A model for auto generating sets of examination items in educational assessment by using fuzzy c-means

Lala S. Riza†, Rabihi Awaludin†, Heri Sutarno†, Munir† & Aji P. Wibawa‡

Universitas Pendidikan Indonesia, Bandung, Indonesia†
Universitas Negeri Malang, Malang, Indonesia‡

ABSTRACT: Generating sets of examination items in educational assessment is not an easy task to do since it must be ensured that all sets are similar in some aspects, such as characteristics, difficulties, qualities, etc. A trivial way to generate the sets is by shuffling the position of items. However, this manner cannot deeply discover students' capabilities and still provides opportunities for cheating. Therefore, this research is aimed at designing a model for generating sets of items automatically by using a clustering method, namely: fuzzy c-means (FCM). This algorithm is used to build cluster centres from data according to the following parameters: Bloom's taxonomy, difficulties levels, expected response time and others (e.g. story, mathematics or programming questions). After obtaining the cluster centres, members were randomly chosen for each set. To evaluate the proposed model, 636 question items collected from five textbooks of computer networking were used. Then, the results, which are some sets of examination items, were analysed statistically along with discussing from the data-mining perspective to measure the similarity in each set.

INTRODUCTION

Educational assessment is an integral subsystem of the overall education system. It becomes an important factor to show the quality of education, so that the education system can be improved continuously. The purpose of education evaluation is a process of collecting data to determine the position of understanding of students/learners in the learning process [1]. Moreover, it can be interpreted as a path involving three stages: input, learning process and output. These processes need to be monitored and reviewed continuously by adding the process of evaluation of these stages as illustrated in Figure 1. It can be seen that most educational evaluation is conducted on three consecutive parts of educational systems. Firstly, it can be used for selecting prospective students entering education. Then, it can be used for providing information related to the ability of learners to the material that has been taught, known as summative and formative assessments. Lastly, educational assessment is conducted as a final test used for measuring whether students can proceed to the next level or not.

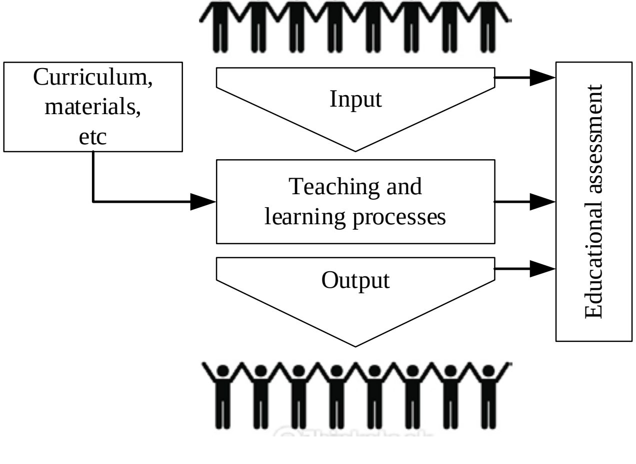

Figure 1: An educational system involving its evaluation.

Basically, the education assessment can be done by many strategies, such as interviews, examinations/tests, etc. Due to the ease and practicality, the assessments are carried out by giving the tests represented by some questions to examinees/students. In this case, some issues should be solved. For example, examiners should create representative questions for showing, evaluating and even filtering the abilities of examinees. Moreover, in some cases, e.g. national examinations held by the government, examinees/committees need to provide a single set of questions. Some sets of questions are prepared to avoid cheating and to comply with the official standard. It should be noted that for fairness, one needs to ensure that the questions included in each set should have the same characteristics as others. This task is not particularly easy and takes a lot of time.

This research attempts to generate sets of examination items automatically that provide the same characteristics in each set. To achieve this objective, the authors consider an algorithm included in machine-learning approaches, which is fuzzy c-means (FCM) [2-4]. It is an algorithm used for grouping data having similar values of properties into one cluster. In this research, the authors considered the following features: the cognitive domain (i.e. Bloom's taxonomy [5]), questions types (i.e. essay, multiple choices, multiple choices with distracter, simple answers, etc), difficulty levels, expected time to answer and other features. So, in short FCM constructs cluster centres and memberships from these values corresponding to each question. After that, sets of questions can be generated by picking items from each cluster. Moreover, this research is carried out by using FCM since many problems have been successfully solved, such as service quality assessment of shared use road segments [6], clustering on learning media preference [7] and an application for medical image segmentation [8].

The remainder of this article is structured as follows. The next section briefly introduces the FCM clustering method. In the following section, the authors describe in detail the design and implementation of the proposed model for auto generating sets of examination items. Then, experimental design, results and discussion are delivered in the next section. Finally, the last section concludes the article.

CLUSTERING WITH FUZZY C-MEANS

In machine learning, clustering is included as unsupervised learning, meaning that a model is built by deducing structures from input data. Fuzzy c-means is a clustering method used for grouping a set of objects, so that members of a cluster are similar to each other in the same cluster and are different from members included in other clusters.

Fuzzy c-means was developed by Dunn [2] and, then, improved by Bezdek in 1981 [3][4]. In contrast to the classical clustering techniques (when an object will only be a member of a particular cluster), in FCM, an object can be a member of multiple clusters. So, the boundaries of the cluster are called soft. It is a clustering technique in which each data point in a cluster is determined by the value of cluster membership, which is between 0 and 1. Firstly, it determines cluster centres by calculating the average of each cluster. These cluster centres will be revised by recalculating cluster centres and their membership values along with given maximum iteration, so that one obtains a minimum of the objective function. The detailed algorithm can be shown as follows:

Input:

- Data training represented in a matrix (*n x m*), where *n* is the number of observations, while *m* is the number of features. For example, X_{ij}means a value on the i^{th} data point with the j^{th} feature.
- The numbers of cluster centres (*c*).
- Maximum iteration (*MaxIter*).
- Minimum tolerated error (ζ).

Output: Cluster centres and their members

Algorithm:

1. Generate initialisation of the matrix $\mu_{ik} \in [0,1]$, where *i* is the total number of data and *k* is the total number of cluster. It represents factors that are the ones taken from the membership functions.
2. Calculate k^{th} cluster centre (v_{kj}):

$$v_{kj} = \frac{\sum_{i=1}^{n}\left(\mu_{ik}{}^{2} * X_{ij}\right)}{\sum_{i=1}^{n}\mu_{ik}{}^{2}}$$

3. Calculate the objective function on the iteration *t*:

$$F_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - v_{kj}\right)^{2}\right]\left(\mu_{ik}\right)^{w}\right)$$, where *w* is any real number greater than 1.

4. Update the matrix μ_{ik}:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij}-v_{kj})^{2}\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij}-v_{kj})^{2}\right]^{\frac{-1}{w-1}}}$$, with *i = 1, 2, …, n*.

5. Check stopping criteria:

   a. If $(|F_t - F_{t-1}| < \zeta)$atau(*t*>*MaxIter*) then stop.
   b. Else, set *t* = *t* + *1* and continue to Step 2.

DESIGNING AND IMPLEMENTING A MODEL FOR AUTO GENERATING SETS OF ITEMS

Figure 2 shows the proposed model used for generating sets of examination items automatically.

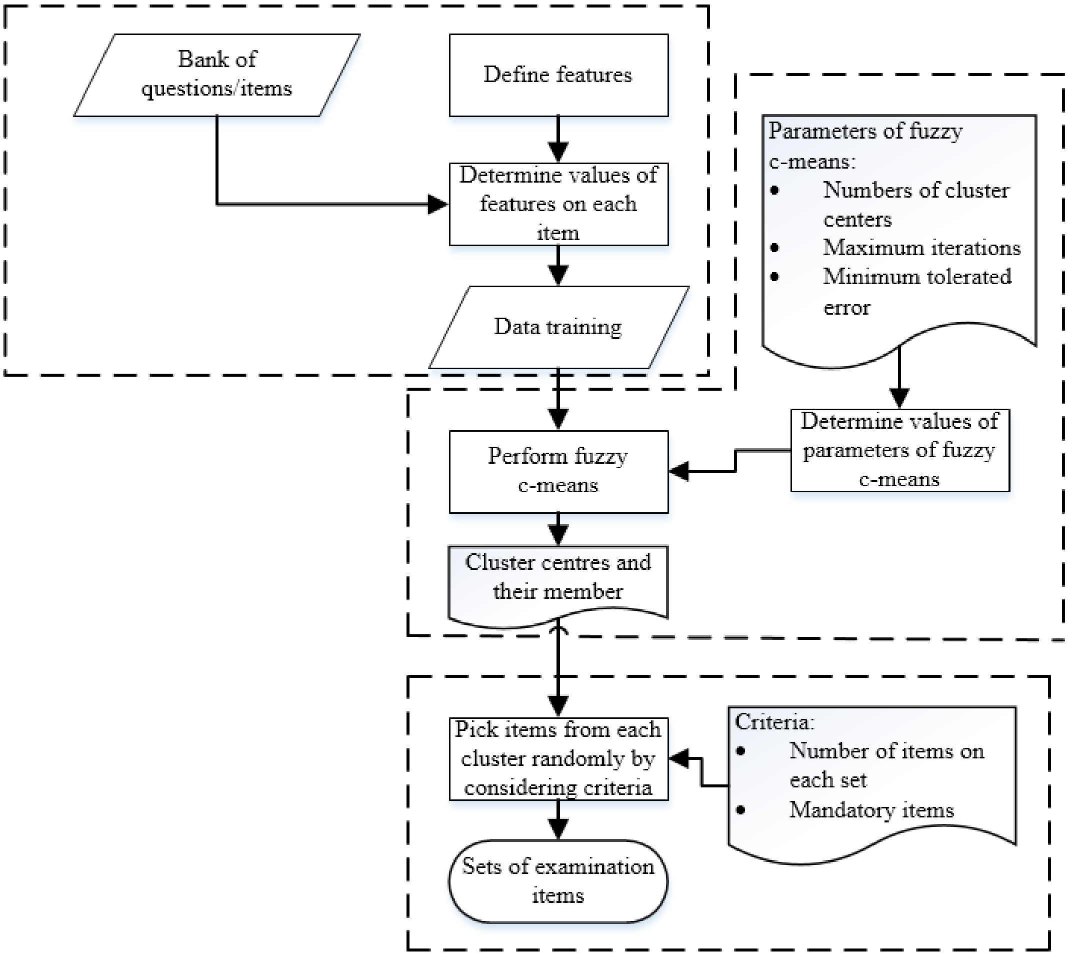
Bank of questions/items
Define features
Determine values of features on each item
Data training
Parameters of fuzzy c-means:
• Numbers of cluster centers
• Maximum iterations
• Minimum tolerated error
Determine values of parameters of fuzzy c-means
Perform fuzzy c-means
Cluster centres and their member
Pick items from each cluster randomly by considering criteria
Criteria:
• Number of items on each set
• Mandatory items
Sets of examination items

Figure 2: The proposed model for generating sets of examination items.

It can be seen that basically the processes involve three big stages as follows:

1. Data collection: in this step, firstly one needs to create or gather questions to be examination items. It should be noted that numbers of items should be enough and representative to be generated. Then, some features corresponding to data should be defined. In this case, the following features are considered:

   a. Six Bloom's taxonomy of cognitive domain [5]: knowledge (*C1*), comprehension (*C2*), application (*C3*), analysis (*C4*), synthesis (*C5*) and evaluation (*C6*). Values of each feature are between 0 and 1.
   b. Question types (*QT*): one considers six values from 1 to 6 for this features that represent essay, correct/wrong, multiple choices with distracter, multiple choices with variations, matching the answer and short answer. For example, if a question is in essay, then, the item should be 1 for its question type.
   c. Picture question (*PC*): it means whether the question contains a picture or not. If it is yes, then, the value is 1, otherwise it is 0.
   d. Story question (*SC*): if the question is a story one, then, one needs to define 1 as the value, otherwise it is 0.
   e. Programming code question (*PE*): if the question contains programming code, then, the value is 1, otherwise it is 0.
   f. Difficulty level (*DF*): one can define the level between 0 and 1.
   g. Expectation time for answering (*DS*): the value of this feature is in minutes, such as 5, which means the question should be done in 5 minutes.
   h. Mathematics question: if the question involves some equations/mathematics, then, the value is 1, otherwise it is 0.
   i. Discussion question (*DQ*): if the question is a story one, then, the value is 1, otherwise it is 0.

Moreover, one can actually define other features that represent some characteristics of questions. After defining these features, their values need to be defined for all questions. It can be done by human experts or teachers, so that the values are objective and representative. The questions along with their feature values are called data training.

2. In the second step, the FCM method is used with some given parameters, i.e. numbers of cluster centres, maximum iterations and minimum tolerated error. The output of this step is cluster centres with their members. In this case, the members contain indices (or ID) of items.
3. After obtaining clusters, questions can be picked randomly from each cluster or random cluster by considering criteria given by examiners/teachers. For example, three sets need to be generated, in which there are 10 questions in each set. So, after generating six cluster centres by using FCM, one cluster is randomly chosen from six clusters. From this cluster, one question is taken randomly for each set. Then, these processes are repeated until the numbers of questions is met. It should be noted that these processes should consider some other criteria given by examiners, such as questions that must be included in each set, no duplicated questions on a particular set, and the proportion of chosen questions from each chapter.

The final result obtained is several sets of questions that provide some characteristics on each set.

EXPERIMENTAL DESIGN

In these experiments, there are 638 questions that have been obtained from three chapters (i.e. Computer and Networking, Application Layer and Transport Layer) of the following famous books on Computer Networking:

1. Computer Network by Tanembaum and Wetherall [9].
2. Computer Network: a System Approach by Peterson and Davie [10].
3. Computer networking: Principles, Protocols, and Practice by Bonaventure [11].
4. Internetworking With TCP/IP, Principles Protocols, and Architecture by Comers [12].
5. Computer Network: a Top Down Approach by Kurose and Ross [13].

Then, three experts in computer networking define values related to features mentioned in the previous section. For example, Table 1 shows three questions with their values in each feature.

Table 1: A part of data training including 3 questions.

ID	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*QT*	*PC*	*SC*	*PE*	*DF*	*DS*	*MT*	*DQ*
1.1	0.1	0.2	0.4	0.3	0	0	1	0	1	0	0.8	20	0	0
1.2	0.6	0.2	0	0.2	0	0	1	0	0	0	0.6	10	0	0
1.3	0	0.2	0.3	0.5	0	0	1	0	0	0	0.7	10	0	0

After data training is ready to be used, some parameters are determined for generating sets of items. For this experiment, the following parameters have been defined: the number of clusters is three, the numbers of sets are three with ten items in each set, the maximum iteration is 100 and the minimum tolerated error is 0.1.

RESULTS AND DISCUSSION

After running the simulation as the design illustrated in the previous section, three sets of items were obtained with a computation cost of around 33.9 seconds. The indices of items on each set can be seen as follows:

1. The 1st set contains the following item IDs: {3.159, 2.188, 3.169, 3.69, 3.182, 1.53, 1.89, 3.20, 2.158, 1.73}.
2. The 2nd set contains the following item IDs: {1.138, 2.136, 1.38, 1.80, 3.159, 1.11, 2.161, 2.151, 3.166, 1.42}.
3. The 3rd set contains the following item IDs: {{2.138, 2.69, 2.196, 3.135, 3.96, 1.28, 2.174, 1.133, 1.75, 3.17}

It should be noted that for example, ID 3.159 means that it is the 159th question on the third chapter.

From this result, one can analyse whether the sets generated by the proposed model met the following aspects:

- Each set contained some questions from all chapters.
- There was no repetition of questions in each set.
- There was no a dominant chapter contributing questions on a particular set.

Moreover, one can deeply take a look at the average values of all the features as illustrated in Tables 2, 3 and 4. Intuitively, one can state that all sets have quite similar average values of the features. Furthermore, the averages of all sets were validated by using 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:

H_1: There is a difference between the average of feature values on Set 1, 2 and 3.
H_0: There is no difference between the average of feature values on Set 1, 2 and 3.

Table 2: Values and their average of features on the 1st set.

ID	C1	C2	C3	C4	C5	C6	QT	PC	SC	PE	DF	DS	MT	DQ
3.159	0	0.2	0	0.5	0.3	0	1	0	0	0	0.8	20	0	0
2.188	0	0.2	0	0.5	0.3	0	1	0	0	0	0.6	15	0	0
3.169	0	0.2	0.2	0.4	0.2	0	1	0	0	0	0.7	15	0	0
3.69	0	0.3	0	0.5	0.2	0	1	0	0	0	0.5	10	1	0
3.182	0	0.3	0	0.5	0.2	0	1	0	0	0	0.5	10	0	0
1.53	0	0.5	0.2	0.3	0	0	1	0	0	0	0.7	15	1	0
1.89	0	0.3	0	0.7	0	0	1	0	0	0	0.2	15	0	0
3.20	0	0.2	0.3	0	0.2	0.3	1	0	0	1	0.9	20	0	0
2.158	0	0.2	0	0.5	0.3	0	1	0	0	0	0.5	10	0	0
1.73	0	0.2	0.4	0.2	0.2	0	1	0	0	0	0.8	20	1	0
Average	0	0.26	0.1	0.36	0.19	0.03	1	0	0	0.1	0.62	16	0.3	0

Table 3: Values and their average of features on the 2ndset.

ID	C1	C2	C3	C4	C5	C6	QT	PC	SC	PE	DF	DS	MT	DQ
1.138	0.2	0.3	0.3	0.2	0	0	1	0	0	0	0.5	15	0	0
2.136	0	0.3	0	0	0	0.7	1	0	0	0	0.8	20	0	0
1.38	0.1	0.2	0.3	0.4	0	0	1	0	0	0	0.8	30	0	0
1.80	0	0.2	0.2	0.4	0.2	0	1	0	0	0	0.8	15	0	0
3.159	0	0.2	0	0.5	0.3	0	1	0	0	0	0.8	20	0	0
1.11	0.5	0.5	0	0	0	0	1	0	0	0	0.3	5	0	0
2.161	0	0.2	0.3	0.3	0.2	0	1	0	0	0	0.6	15	0	0
2.151	0	0.3	0	0.5	0.2	0	6	0	0	0	0.5	10	0	0
3.166	0.3	0.2	0	0.3	0.2	0	2	0	0	0	0.7	5	0	0
1.42	0	0.2	0.2	0.3	0.3	0	1	0	0	0	0.8	15	1	0
Average	0.11	0.26	0.16	0.29	0.14	0.07	1.6	0	0	0	0.66	15	0.1	0

Table 4: Values and their average of features on the 3rd set.

ID	C1	C2	C3	C4	C5	C6	QT	PC	SC	PE	DF	DS	MT	DQ
2.138	0	0.3	0.3	0	0	0.4	1	0	0	0	0.6	15	0	0
2.69	0.7	0	0	0.3	0	0	2	0	0	0	0.4	5	0	0
2.196	0	0.3	0	0.4	0.3	0	1	0	0	0	0.6	15	0	0
3.135	0	0.2	0.2	0	0.2	0.4	1	0	0	0	0.8	15	0	0
3.96	0	0.2	0	0.3	0.2	0.3	1	0	0	0	0.8	15	0	0
1.28	0	0.2	0.2	0.2	0.3	0.1	1	0	0	0	0.8	25	0	0
2.174	0	0.3	0	0.7	0	0	1	0	0	0	0.3	10	0	0
1.133	0	0.3	0	0.5	0.2	0	1	0	0	0	0.6	10	0	0
1.75	0.2	0.3	0	0.5	0	0	1	0	0	0	0.2	20	0	1
3.17	0	0.3	0.3	0.2	0.2	0	1	0	0	0	0.7	20	1	0
Average	0.09	0.24	0.12	0.31	0.14	0.12	1.1	0	0	0	0.58	13.5	0.1	0.1

Table 5 shows the ANOVA test results. It can be seen that H_0 is accepted, and that means, there is no difference in the average of feature values in Sets 1, 2 and 3. In other words, sets generated by fuzzy c-means have similar characteristics.

Table 5: Results of ANOVA.

Source of variation	SS	df	MS	F	*p*-value	F crit.
Between groups	0.258062	2	0.129031	0.008374	0.991663	3.238096
Within groups	600.9294	39	15.40845			
Total	601.1875	41				

CONCLUSIONS AND FUTURE WORK

In this research, a model used for generating sets of examination items that provide the same characteristics in each set is presented. It can be done by using a machine-learning method, called FCM, for clustering the data. Some characteristics are used to determine similarity on each set, such as Bloom's Taxonomy, difficulty levels, question types, etc. An experiment showing and validating the model was presented along with its analysis using ANOVA.

As future work, the authors plan to use other methods for comparison, such as fuzzy rule-based systems [14] and rough sets [15]. Afterwards, these techniques will also be implemented for other emerging technologies, such as intelligent tutoring systems [16] and remote laboratories [17].

ACKNOWLEDGEMENTS

The authors acknowledged RISTEKDIKTI (Grant: Programme Unggulan Perguruan Tinggi (PUPT), Universitas Pendidikan Indonesia).

REFERENCES

1. Tyler, R.W., General statement on evaluation. *The J. of Educational Research*, 35, **7**, 492-501 (1942).
2. Dunn, J.C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. of Cybernetics*, 3, 32-57 (1973).
3. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York Plenum Press (1981).
4. Bezdek, J.C., Ehrlich, R. and Full, W., FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10, **2-3**, 191-203 (1984).
5. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R., *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. David McKay Company, Inc. (1956).
6. Beura, S.K., Chellapilla, H., Jena, S. and Bhuyan, P.K., Service quality assessment of shared use road segments: a pedestrian perspective. *Proc 2nd Inter. Conf. on Intelligent Computing and Applications*, Springer Singapore, 653-669 (2017).
7. Jiang, Y.H., The application research of fuzzy c means clustering algorithm in middle school students' network learning media preference. *Applied Mechanics and Materials*, 644, 2051-2054 (2014).
8. Gefeng, Y., Xu, O. and Zhisheng, L., Fuzzy clustering application in medical image segmentation. *Proc. 2011 6th Inter. Conf. on Computer Science and Educ.* (ICCSE), 826-829 (2011).
9. Tanembaum, S.A. and Wetherall, J.D., *Computer Network Fifth Edition:* Boston: Pearson (2011).
10. Peterson, L.L. and Davie, S.B., *Computer Network a System Approach.* Morgan Kauffman (2011).
11. Bonaventure, O., *Computer Networking: Principles, Protocols, and Practice*. The Saylor Foundation (2011).
12. Comer, D., *Internetworking with TCP/IP, Principles Protocols, and Architecture*. Englewood Cliffs: Prentice Hall, 1 (2006).
13. Kurose, F.J. and Ross, W.K., *Computer Networking: a Top-Down Approach.* New Jersey: Pearson (2013).
14. Riza, L.S., Bergmeir, C., Herrera, F. and Benıtez, J.M., FRBS: fuzzy rule-based systems for classification and regression in R. *J. of Stat. Softw*. 65, **1**, 1-30 (2015).
15. Riza, L.S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Slezak, D. and Benıtez, J.M., Implementing algorithms of rough set theory and fuzzy rough set theory in the R package RoughSets. *Info. Sciences*, 287, 68-89 (2014).
16. Wibawa, A.P. and Nafalski, A., Intelligent tutoring system: a proposed approach to Javanese language learning in Indonesia. *World Trans. on Engng. and Technol. Educ.*, 8, **2**, 216-220 (2010).
17. Gadzhanov, S., Nafalski, A. and Wibawa, A.P., Remote and proximal delivery of the laboratory component of electrical and energy systems course. *J. Pendidikan Sains*, 5, **1**, 6-10 (2017)