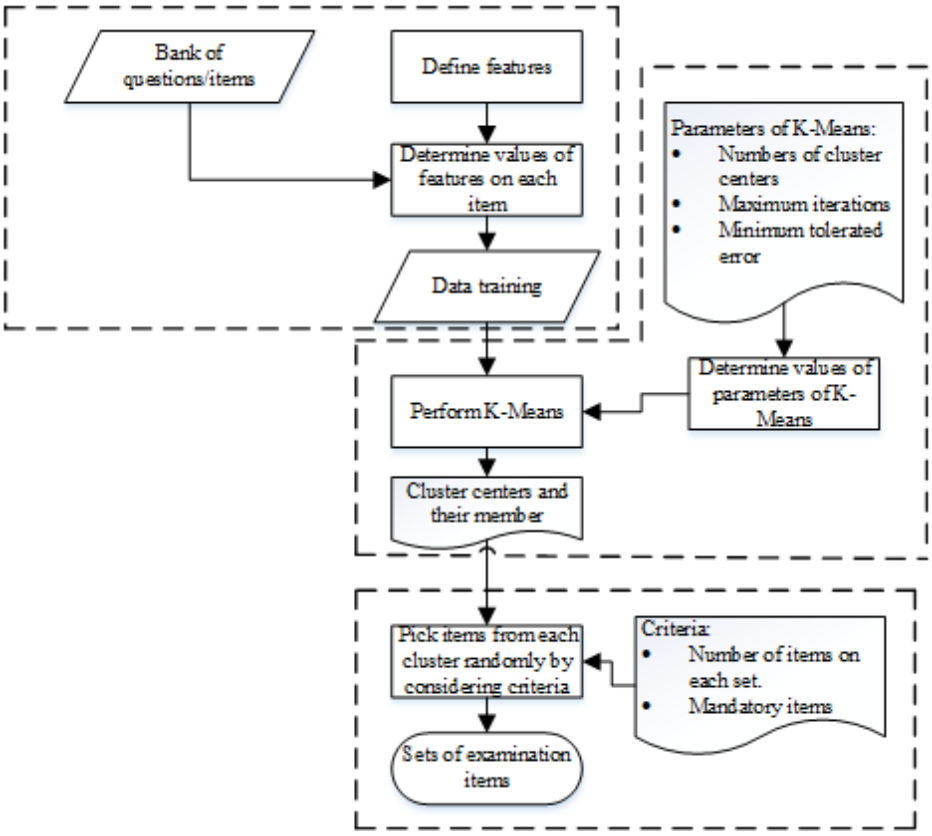

Figure 1. The proposed computational model for generating sets of items adopted from (Riza *et al*., 2017)

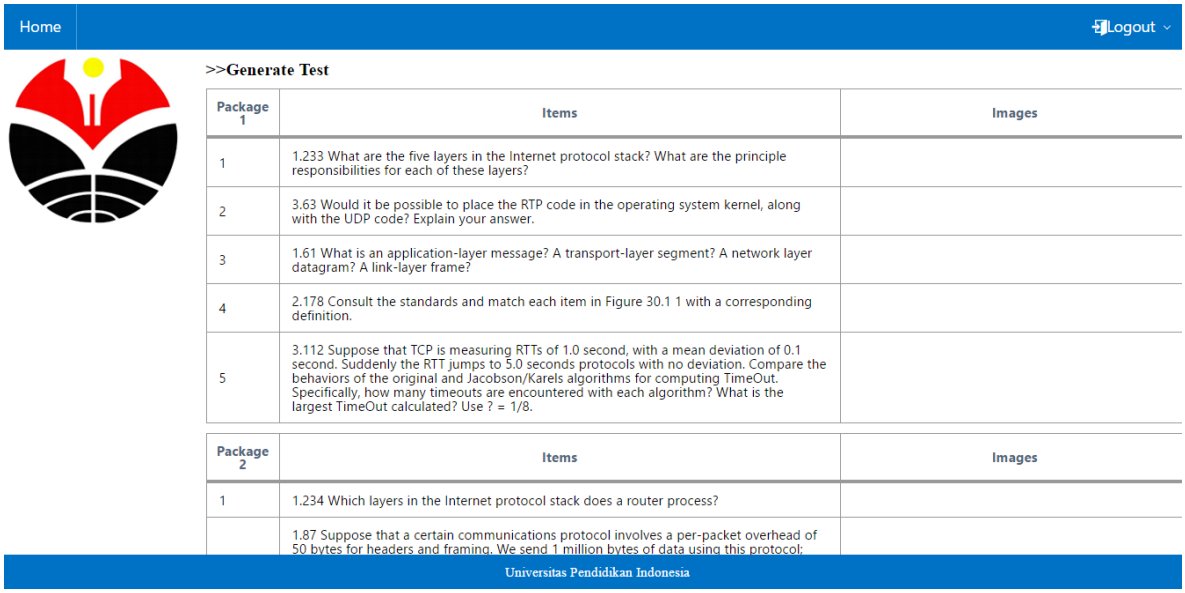

Figure 2. Graphical user interface of the result page in the proposed system.

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

Tabel 1. The average of features values on all generated sets

Sets	TS	C1	C2	C3	C4	C5	C6	BR	SC	PMG	TK	BW	MTS	DSK
1	11	0.5	2.4	1.5	3.7	1.9	0	0	0	0	6.3	134	3	1
2	10	0.7	2.7	1	3.4	2.2	0	1	1	0	5.6	130	1	1
3	10	1.1	3	0.5	4	1.1	0.3	0	1	1	5.1	110	0	1

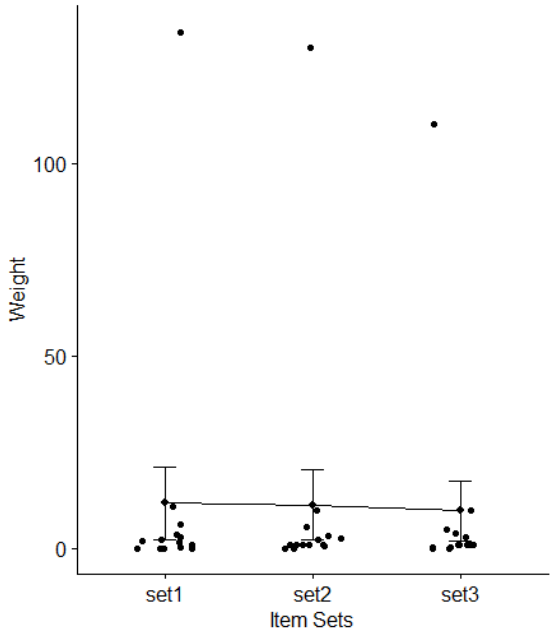

Figure 3. Data visualization of the average values of features

According to the average value in **Table 1**, we obtain mean for all sets: 11.8, 11.4, and 9.86 and standard deviation for all sets: 35.3, 34.2, and 28.9. It means that all sets relatively have similar characteristics on 14 features. To explain in more detail, the data visualization of the average values of all features can also be seen in **Figure 3**.

Additionally, we perform the analysis of variance (ANOVA) test with $\alpha = 0.05$. The following are hypotheses constructed to prove that items in each set have similar characteristics:

- H0: There is no difference between the average of feature values on Set 1, 2 and 3.
- H1: There is a difference between the average of feature values on Set 1, 2 and 3.

After running ANOVA we obtained p-value: 0.987. It means that H0 is accepted. Therefore, we can state that the characteristics of equations in all sets are relatively similar/same so that the fairness of examination can be kept. Additionally, we can also compare with our previous research (Riza *et al.*, 2017) that shows that by using Fuzzy C-Means the system provides the same results

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

even though the question indices are different from this research.

In the future, we have a plan to improve the model by using different alternative methods, such as Rough Sets (Riza *et al.*, 2014), Naïve Bayes (Mulyani *et al.*, 2016), and Fuzzy Sets (Riza *et al.*, 2015). These methods are included in Machine-Learning methods so that the computational model built can be smart. Moreover, we also propose a computational model to generate the bank of questions (Riza *et al.*, 2019) and values of features of the questions automatically. Various intelligent classifiers can be used for improve the computational model (Alasker *et al.*, 2017). We can also improve the computational cost by implementing data streaming (Mediayani *et al.*, 2013).

4. CONCLUSION

The contributions of this research are that firstly we provide a computational model using K-Means for generating sets of items that have the same characteristics to ensure the fairness of the examination. Before performing the K-Means, we also proposed 14 features to be used for building data training. The 14 defined features, such as Bloom's taxonomy, types of questions, etc, represents inside characteristics on questions. Moreover, an experiment was done to validate the model. According to the results and their analysis using descriptive (i.e., mean, standard deviation, and data visualization) and inference (i.e., ANOVA) statistics, we can state that the proposed system produced the sets of items as required.

5. AUTHORS' NOTE

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article. Authors confirmed that the data and the paper are free of plagiarism.

6. REFERENCES

Alasker, H., Alharkan, S., Alharkan, W., Zaki, A., and Riza, L. S. (2017). Detection of kidney disease using various intelligent classifiers. In *2017 3rd International Conference on Science in Information Technology (ICSITech)* (pp. 681-684). IEEE.

Bansal, A., Sharma, M., and Goel, S. (2017). Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining. *International Journal of Computer Applications*, *157*(6), 0975-8887.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). Taxonomy of educational objectives-handbook 1: Cognitive domain. New York: David McKay Company.

Bonaventure, O. (2011). Computer networking: principles, protocols and practice (pp. 41-45). Washington: Saylor foundation.

Comer, D. (2006). Internetworking with TCP/IP, principles protocols, and architecture. Englewood Cliffs: Prentice Hall.

Harlen, W., and James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy* and *Practice*, *4*(3), 365-379.

Kurose, F. J., and Ross, W. K. (2013). Computer Networking: a Top-Down Approach. New Jersey: Pearson.

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |

Mediayani, M., Wibisono, Y., Riza, L. S., and Pérez, A. R. (2019). Determining trending topics in twitter with a data-streaming method in R. *Indonesian Journal of Science and Technology*, *4*(1), 148-157.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., and Waibel, A. (1990). Machine learning. *Annual Review Of Computer Science*, 4(1), 417-33.

Mulyani, Y., Rahman, E. F., and Riza, L. S. (2016, October). A new approach on prediction of fever disease by using a combination of Dempster Shafer and Naïve bayes. *2016 2nd International Conference on Science in Information Technology (ICSITech)* (pp. 367-371). IEEE.

Munir, Farasyi, G., Setiawan, W., Fahsi, M., and Riza, L. S. (2018). The association rule method for mapping and recommendation system on students' difficulties. *Transylvanian Review*, 26(32), 1-9.

Na, S., Xumin, L., and Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. *2010 Third International Symposium on intelligent information technology and security informatics* (pp. 63-67). IEEE.

Peterson, L. L., and Davie, S. B. (2011). Computer network a system approach. Morgan Kauffman.

Riza, L. S., Firmansyah, M. I., Siregar, H., Budiana, D., and Rosales-Pérez, A. (2018). Determining strategies on playing badminton using the Knuth-Morris-Pratt algorithm. *Telkomnika (Telecommunication Computing Electronics and Control)*, 16(6), 2763-2770.

Riza, L. S., Awaludin, R., Sutarno, H., Munir, and Wibawa, A. P. (2017). A model for auto generating sets of examination items in educational assessment by using fuzzy c-means. *World Transactions on Engineering and Technology Education*, 15(2), 114-119.

Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślezak, D., and Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". *Information Sciences*, 287, 68-89.

Riza, L. S., Bergmeir, C. N., Herrera, F., and Benítez Sánchez, J. M. (2015). frbs: Fuzzy rule-based systems for classification and regression in R. *Journal of Statistical Software*, 65(6), 1-30.

Riza, L. S., Pertiwi, A. D., Rahman, E. F., and Munir, Abdullah, C. U. (2019). Question generator system of sentence completion in toefl using nlp and k-nearest neighbor. *Indonesian Journal of Science and Technology*, *4*(2), 294-311.

Scheerens, J., and Glas, C. A. (2003). Thomas SM, Thomas S. Educational evaluation, assessment, and monitoring: A systemic approach. *Taylor and Francis*.

Tanembaum, S. A., and Wetherall, J. D. (2011). Computer network fifth edition. Boston: Pearson.

Tyler, R. W. (1942). General statement on evaluation. *The Journal of Educational Research*, 35(7), 492-501.

DOI: https://doi.org/10.17509/ijost.v6i1.31523 |
p- ISSN 2528-1410 e- ISSN 2527-8045 |