A Benchmarking Study of K-Means and Kohonen Self-Organizing Maps Applied to 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. Their heterogeneity represents an opportunity to identify the underlying relationships present in the internal structure of the data. This paper aims at identifying and analysing MOOC participants' profiles by running two unsupervised clustering techniques: K-Means as a partitioning approach and Kohonen's Self-Organizing Maps (SOM) as a competitive learning technique.

The dataset comes from MOOCKnowledge project data collection. After the execution of both clustering algorithms, the evaluation stage is performed with the validation measures: an intra-cluster measure and an overall quality criterion for K-Means, and two measures related to topological ordering for SOM.

The interpretation of the resulting profiles is made with the help of a matrix of prevalence levels. The similarities between the two resulting clustering on the one hand, and some pinpointed differences on the other are highlighted. They cannot be evaluated in advance without the opinion of an expert familiarized with the specifications of the MOOC.

After a preliminary study the results are not considered conclusive. For sure there is a long way in order to help stakeholders on how to identify and select the appropriate clustering according to several quality criteria.

Abstract in Spanish

Este artículo tiene como objetivo la identificación y el análisis de un conjunto de perfiles de participantes MOOC con la aplicación de dos técnicas de agrupamiento no supervisadas: K-Means como algoritmo particional y Kohonen Self-Organizing Maps (SOM) como una técnica representativa de aprendizaje competitivo.

El conjunto de datos del estudio tiene su origen en el proyecto MOOCKnowledge. Tras la etapa de ejecución, la evaluación del

agrupamiento se realiza con una selección de medidas de validación: una de dentro de cada grupo (intra-cluster) y la segunda de calidad global del agrupamiento (average Silhouette width) para K-Means, así como dos medidas relacionadas con el orden topológico para SOM.

La interpretación de los perfiles resultantes de los dos agrupamientos se realiza con la ayuda de una matriz de niveles de prevalencia. Tanto las similitudes como las diferencias identificadas en las matrices no pueden evaluarse de antemano sin la opinión de un experto familiarizado con el formato MOOC.

Tras un estudio preliminar los resultados no se consideran concluyentes. Sin duda, hay todavía un largo camino para ayudar a las partes interesadas sobre cómo identificar y seleccionar la agrupación adecuada de acuerdo con varios criterios de calidad.

Keywords: MOOC profiles, K-Means, Kohonen's Self-Organizing Maps, SOM, cluster analysis, clustering

Introduction

This paper has the final purpose of dealing with a comparative study of two clustering approaches (K-Means and SOM) on a subset of participants' features of a MOOC in the scope of the personal development. According to this study, clustering can be discovered as a useful exploratory technique for identifying and analysing MOOC participants' profiles, a format characterized by a great diversity of enrolled people that 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 challenging their clustering.

In the field of MOOC format 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, Adams, & Williams, 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 variables include gender, date of birth, nationality, educational level, employment status, previous MOOC experience, goals setting and, finally, the role of interaction in the learning process. 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, their further interpretation with a view to identifying underlying relationships in the internal structure of the features that make up participants' profiles. In conclusion, clustering results 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. Then evaluation and interpretation clustering are outlined. 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 Open Education research. 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, where their diversity and 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 (Farías, Durán, & Figueroa, 2008). This paper is focused on two clustering techniques, K-Means and its four methods (Lloyd, 1982; 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) (Kohonen, 1989). Clustering can be a useful exploratory technique for identifying and analysing 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, Murty, & Flynn, 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, Krogh, & Palmer, 1991) and is performed to identify, classify and extract features of high-dimensional data (Deligiorgi, Philippopoulos, & Kouroupetroglou, 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, which is built up of a set of stages (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Figure 1 shows the main stages of KDD methodology, especially Data Mining and its three main grouped tasks (execution, analysis and model refining) regarding clustering model. The execution task is focused on K-Means and SOM clustering approaches, analysis is in charge of fitting the parameterization of clustering algorithms and model refining is committed to the experiment improvement. Then, the next stage (clustering evaluation) is involved in the validation of quality measures and subsequent the interpretation of the (sub-) optimal clustering. Finally, a matrix of prevalence levels for

each feature reflects the weights for both K-Means and SOM approaches. Previously, two interrelated stages have been carried out: Data cleaning and Data transformation. Their tasks are extremely important by tackling the initial dataset to input data to clustering stage.

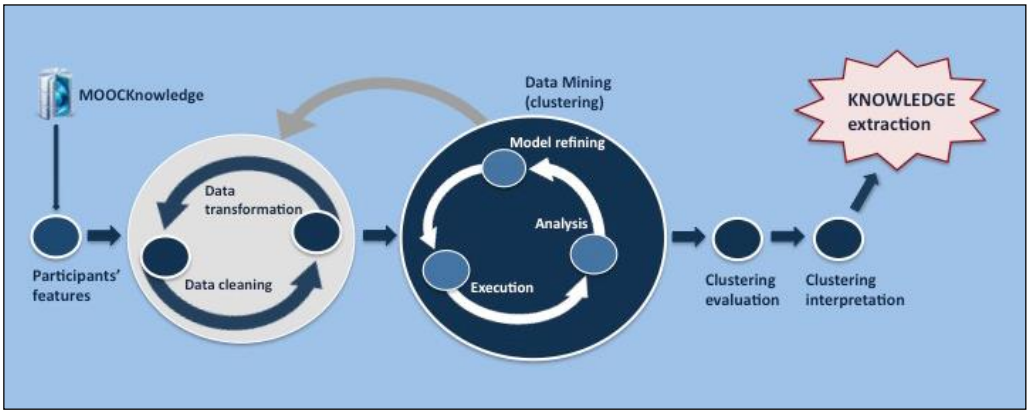

Figure 1. Stages of Knowledge Discovery in Databases (KDD) Process

MOOCKnowledge project conducted an online multilingual survey although for this paper it was only selected 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 undoubtedly an opportunity for applying K-Means and SOM clustering algorithms with real-world data from hundreds, even thousands of people. Finally, the data sample was made up of the following participants' features:

- Demographics (gender, date of birth is transformed into a new variable identified as age).
- Nationality is transformed into a new variable based on Human Development Index (HDI). This indicator represents a summary measure in three key dimensions (life expectancy, education, income) of human development with four levels (very high, high, medium and low) (Selim, 2015).
- 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 of 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 types of interaction (learner-learner, learner-instructor and learner-content) identified by (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 (RStudio Team, 2015).

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 so they were finally rejected in order to deal with a more consistent data exploitation. This study had mixed data types (continuous and categorical) and, consequently, standardization and transformation tasks were performed. The chosen techniques were to apply the Z-score standardization method for continuous data and replace categorical data with binary 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 (5,553.208), which matched with Hartigan-Wong's method and $K = 4$. 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) (Nguyen & Rayward-Smith, 2008), 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) (Wehrens & Buydens, 2007). 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 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 knowledge in advance about their number within the specific MOOC of this study. The application of unsupervised clustering techniques allowed the selection of the best of all the resulting clustering from both algorithms with the help of the established quality criteria. All those clusters show to what extent each of the participants' features contributes to the internal structure for the identified MOOC participants' profiles by running K-Means with the method Hartigan-Wong and SOM.

An overview into the different profiles evinced significant similarities between K-Means and SOM approaches in the number of participants, as is shown in Table 1. However, it would be necessary a deeper analysis of the features that comprise the different clusters in order to verify this first impression.

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

The demographic information (age and gender) and the 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 was 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 at first sight. It could seem that its participants had taken a significant number of courses although, of course, a deeper analysis is required.

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

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. HDI values are shown in 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

For 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 6, Table 8, Table 10, Table 12 for K-Means and Table 7, Table 9, Table 11, Table 13 for SOM.

Among the items that make up the educational level feature 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. Participants' educational level values are shown in 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_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	MEDIUM	LOW	LOW	LOW
	Second stage of tertiary education	VERY HIGH	VERY HIGH	VERY HIGH	MEDIUM

The items student and employed for wages had high prevalence in K-Means 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 wage in SOM has similar values for the four profiles. The element out of work and looking for work had the same weight for six of the eight profiles, except for a low value and a medium value for K_Profile3 and S_Profile3, respectively. Participants' employment status values are shown in 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 looking 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_Profile1	S_Profile2	S_Profile3	S_Profile4
employment status	homemaker	LOW	LOW	LOW	LOW
	student	MEDIUM	MEDIUM	MEDIUM	LOW
	employed for wages	MEDIUM	MEDIUM	MEDIUM	MEDIUM
	out of work and looking 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 all their weights had equal or lower prevalence. Participants that belonged to S_Profile1 had the highest prevalence with a high value to the element participant's confidence in the quality assurance of their work. The rest of the weights were uniform in each of the profiles. Therefore, this feature should be analysed in a more detailed way. Participants' goals values are shown in 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_Profile1	S_Profile2	S_Profile3	S_Profile4
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 the three types of interaction, on K-Means clustering the range of weights took values from very high to medium. On the other hand, the range of weights in SOM was from high to low. In K-Means, Learner-Content interaction was the element with a very high prevalence on K_Profile1 and a high value for K_Profile2 and K_Profile3. Learner-content interaction in SOM was depicted with a high weight except for a medium prevalence for S_Profile4. 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. 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. The values for the three types of interactions are shown in 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
interaction	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_Profile1	S_Profile2	S_Profile3	S_Profile4
interaction	learner-learner	MEDIUM	MEDIUM	MEDIUM	LOW
	learner-content	HIGH	HIGH	HIGH	MEDIUM
	learner-teacher	HIGH	MEDIUM	HIGH	LOW

The above comparative of participants' features does not allow a generalization of the partial results to the whole data collection because this study represents a preliminary stage that requires both an additional analysis of resulting clustering and the help of an expert that guides and contextualizes the interpretation process for both approaches (K-Means and SOM) and finally determines which one is closer to the real picture of MOOC participants.

In conclusion, the results bring 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 (Lloyd, Forgy, MacQueen, and Hartigan-Wong) 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 an enhanced understanding of a subset of features of participants in a MOOC in the field of personal development. They definitely might represent a way of discovering the intrinsic structure of data sample and, consequently, designers and other policy-makers could also have a deeper knowledge 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 the results of clustering are largely influenced by data sample, the selected variables and the clustering algorithm used.

As conclusion, therefore, it can be said that the results bring to light that it is not possible to determine which one is the best clustering (K-Means and SOM) without an additional analysis where the role of MOOC experts is more than relevant.

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 and KDD methodology, especially evaluation and interpretation clustering stages, with the involvement of the whole MOOCKnowledge data collection.

References

1. 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. Retrieved from http://www3.stat.sinica.edu.tw/statistica/oldpdf/A12n112.pdf
2. Commission of the European Communities. (2001). *Making a European Area of Lifelong Learning a Reality* (No. COM(2001) 678 final). Brussels.
3. Deligiorgi, D., Philippopoulos, K., & Kouroupetroglou, G. (2014). An Assessment of Self-Organizing Maps and k-means Clustering Approaches for Atmospheric

Circulation Classification. *Proceedings of the 2014 International Conference on Environmental, Venice, Italy,* Vol. 17, 17–23.

4. Farías, R., Durán, E. B., & Figueroa, S. G. (2008). *Las Técnicas de Clustering en la Personalización de Sistemas de e-Learning.* Paper presented at the XIV Congreso Argentino de Ciencias de la Computación. Retrieved from http://sedici.unlp.edu.ar/bitstream/handle/10915/21990/Documento_completo.pdf?sequence=1

5. 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 http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf

6. Forgy, E. W. (1965). Clustering analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics, 21*, 768–769.

7. Hartigan, J. A. & Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100–108. Retrieved from http://www.labri.fr/perso/bpinaud/userfiles/downloads/hartigan_1979_kmeans.pdf

8. Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Santa Fe Institute Studies in the Sciences of Complexity lecture notes. Redwood City, CA: Addison-Wesley Longman: Basic Books. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.35.978&rep=rep1&type=pdf

9. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys (CSUR), 31*(3), 264–323. Retrieved from https://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf

10. 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. *International Review of Research in Open and Distributed Learning, 16*(6), 62–77. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/2150/3553

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

12. Liyanagunawardena, T., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning, 14*(3), 202–227. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1455/2602

13. Lloyd, S. P. (1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. *EEE Transactions on Information Theory, 28*(2), 128–137. Retrieved from http://www.cs.nyu.edu/~roweis/csc2515-2006/readings/lloyd57.pdf

14. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, Calif.*, 281–297. Retrieved from https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992

15. Moore, M. (1989). Three Types of Interaction. *American Journal of Distance Education, 3*(2), 1–7. Retrieved from http://aris.teluq.uquebec.ca/portals/598/t3_moore1989.pdf

16. Nguyen, Q. H., & Rayward-Smith, V. J. (2008). Internal quality measures for clustering in metric spaces. *International Journal of Business Intelligence and Data Mining, 3*(1), 4–29. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.205.5025

17. RStudio Team. (2015). *RStudio: Integrated Development for R.* Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com

18. Selim, J. (2015). *Human Development report 2015.* Work for Human Development. selim - Google Académico. New York, NY: United Nations Development Programme (UNDP). Retrieved from http://hdr.undp.org/sites/default/files/2015_human_development_report_1.pdf

19. UNESCO. (2012). *2012 PARIS OER DECLARATION.* Presented at the 2012 WORLD OPEN EDUCATIONAL RESOURCES (OER) CONGRESS, Paris.

20. Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software, 21*(5), 1–19. Retrieved from http://www.jstatsoft.org/v21/i05/paper

Acknowledgement

This work has been partially funded by a tender (JRC/SVQ/2013/J.3/0035/NC) of the European Commission's Institute for Prospective Technological Studies (IPTS) and by Regional Government of Madrid (eMadrid S2013/ICE-2715).