Re-Imaging Learning Environments
Proceedings of the European Distance and E-Learning Network 2016 Annual Conference
Budapest, 14-17 June, 2016
ISBN 978-615-5511-10-3
ISSN: 2707-2819
doi: https://doi.org/10.38069/edenconf-2016-ac-0024

A BENCHMARKING STUDY OF K-MEANS AND SOM APPROACHES APPLIED TO A SET OF FEATURES OF MOOC PARTICIPANTS

Rosa Cabedo Gallén, Edmundo Tovar Caro, Technical University of Madrid, Spain

Abstract

MOOC format is characterized by the great diversity of enrolled people. This heterogeneity of participants represents a challenging opportunity in order to identify underlying relationships in the internal structure of features that make up participants' profiles. This paper has the aim of identifying and analyzing a feasible set of MOOC participants' profiles with the use of two unsupervised clustering techniques, K-Means as a partitional clustering algorithm and Kohonen's Self-Organizing Maps (SOMs), hereinafter SOM, as a representative technique of Artificial Neural Networks (ANNs).

The selected dataset comes from MOOCKnowledge project data collection, which provides the opportunity to work with real-world data from hundreds of people. The clustering approach is performed by running both algorithms with a subset of participants' features. The clustering evaluation is achieved with some indices, an intra-cluster measure and an overall quality criterion for K-Means, and two measures related to topological ordering for SOM.

The analysis of internal structure with the help of the matrix of prevalence levels shows that there are similarities between the two resulting clustering on the one hand and some pinpointed differences that cannot be evaluated in advance without the opinion of an expert familiarized with the specifications of the MOOC on the other.

The comparison of matrix of prevalence levels of participants' features for the resulting profiles of both K-Means and SOM clustering cannot be considered conclusive after a preliminary study of the results of the clustering, and for sure there is a long way in order to help designers and other policy-makers to provide a methodological guide on how to identify and select the appropriate clustering according to several quality criteria and therefore, to raise the likelihood of finding a clustering that best fits.

Introduction

This paper has the final purpose of dealing with a comparative study of two different clustering approaches (K-Means and SOM) on participants' selected features of a MOOC in the scope of the personal development. Clustering can be discovered as a useful exploratory technique for identifying and analyzing MOOC participants' profiles, a format characterized by the great diversity of enrolled people, which come from different personal and professional backgrounds, a very large range of knowledge levels, dissimilar motivations and goals, as well as many other heterogeneous issues that make more changing their clustering.

In the field of MOOC format, the understanding of participants' behaviour and the knowledge of participants' profiles are rather limited and just confined to a description of participants' features and their percentage of presence in the courses. Definitely, and according to Liyanagunawardena et al. (2013), the lack of information about MOOC participants for sure represents a challenge for researchers.

Clustering technique in this study is performed by running K-Means and SOM algorithms with a subset of variables collected from a survey with the aim of grouping the participants of a MOOC in a cohesive way. Participant's features include gender, date of birth, educational level, employment status, previous MOOC experience, goals setting and finally the role of interaction in the learning process from participants' perspective. Two aspects are addressed, firstly the clustering evaluation by applying quality criteria to the resulting clustering of K-Means (intra-cluster value and average Silhouette width) and SOM (estimated topographical accuracy and average distortion measure) and, secondly, its further interpretation in order to identify underlying relationships in the internal structure of features that make up participants' profiles, which may help designers and other policy-makers to have a deeper understanding of the diversity of participants' profiles.

The paper is structured as follows. Firstly it is briefly described Open Education movement and introduced MOOCKnowledge project. Next, K-Means and SOM techniques are proposed. Afterwards a description of KDD-based methodology is detailed. This is followed by evaluation and interpretation clustering. Finally, this paper presents the most relevant preliminary conclusions of the comparison of internal structure of both K-Means and SOM clustering and possible lines of future work are discussed.

Open Education movement

The Declaration of Paris on Open Educational Resources (OER) recommends promoting the knowledge and using of open and flexible education from a lifelong learning perspective (UNESCO, 2012), which for the Lisbon European Council represents a basic component of European social model in order to build a more inclusive, tolerant and democratic society (Commission of the European Communities, 2001). In the same way, OpenCourseWare (OCW) program initiative represents one step further and Massive Open Online Courses (MOOC) alternative provides an opportunity to access to Open Education scenario to a great number of people from any place in the world. The desire of learning without constraints leads to identify a diversity of profiles that considers people intentions, needs, motivations and goals, among others. All these features play an important role in the new educational trends and have the support of the European institutions (Commission of the European Communities, 2001), but unfortunately they have little prominence in research scope. MOOCKnowledge project, an initiative of the European Commission's Institute of Prospective Technological Studies (IPTS), aims to establish large-scale cross-provider data collection on European MOOCs to cover partially the participants' underrepresentation from a participant-centred perspective, where the diversity of the participants and the variety of profiles represent a relevant issue (Kalz et al., 2015).

Clustering techniques: K-Means and SOM algorithms

Clustering is an example of unsupervised learning that aims to find natural partitions into groups (Farias et al., 2008). This paper is focused on two clustering techniques, K-Means and its four methods (Lloyd, 1957; Forgy, 1965; MacQueen, 1967; Hartigan-Wong, 1979) as a partition-based clustering algorithm and Kohonen's Self-Organizing Maps (SOMs) as a representative technique of Artificial Neural Networks (ANNs). Clustering can be a useful exploratory technique for identifying and analyzing MOOC participants' profiles with the purpose of discovering underlying relationships in the internal structure of participants' features that could provide support for MOOC designers and other policy-makers.

K-Means takes as input parameters a set S of entities and an integer K (number of clusters), and outputs a partition of S into subsets $S_1,...,S_k$ according to the similarity of their attributes (Chen et al., 2002). The main points of interest for this paper are the four K-Means methods, the estimation of the number of clusters (K) (Jain et al., 1999) and the minimization of the total distance between the group's members and their centroids (intra-cluster distance).

SOM technique, developed by Teuvo Kohonen in 1982, is a type of Artificial Neural Network (ANN) model inspired by a kind of biological neural network (Hertz et al., 1991) and is performed to identify, classify and extract features of high-dimensional data (Deligiorgi et al., 2014). This network architecture considers on the one hand a neurons' learning network and on the other hand the training vectors (input layer) of dimension n. The elements of these two layers are fully connected and the training set is mapped into a two-dimensional lattice (Kohonen, 1989).

Methodology

This methodological proposal is based on Knowledge Discovery in Databases (KDD) system and is built up of a set of stages (Fayyad et al., 1996).

Within MOOCKnowledge project was implemented an online multilingual survey although for this paper it was only selected the one of a MOOC in the field of personal development that was offered by a Spanish higher education institution and provided by MiriadaX in the autumn of 2014. The number of enrolled population was about 10,000 and the number of fully filled out pre-questionnaires was 715. This is an opportunity for applying K-Means and SOM clustering algorithms with real-world data from hundreds, even thousands of people. This data sample was made up of the following participants' features:

- demographics (gender, age),

- Human Development Index (HDI), a summary measure in key dimensions (life expectancy, education, income) of human development (Jahan, 2015) with four levels (very high, high, medium and low),
- educational level (Pre-primary education, Primary education or first stage of Basic education, Low secondary or second stage of basic education, (Upper) secondary education, Post-secondary non-tertiary education, First stage of tertiary education, Second stage of tertiary education),
- employment status (employed for wages, self-employed, out of work and looking for work, out for work but not currently looking for employment, student, military, retired, unable to work),
- previous experience in MOOC format,
- setting of participants' goals regarding their enrolment in a MOOC (establishment of standards for assignments, establishment of short- and long-term goals, maintenance of high standards in learning, management of temporal planification, confidence in the work quality assurance),
- importance of the three kinds of interaction (learner-learner, learner-instructor and learner-content) identified by Michael Moore (1989) from participants' perspective.

The interface used in this study is RStudio Version 0.99.491 licensed under the terms of version 3 of the GNU Affero General Public License. Furthermore, R 3.2.3 GUI 1.66 Mavericks build (7060), part of the Free Software Foundation's GNU Project, is the selected environment for performing this study.

As a reflection of real-world data, it was needed an additional effort in data cleaning process for dealing with extreme outliers. Most of the fields of a set of records were empty, they was finally rejected in order to perform more consistent data exploitation. This study had mixed type data (continuous and categorical) and, consequently, standardization stage was performed. The chosen technique was to replace categorical data with binary data and to apply the Z-score standardization method for continuous data. On that point, data sample was ready for a clustering analysis with 657 resulting records.

The number of iterations running K-Means for each method was 120 times and SOM was iteratively performed 480 times. In order to evaluate the quality of K-Means clustering, it was applied an intra-cluster measure and the average Silhouette width, respectively. The chosen K-Means clustering was the one with the minimum intra-cluster value (5553.208), which matched with Hartigan-Wong's method and with K = 4. The clustering candidate had a value close to zero (0.09) for average Silhouette width criterion, which revealed it could not be ensured that all participants were properly grouped (a value close to 0 in a range value between -1 and 1), although was the highest value of all the implementations. The estimated topographical accuracy and the average distortion measure, which should be minimized and maximized respectively, were the two selected quality measures to evaluate the resulting SOM clustering, with values of 38.136 and 0.98. Both indicators were referred to what degree the topology reflects the relationships in input data (sample data). These statistics evaluated clusters without any previous knowledge related to MOOC participants' features and as result it could be chosen the local (sub)-optimal clustering and afterwards extracted the meaningful information about MOOC participants.

Measure criteria of previous stage were focused on data themselves and evaluated clusters without prior knowledge of MOOC participants. This stage, clustering interpretation, was the process that made possible the extraction of previously unknown knowledge and useful information from a subset of variables from the MOOC pre-questionnaire.

Results and discussions

Due to the heterogeneity of MOOC participants' profiles, there was no prior knowledge in advance about their number within a specific MOOC. The application of unsupervised clustering techniques allowed the selection of the best of all resulting clusters for both algorithms, which were based on the established quality criteria. These set of clusters show to what extent every participants' feature contributes in the internal structures for the identified MOOC participants' profiles by running K-Means with the method Hartigang-Wong and SOM.

The segmentation of participants into the different profiles evinced significant similarities between K-Means and SOM clustering, as is shown in Table 1. However, it would be necessary a deeper analysis in order to verify this behaviour.

Table 1: Number of participants per profile

Number of participants	Profile 1	Profile 2	Profile 3	Profile 4
K-Means	105	277	48	227
SOM	42	278	120	217

Demographic information (age and gender) and MOOC experience of participants are shown in Table 2 and Table 3, respectively. The ages of participants varied over a very fairly similar range of weights for the eight clusters. It was highlighted that the maximum age was located in K-Means, while the minimum in SOM. The weights of gender belong to women and it was noteworthy their greater presence except in S_Profile4, where the majority were men. Finally, regarding the MOOC experience of participants, only a profile, K_Profile3, had an inexplicable weight. It seemed that its participants had taken 24 MOOCs on average.

Table 2: Demographics and MOOC experience of participants for K-Means clustering

Features	K_Profile1	K_Profile2	K_Profile3	K_Profile4
Age	38	49	40	28
Gender (Female)	0,638	0,635	0,604	0,722
MOOC experience	5	5	24	8

Table 3: Demographics and MOOC experience of participants for SOM clustering

Features	S_Profile1	S_Profile2	S_Profile3	S_Profile4
Age	37	39	42	22
Gender (Female)	0,738	0,669	0,658	0,387

MOOC experience	8	5	6	6

With the purpose of making a preliminary analysis, each of features' weights that contributed to shape those eight profiles set above (Table 1) were mapped to VERY HIGH, HIGH, MEDIUM and LOW values. These new tables, called matrix of prevalence levels, are shown in Table 4, Table 6, Table 8, Table 10, Table 12 for K-Means and Table 5, Table 7, Table 9, Table 11, Table 13 for SOM.

Human Development Index (HDI) had a similar weight for both techniques, although it seemed that in SOM could prevail with the weight very high. In any case, one reason could be that these weights reflect that most participants came from countries mapped with a very high- and high-HDI index. (Table 4 and Table 5)

Table 4: Matrix of prevalence levels of participants' HDI for K-Means

Feature	K_Profile1	K_Profile2	K_Profile3	K_Profile4
HDI	HIGH	VERY HIGH	HIGH	HIGH
	MEDIUM	LOW	MEDIUM	MEDIUM
	LOW	LOW	LOW	LOW

Table 5: Matrix of prevalence levels of participants' HDI for SOM

Feature	S_Profile1	S_Profile2	S_Profile3	S_Profile4
HDI	VERY HIGH	VERY HIGH	VERY HIGH	HIGH
	LOW	MEDIUM	LOW	LOW
	LOW	LOW	LOW	LOW

Among the elements for the feature educational level of a participant, the only one with a predominant weight was *Second stage of tertiary education* for both clustering. This variable had a high or very high prevalence weight for all profiles except for one on SOM clustering. (Table 6 and Table 7)

Table 6: Matrix of prevalence levels of participants' educational level for K-Means

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
Educational Level	Pre-primary education	LOW	LOW	LOW	LOW
	Primary education or first stage of basic education	LOW	LOW	LOW	LOW
	Lower secondary or second stage of basic education	LOW	LOW	LOW	LOW
	(Upper) secondary education	LOW	LOW	LOW	LOW
	Post-secondary non-tertiary education	LOW	LOW	LOW	LOW
	First stage of tertiary education	LOW	LOW	LOW	LOW
	Second stage of tertiary education	HIGH	HIGH	HIGH	HIGH

Table 7: Matrix of prevalence levels of participants' educational level for SOM

Feature		S_Profile1	S_Profile2	S_Profile3	S_Profile 4
Educational level	Pre-primary education	LOW	LOW	LOW	LOW
	Primary education or first stage of basic education	LOW	LOW	LOW	LOW
	Lower secondary or second stage of basic education	LOW	LOW	LOW	LOW
	(Upper) secondary education	LOW	LOW	LOW	LOW
	Post-secondary non-tertiary education	LOW	LOW	LOW	LOW
	First stage of tertiary education	MEDIUM	LOW	LOW	LOW
	Second stage of tertiary education	VERY HIGH	VERY HIGH	VERY HIGH	MEDIUM

The elements student and employed for wages had high prevalence on K_Profile4 and K_Profile2 respectively. It stood out that it could characterize young students on K_Profile4 a high student's weight combined with the fact that the average age was 28 years, although it would be needed further analysis in order to verify this hypothesis. K_Profile2 showed the same circumstance with the element employed for wages and the average age 49 years that could characterize middle age employed people. The weight of element employed for wages on SOM was not as prevalent as on K-Means, although had a certain prevalence on every profile. (Table 8 and Table 9)

Table 8: Matrix of prevalence levels of participants' employment status for K-Means

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
Employment status	Homemaker	LOW	LOW	LOW	LOW
	Student	LOW	LOW	LOW	HIGH
	Employed for wages	MEDIUM	HIGH	MEDIUM	LOW
	Out of work and looping for work	MEDIUM	MEDIUM	LOW	LOW
	Out of work but not currently looking for wages	LOW	LOW	LOW	LOW
	Retired	LOW	LOW	LOW	LOW
	Self-employed	LOW	LOW	LOW	LOW
	Unable to work	LOW	LOW	LOW	LOW

Table 9: Matrix of prevalence levels of participants' employment status for SOM

Feature		S_Profile 1	S_Profile 2	S_Profile 3	S_Profile 4
Employment status	Homemaker	LOW	LOW	LOW	LOW
	Student	MEDIUM	MEDIUM	MEDIUM	LOW
	Employed for wages	MEDIUM	MEDIUM	MEDIUM	MEDIUM
	Out of work and looping for work	MEDIUM	MEDIUM	MEDIUM	LOW
	Out of work but not currently looking for wages	LOW	LOW	LOW	LOW
	Retired	LOW	LOW	LOW	LOW
	Self-employed	LOW	LOW	LOW	LOW
	Unable to work	LOW	LOW	LOW	LOW

One of the most interesting features for this study was the setting of participant's goals because of its specific distribution of the weights on every cluster. K-Means preserved the same prevalence in each of the profiles, although K-Profile1 attracted the attention with its very high weight to all and each of the five elements. SOM had a quasi-identical circumstance in terms of profiles' behaviour, although participants that belonged to S_Profile1 gave a high prevalence to the element participant's confidence in the quality assurance of their work. Therefore, this feature should be analyzed in a more detailed way. (Table 10 and Table 11)

Table 10: Matrix of prevalence levels of participants' goals for K-Means

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
Goals setting.	Standards establishment	VERY HIGH	MEDIUM	MEDIUM	MEDIUM
	Short- and long-term goals establishment	VERY HIGH	MEDIUM	MEDIUM	MEDIUM
	High standards maintenance	VERY HIGH	MEDIUM	MEDIUM	MEDIUM
	Temporal planification management	VERY HIGH	MEDIUM	MEDIUM	MEDIUM
	Confidence in work quality assurance	VERY HIGH	MEDIUM	MEDIUM	MEDIUM

Table 11: Matrix of prevalence levels of participants' goals for SOM

Feature		S_Profile 1	S_Profile 2	S_Profile 3	S_Profile 4
Goals setting.	Standards establishment	MEDIUM	MEDIUM	MEDIUM	LOW
	Short- and long-term goals establishment	MEDIUM	MEDIUM	MEDIUM	LOW
	High standards maintenance	MEDIUM	MEDIUM	MEDIUM	LOW
	Temporal planification management	MEDIUM	MEDIUM	MEDIUM	LOW
	Confidence in work quality assurance	HIGH	MEDIUM	MEDIUM	LOW

Focused on interaction feature, on K-Means clustering the range of weights took values from very high to medium. Learner-Content interaction was the element with a very high prevalence on K_Profile1, Learner-Learner interaction was the least representative interaction for the eight clusters and, finally, Learner-Teacher interaction did not show such a regular behaviour as the other two characteristics described above. On SOM the range of weights was from high to low and Learner-Content interaction was depicted with a greater weights. Undoubtedly, the three interactions played their role in each and every one of the profiles, even on those where the prevalence was low, and also a deeper analysis should be accomplished. (Table 12 and Table 13)

Table 12: Matrix of prevalence levels of types of interactions of participants for K-Means

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
Interactions	Learner-Learner	MEDIUM	MEDIUM	MEDIUM	MEDIUM
	Learner-Content	VERY HIGH	HIGH	HIGH	MEDIUM
	Learner-Teacher	HIGH	MEDIUM	MEDIUM	MEDIUM

Table 13: Matrix of prevalence levels of types of interactions of participants for SOM

Feature		S_Profile 1	S_Profile 2	S_Profile 3	S_Profile 4
Interactions	Learner-Learner	MEDIUM	MEDIUM	MEDIUM	LOW
	Learner-Content	HIGH	HIGH	HIGH	MEDIUM
	Learner-Teacher	HIGH	MEDIUM	HIGH	LOW

In conclusion, the results brings to light that it is not possible to determine the best clustering without additional analysis.

Conclusions

In this study it was chosen two types of algorithms from two different approaches, a partitional clustering algorithm and an artificial neural network. The comparison of K-Means and SOM was performed with the aim of finding out which of them fitted better. These clustering techniques were applied under some specific conditions to a better understanding of MOOC participants' subset of features and might represent a way of discovering the intrinsic structures within the data sample and, consequently, designers and other policy-makers could also have a deeper understanding of the diversity of participants' profiles. It should be emphasized that the role played by experts in MOOC format has a critical subjective component and their relevance is even greater because clustering result is largely influenced by data sample, the selected variables and the used clustering algorithm.

A more realistic understanding of people profiles is a step forward for many disciplines that call for a more in-depth knowledge of their customers and Open Education is no exception. Therefore, future work in the short to medium term involves a deeper research of clustering techniques, especially both evaluation and interpretation of clustering, with the involvement of the whole data collection of MOOCKnowledge project.

References

1. Brachman, R. J., & Anand, T. (1994). *The Process of Knowledge Discovery in Databases: A First Sketch.* AAAI Technical Report WS-94-03. Atlanta: AT&T Bell Laboratories.
2. Chen, G., Jaradat, S., Banerjee, N., Tanaka, T., Ko, M., & Zhang, M. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica, 12*, 241-262.
3. Commission of the European Communities (2001). *Making a European Area of Lifelong Learning a Reality.* COM(2001) 678 final, Brussels.
4. Deligiorgi, D., Philippopoulos, K., & Kouroupetroglou, G. (2014). An Assessment of Self-Organizing Maps and k-means Clustering Approaches for Atmospheric Circulation Classification. In V. S. Bulucea (Ed.), *Proceedings of the 2014 International Conference on Environmental. Recent Advances in Environmental Science and Geoscience* (pp. 17-23). Venice.
5. Farias, R., Durán, E., & Figueroa, S. (2008). *Las Técnicas de Clustering en la Personalización de Sistemas de e-Learning.* XIV Congreso Argentino de Ciencias de la Computación (CACIC).
6. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, 17*(3), 37-54. Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131
7. Forgy, E. W. (1965). Clustering analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics, 21*, 768-769.
8. Hartigal, J., & Wong, M. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 28*(1), 100-108.
9. Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the Theory of Neural Computation.* Reading: Addison-Wesley Longman.
10. Jahan, S. (2015). *Human Development report 2015. Work for Human Development.* United Nations Development Programme (UNDP), New York.
11. Jain, A., Murty, M., & Flynn, P. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*(3), 264-323.

12. Kalz, M., Kreijns, K., Wahlout, J., Castaño-Muñoz, J., Espasa, A., & Tovar, E. (2015). Setting-up a European Cross-Provider Data Collection on Open Online Courses. *The International Review of Research in Open and Distributed Learning, 16*(6), 62-77. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/2150

13. Kohonen, T. (1989). *Self-Organization and Associate Memory* (3rd ed.). New York: Springer-Verlag.

14. Liyanagunawardena, T., Adams, A., & Williams, S. (2013). MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distance Learning, 14*(3), 202-227. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1455

15. Lloyd, S. P. (1957). *Least squares quantization in PCM.* Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory, 28, 128–137.

16. MacQueen, J. (1967). Some methods for classification and analysis of multivariante observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297. Berkeley: University of California Press.

17. Moore, M. (1989). Editorial: Three Types of Interaction. *The American Journal of Distance Education, 3*(2), 1-7.

18. UNESCO. (2012). *2012 Paris OER Declaration.* 2012 World Open Educational Resources (OER) Congress UNESCO. Paris.