ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

New Orleans JAZZED about Engineering Education

Paper ID #16718

A Benchmarking Study of Clustering Techniques Applied to a Set of Characteristics of MOOC Participants

Ms. Rosa Cabedo, Universidad Politecnica de Madrid

Rosa Cabedo is Engineer in Computer Science and currently Ph.D. Student at Technical University of Madrid (Spain) in the field of Open Education. The final purpose of her research is the identification and analysis of the profiles of language MOOC participants and the features of language learning (interaction, feedback, evaluation, certification) in order to adequate the design to MOOC format to facilitate the linguistic and communicativa competences improvement of LMOOC participants and their professional development. Her research focuses on the analysis of the heterogeneity of (L)MOOCs participants with the help of clustering techniques.

Dr. Tovar Caro Edmundo, Universidad Politecnica de Madrid

Edmundo Tovar, computer engineering educator, has a Ph.D. (1994) and a bachelor's degree (1986) in computer engineering from the Universidad Politécnica de Madrid (UPM). He is a certified Software Development Professional (CSDP) from the IEEE Computer Society. He is Associate Dean for Quality and Strategic Planning in the Computing School of the Universidad Politécnica de Madrid. From this last position, he is in charge of the training for academic staff, the introduction of innovative solutions including new pedagogies, new approaches that improve student learning of technical skills and cultural skills, improved methods of blended learning, and others. He works in the open educational resources area. He is leader of an Innovation Group in Education in the UPM. He is Executive Director of the OCW UPM Office and an elected member of the Board of Directors of the OpenCourseWare Consortium. He is author of many papers in engineering education, a member of the Steering Committee of and Coc-hair for Europe of Frontiers Education Conference (FIE), and member of the IEEE RITA Editorial Committee. He is IEEE Senior Member, Past Chairman of the Spanish Chapter, and, as member of the Board of Governors of the IEEE Education Society, he is currently Chair of the Distinguished Lectures Program for the IEEE Education Society.

Prof. Manuel Castro, Universidad Nacional de Educacion a Distancia

Manuel Castro, Electrical and Computer Engineering educator in the Spanish University for Distance Education (UNED) has a doctoral industrial engineering degree from the ETSII/UPM. Full Professor of Electronics Technology inside the Electrical and Computer Engineering Department. He is Head of Department of Electrical and Computer Engineering at UNED. Was co-chair of the conference FIE 2014 (Frontiers in Education Conference) organized in Madrid, Spain, by the IEEE and the ASEE, and will co-chair REV 2016 (Remote engineering and Virtual Instrumentations) in Madrid, Spain. He is Fellow member of IEEE (for contributions to distance learning in electrical and computer engineering education) and member of the Board of Governors (BoG) (2005–2018) of the IEEE Education Society, President (2013-2014) and Jr Past-President (2015-2016) of the IEEE Education Society; Founder and Past-Chairman (2004-2006) of the Spanish Chapter of the IEEE Education Society, Past-Chair of the IEEE Spain Section (2010-2011) and IEEE Region 8 Educational Activities Subcommittee Chair. He has been awarded with the 2012 TAEE (Technologies Applied to Electronic Education) Professional Career Award, IEEE EDUCON 2011 Meritorious Service Award (jointly with Edmundo Tovar) of the EDUCON 2011 conference; 2010 Distinguished Member Award of the IEEE Education Society; 2009 Edwin C. Jones, Jr. Meritorious Service Award of the IEEE Education Society; with the 2006 Distinguished Chapter Leadership Award and for the collective work inside the Spanish Chapter of the IEEE Education Society with the 2011 Best Chapter Award (by the IEEE Region 8) and with the 2007 Chapter Achievement Award (by the IEEE Education Society). He is Member of the Board of the Spanish International Solar Energy Society (ISES).

A benchmarking study of clustering techniques applied to a set of features of MOOC participants

Abstract
Massive Open Online Courses (MOOC) format is characterized by the great diversity of enrolled people. Moreover, the lack of prior knowledge of their profiles constitutes an important barrier with a view to identifying and getting a better understanding of underlying relationships in the internal structure of the features that make up the profile of the participants in those courses. This paper has the aim of identifying and analyzing the feasible set of MOOC participants' profiles by running 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 for this paper comes from the MOOCKnowledge project data collection, which provides an opportunity to work with real-world data from hundreds of people. K-Means and SOM algorithms are performed with a subset of participants' features as input data. The clustering evaluation, meanwhile, is achieved with a selection of indices, an intra-cluster measure and an overall quality criterion for K-Means, and two measures related to topological ordering for SOM.

The comparison of internal structure of both clustering (set of profiles) shows that there are similarities between them on the one hand and some pinpointed differences that can not be evaluated in advance without the opinion of an expert familiarized with the specifications of the MOOC on the other.

Therefore, this comparison can not be considered conclusive until after a preliminary study of the results of the clustering interpretation for both algorithms. Finally, although it is not determined the clustering that best fits between K-Means and SOM, this study might help to provide a methodological guide on how to identify and select the appropriate clustering according to several quality criteria.

Key Words
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 different clustering approaches (K-Means and SOM) on a selected set of participants' features of a MOOC in the scope of the personal development. With this study, clustering could be discovered as a useful exploratory technique for identifying and analyzing MOOC participants' profiles, a format characterized by the great diversity of enrolled people. The heterogeneity of the population has its origin in different personal and professional backgrounds, a range of knowledge levels very large, dissimilar motivations and goals, as well as many other different issues that make more challenging a clustering of MOOC participants.

Clustering and patterns recognition is a technique applied in many disciplines, such as customer segmentation in marketing, medicine or engineering. In the field of MOOC format, the understanding of participants' behavior and their degree of engagement with resources are examples of recognition of patterns. However, the knowledge of participants' profiles is rather limited and is just confined to a description of participants' features and their percentage of presence in the courses. Definitely, and according to Liyanagunawardena[1], the lack of information about MOOC participants for sure represents an open line of research.

Clustering technique in this study is performed by running K-Means and SOM 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, the goals setting process and the role of interaction in their learning process. The paper addresses two aspects, firstly the clustering evaluation by applying quality criteria of both K-Means and SOMs algorithms and, secondly, their further interpretation in order to identify underlying relationships in the internal structure of features that make up the participants' profiles. The evaluation of K-Means clustering is performed with an internal validity criterion and a mixed measure. Similarly, SOM is carried out with the value of the estimated topographical accuracy and the average distortion measure. The clustering interpretation facilitates the identification of underlying relationships in the internal structure of participants' features that may help designers and other policy-makers to reach 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 the MOOCKnowledge project. Next, K-Means and SOM techniques are proposed, followed by a comparison of both approaches. Afterwards a description of KDD-based methodology is detailed, which also includes the stages of 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[2], 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[3]. In the same way, OpenCourseWare (OCW) program initiative represents one step further, since it is focused on the inclusion of OERs in educational activities[4]. MOOC alternative also provides an excellent opportunity to access to Open Education scenario to a great number of people from any place in the world, a phenomenon that attracts once again the attention of scientific and educational community through OERs. The desire of learning without demographic, geographical and socioeconomic constraints leads to identify a diversity of profiles that considers, in addition to these set of specific features that characterize potential participants, their intentions, needs, motivations and goals, among others. All these features play an important role in the new educational trends, belong or not to formal education, and have the support of the European institutions[3].

MOOC participants' perspective, and specifically the set of their profiles, has little prominence in research on MOOC format. 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 their perspective, where the diversity of the participants and the variety of their profiles represent a relevant issue[5].

Clustering techniques
The data size is increased day by day and researchers are overwhelmed with mountains of data somewhat disconcerting when are viewed as a whole. A wide range of data processing techniques, including clustering, have been developed with the purpose of a more meaningful data management and a subsequent process by making sense of them. Clustering is an example of unsupervised learning which aims to find natural partitions into groups[6], an automatic grouping of coherent data subsets without the help of a response variable.

This paper is focused on two clustering techniques, K-Means and its four methods (Lloyd[7], Forgy[8], MacQueen[9], Hartigan-Wong[10]) as a partitional clustering algorithm and Kohonen's Self-Organizing Maps (SOMs) as a representative technique of Artificial Neural Networks (ANNs).

Clustering could be discovered as a useful exploratory technique for identifying and analyzing MOOC participants' profiles, a format characterized by the great diversity of enrolled people that come from different personal and professional backgrounds, have a range of knowledge levels very large, with dissimilar motivations and goals, as well as many other heterogeneous issues that make more challenging the clustering process of MOOC participants. The identification of underlying relationships in this internal structure of participants' features might help designers to identify the trully defining features that impact in a decisive way on MOOC design.

In almost every disciplines clustering is showed as a representative technique by exploring the features of data collections. Some common applications are market segmentation in order to offer a better service to customers, analysis of social networks by grouping their users, or fields such as spatial data analysis, image processing, medical analysis, economics, bioinformatics oder biometrics, and so on[6]. In the field of MOOC format, it is highlighted some clustering applications such as the recognition of patterns by grouping features of MOOC participants in order to have a better understanding of their behavior[11,12,13] or the identification of engagement patterns in videos and assessment[14].

SOMs are applied to different fields such as census data[15], purchase transactions of a company[15], customer segmentation profiles[16,15], language recognition with the study of specific patterns from bilingual speakers[17], classification of species, and many other disciplines including medicine, biology, image classification, speech recognition, computer science, insurance, among others[18,19].

K-Means algorithm
K-Means is a partition-based clustering algorithm that 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[20]. Although there are several different variations and optimizations of K-Means algorithm[21], this paper is focused on its four methods (Lloyd, Forgy, MacQueen and Hartigan-Wong).

The estimation of the number of clusters in a data collection represents a tricky process for partitional algorithms as K-Means[22] and the way of choosing K parameter is often a somewhat misleading process. In order to take that decision, diverse methods are available such as the most common ones, by hand and the elbow method, or even the proposed by Hartigan.

The iterative implementation of K-Means pursues to maximize the distances between clusters (inter-cluster distance) and minimize the total distance between the group's members and their centroids (intra-cluster distance). In other words, the resulting K groups are expected to have great similarity within each group but little similarity (dissimilarity) across groups[19].

Self-Organizing Maps (SOMs)
The Self-Organizing Maps technique, developed by Teuvo Kohonen in 1982, is a type of Artificial Neural Network (ANN) model, called Kohonen Neural Network, and is inspired by a kind of biological neural network[23]. From a philosophical perspective, it could be highlighted that ANNs might seem the brain, and imitate its innate ability to build topological maps from external information.

SOM is performed to identify, classify and extract features of high-dimensional data[24]. This network architecture (Figure 1) considers on the one hand a neurons' learning network and on the other 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. SOM is implemented iteratively so that different areas of the lattice have similar reactions to certain input layer and finally input similarities[25] are extracted and represented as the end point of the process[24,16].

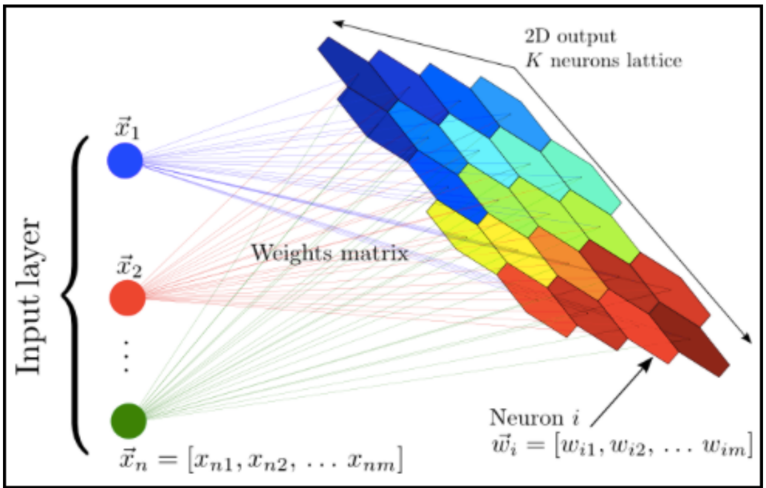

Figure 1 A schematic representation of a Self-Organizing Map[26]

Comparison of K-Means and SOM algorithms
There are no very full conclusions in comparative research of SOM and K-Means approaches. Some authors affirm that the performance of K-Means outperforms SOM[27], whereas others state exactly the opposite[21]. It is also possible to find studies where both algorithms outperform equally well[28]. It is important to highlight that all of them are addressed with different both quality measures and datasets.

A collection of studies focused on comparison between K-Means and SOM algorithms have been detailed such as applications in the scope of image segmentation[19] or atmospheric circulation classification with very similar results between both approaches[24].

Methodology
The methodological proposal is based on Knowledge Discovery in Databases (KDD) system, whose workflow is shown in Figure 2. Firstly, the study goals were set with the support of the problem analysis, the MOOC selection, as well as the software tool used. Then the preparation of data (data cleaning stage) was focused on outliers and missing data. The

standardization, identification and processing of the types of variables were also addressed. Afterwards K-Means and SOM algorithms were performed, and the evaluation of quality clustering hereafter was carried out by applying quality criteria within both algorithms. Finally, the data-driven discovery stage[29,30] was represented by the comparison between the set of clusters for both K-Means and SOM techniques.

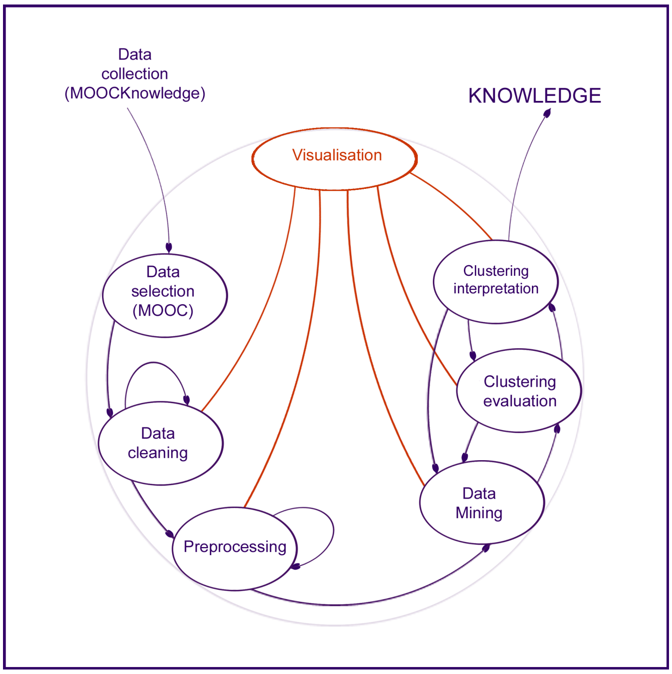

Figure 2 A methodological approach based on Knowledge Discovery in Databases (KDD) system

Problem analysis
MOOCKnowledge project has the purpose of building a large scale data collection that provides information related to profiles, experiences and behaviors of (European) MOOCs participants from an European perspective, as well as analyzing the Open Education impact of participants' subgroups such as those with a specific cultural background[5]. The implemented online multilingual survey, comprised by a pre- and post-questionnaire, was expected to reflect the high level of heterogeneity of MOOC participants' profiles, although for this paper it was only selected the survey of an isolated course.

Data selection
This diversity of MOOC participants represents an opportunity of applying K-Means and SOM clustering algorithms with real-world data from hundreds, even thousands of people. The selected data sample for this paper came from MOOCKnowledge data collection, 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. According to response rate, the amount of participants that accessed voluntary to the survey was 13% and it was completed by 7%.

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[31] with four levels (very high, high, medium, 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 of work but not currently looking for employment, student, militay, retired, unable to work),
- previous experience in MOOC format,
- setting of participants' goals regarding their enrollment 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, from a participants' perspective, of the three types of interaction (learner-learner, learner-instructor, learner-content) identified by Michael Moore[32].

Materials

The interface used is RStudio Version 0.99.491 licenced 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.

Data cleaning

The dataset for this study was a reflection of real-world data, so in order to a successful KDD, it was needed an arduous effort in the data cleaning process. Data cleaning seeks an unified logical view of databases with issues such as encouraging a single naming convention or provision of strategies for data handling such as outliers or missing data[30]. This stage included to deal with extrem outliers and in order to reduce their impact, they adopted a new value (the statistical average) because of clustering analysis is very sensitive to their presence. Most of the fields of a set of records were empty. They were finally rejected in order to perform a more consistent data exploitation.

Preprocessing

Standardization of variables aims to provide a common value range so that all the features of the data sample have the same impact on the clustering process, so it is recommended the standardization of data sample variables before starting the clustering process. This study had mixed type data (continuous and categorical) and, consequently, standardization stage was performed. The technique chosen was to replace categorical data with binary data and apply the Z-score standardization method for continuous data. On that point, data sample was ready for a clustering analysis.

The size of the data sample was an important issue. Jain et al. considered a small size a collection with fewer than 200 objects or individuals[22]. Therefore, and according to Jain et al., the resulting 657 records after cleaning and pre-processing stages should not be initially considered a sample of small size.

Data Mining (clustering)

The number of iterations running K-Means for each method was 120 times and SOM was iteratively performed 480 times. It is emphasized that the choice of K-Means method and the number of clusters (K) was made on the basis of clustering quality criteria.

Data Mining and next two stages, clustering evaluation and clustering interpretation, complemented each other. The workflow, before extracting useful knowledge from underlying data structure, went forward and back as often as it was needed.

Clustering evaluation

The evaluation of a clustering, in other words, the evaluation of the quality of the resulting clusters, faces a significant barrier. And besides, it is not a simple task to find an algorithm-independent quality measure[20].

In this study it was applied an internal validity criterion (intra-cluster measure) and a mixed measure (the average Silhouette width) in order to evaluate the quality of K-Means clustering. As is showed in Figure 3 and Figure 4, the minimization of the intra-cluster measure always involves the maximization of inter-cluster measure. In short, as clustering quality declines, intra-cluster measures tend to increase while inter-cluster measures have the opposite trend[33]. The second measure used to evaluate the clustering was a mixed measure, a combination of inter- and intra-cluster measures, named Silhouette width index, whose average reflects the overall quality of the result of clustering[20]. Average Silhouette width can be used as a single index for the clustering's quality in order to reflect the compactness and separation of the clusters[33].

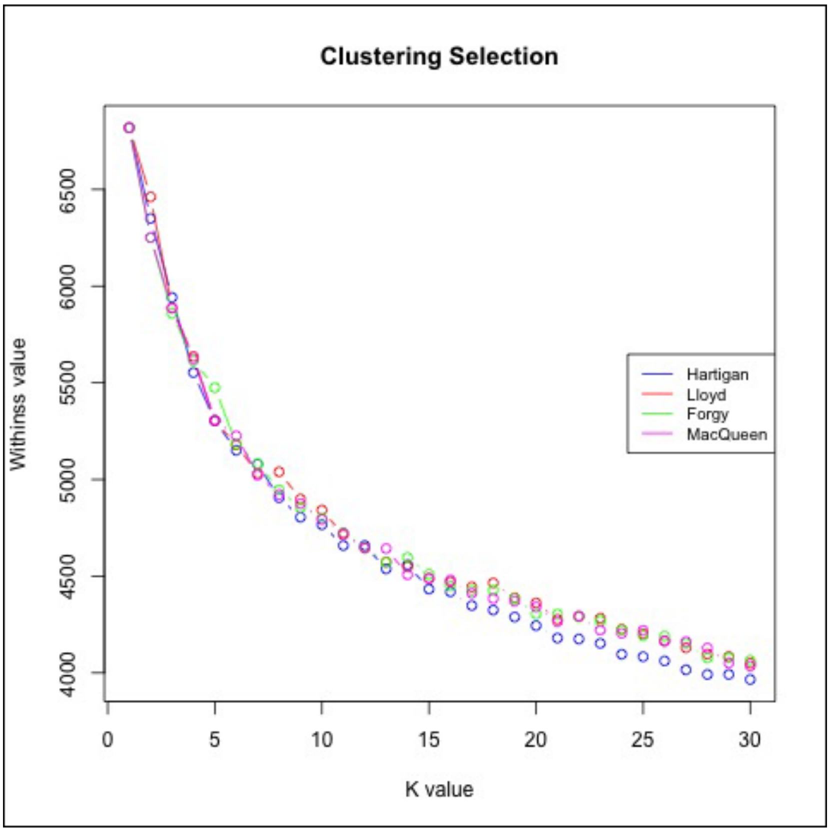

Figure 3 Evolution of the intra-cluster measure by running K-Means with its four methods

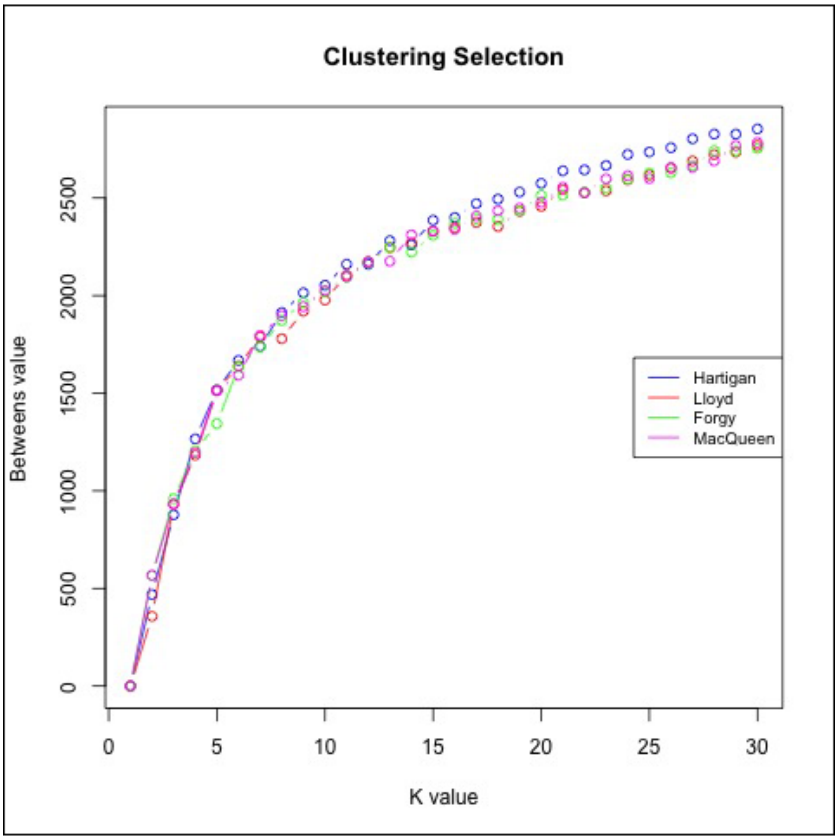

Figure 4 Evolution of the inter-cluster measure by running K-Means with its four methods

The chosen K-Means clustering was the one with the minimum intra-cluster value (5553,208), which matched with Hartigan-Wong's method and K=4 (Figure 3). The clustering candidate had a value close to zero (0,09) for the second quality criterion, average Silhouette width, 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). Thus, a high value for this index represents a desirable scenario of clusters number[34].

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. Both indicators were referred to what degree the topology reflected the relationships in input data (sample data). SOM lattice is highlighted by its topological ordering, whose relations intend to be preserved by mapping process, in such a way that two very similar high-dimensional entities should also have a similar position in a two-dimensional space[18]. The chosen SOM clustering was the one with the minimum estimated topographical accuracy (38,136) and the maximum average distortion measure (0,98). In order to maintain a parallelism between both algorithms, the number of clusters for SOM implementation was also of K=4.

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.

Interpretation of clustering

Measure criteria of the previous stage were focused on data themselves and clusters were evaluated without prior knowledge of MOOC participants. This stage, clustering interpretation, was the process that made possible the ultimate goal of KDD system, the extraction of unknown knowledge and useful information from a subset of variables from the MOOC pre-questionnaire. According to Brachman & Anand[30], this is an extremely difficult task from a technical perspective and, moreover, the lack of information about the optimal structure clustering constitutes a significant obstacle [20].

Results and Discussions

Due to the heterogeneity of MOOC participants' profiles, there was no prior knowledge about their number within the specific MOOC of this study. The application of unsupervised clustering techniques allowed the selection of the best of all resulting clustering from both algorithms with the help of the established quality criteria. These two sets of clusters showed to what extent every feature contributed to the internal structure for the identified MOOC participants' profiles by running K-Means with the method Hartigang-Wong and SOM.

An overview into the different profiles evince 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 comprises the different clusters in order to verify this first impression.

Number of participants	Profile1	Profile2	Profile3	Profile4
K-Means	105	277	48	227
SOM	42	278	120	217

Table 1 Number of participants per profile

One of the strengths of SOM is to encourage the visualisation of unknown relationships into topological representations. This study also exploits the advantage that provides SOM in order to show a circular heatmap of the resulting clustering with a differentiation through colours of the four numbered profiles detailed in Table 1 (Figure 5).

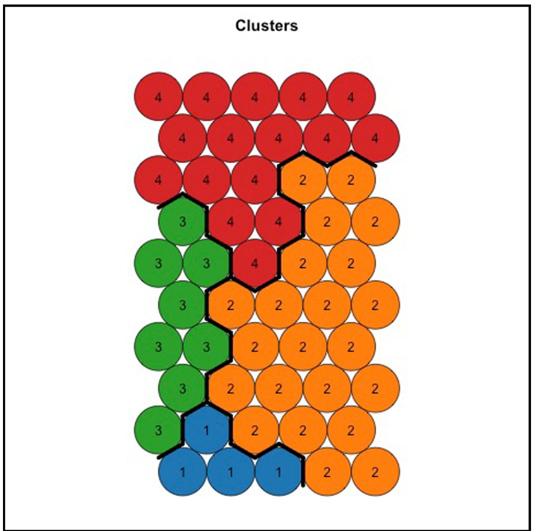

Figure 5 Visualisation of resulting SOM clustering[35]

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 fairly similar range for the clusters of both approaches. The weights of gender belonged 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 unexplainable weight at first sight. It might seem that its participants had taken a significant number of MOOCs although, of course, an in-depth analysis should be required.

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 2 Demographics and MOOC experience of participants for K-Means 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

Table 3 Demographics and MOOC experience of participants for SOM clustering

The feature identified as Human Development Index (HDI) had similar weights for both techniques, although it seemed that in SOM could prevail slightly higher weights. However, the weights reflected that the countries of residence of most participants were those with a high- or medium-HDI indexes. HDI weights are shown in Table 4 and Table 5.

Feature	K_Profile1	K_Profile2	K_Profile3	K_Profile4
HDI_very high	0,133	0,076	0,063	0,097
HDI_high	0,486	0,801	0,563	0,599
HDI_medium	0,333	0,108	0,333	0,278
HDI_low	0,048	0,014	0,042	0,026

Table 4 HDI of participants' countries of residence for K-Means

Feature	S_Profile1	S_Profile2	S_Profile3	S_Profile4
HDI_very high	0,095	0,097	0,100	0,000
HDI_high	0,714	0,637	0,700	0,396
HDI_medium	0,167	0,237	0,167	0,147
HDI_low	0,024	0,029	0,033	0,009

Table 5 HDI of participants' countries of residence for SOM

Among the items for the educational level of a participant, the only one with a predominant weight was second stage of tertiary education for both clustering, with a major weight for all profiles except for one on SOM clustering. The weights of participants' educational level values are shown in Table 6 and Table 7.

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
educational level	pre-primary education	0	0	0,021	0,004
	primary education or first stage of basic education	0,019	0,018	0	0,004
	lower secondary or second stage of basic education	0,038	0,036	0,021	0,048
	(upper) secondary education	0,076	0,101	0,021	0,123
	post-secondary non-tertiary education	0,067	0,051	0,021	0,079
	first stage of tertiary education	0,114	0,188	0,229	0,225
	second stage of tertiary education	0,686	0,606	0,688	0,515

Table 6 Participants' educational level for K-Means

Feature		S_Profile1	S_Profile2	S_Profile3	S_Profile4
educational level	pre-primary education	0,000	0,000	0,008	0,005
	primary education or first stage of basic education	0,000	0,011	0,000	0,005
	lower secondary or second stage of basic education	0,071	0,032	0,067	0,014
	(upper) secondary education	0,071	0,101	0,108	0,051
	post-secondary non-tertiary education	0,048	0,065	0,050	0,041
	first stage of tertiary education	0,214	0,180	0,167	0,111
	second stage of tertiary education	0,595	0,612	0,600	0,327

Table 7 Participants' educational level for SOM

The items student and employed for wages had the highest weights in K-Means for K_Profile4 and K_Profile2, respectively. It stood out that it could be characterized young students for K_Profile4 (its average age was 28 years), although it would be needed a further analysis in order to verify this hypothesis. K_Profile2 showed the same circumstance with the item employed for wages and the average age 49 years, that could characterize middle age employed for wages people. In short, in K-Means, except for K_Profile4 with a strong presence of students, the other three profiles stood out for a more meaningful presence of people employed for wages and out of work and looking for work. The highest weights in SOM were the items employed for wages and out of work and looking for work, that had similar weights for S_Profile1. At a certain distance it was also highlighted the presence of students. A similar behavior was observed for S_Profile2 and S_Profile3, although the relationship between weights varied from one profile to another. The weights of participants' employment status are shown in Table 8 and Table 9.

Feature		K_Profile1	K_Profile2	K_Profile3	K_Profile4
emploiment status	homemaker	0,019	0,007	0	0,013
	student	0,229	0,011	0,167	0,476
	employed for wages	0,381	0,473	0,458	0,181
	out of work and looking for work	0,248	0,267	0,208	0,207
	out of work but not currently looking for wages	0,01	0,029	0	0,013
	retired	0	0,036	0,021	0,018
	self-employed	0,076	0,116	0,083	0,044
	unable to work	0,01	0,025	0	0,004
	others	0,029	0,036	0,063	0,044

Table 8 Participants' employment status for K-Means

Feature		S_Profile1	S_Profile2	S_Profile3	S_Profile4
employment status	homemaker	0,024	0,011	0,000	0,005
	student	0,214	0,212	0,225	0,120
	employed for wages	0,310	0,342	0,283	0,230
	out of work and looking for work	0,333	0,252	0,275	0,111
	out of work but not currently looking for wages	0,000	0,029	0,008	0,009
	retired	0,000	0,007	0,058	0,014
	self-employed	0,000	0,104	0,083	0,028
	unable to work	0,000	0,014	0,025	0,005
	others	0,119	0,029	0,042	0,032

Table 9 Participants' employment status for SOM

The features setting goals and type of interactions had a different structure to those described above. In these two features, participants were asked to express their views through a 7-Likert scale, so that they assessed each of the items that made up both features based on their subjective criterion.

One of the most interesting features for this study was the setting of participant's goals because of its distribution of weights on every cluster. K_Profile1 attracted the attention with the highest weights for all and each of the five items, supported by a very favourable attitude (always true) of participants. The most significant weights in the items that made up this feature in K-Means reflected the positive attitudes of participants (very often and fairly often true). SOM, on the other hand, had a quasi-identical circumstance in terms of profiles' behavior, although all their weights were very similar or lower. S_Profile1 had the highest weight in SOM for the item participant's confidence in the quality assurance of their work with a very favourable attitude (always true) and the rest of the weights were more or less similar in each of the items, where participants' attitude was represented by a positive attitude (very often true). Therefore, this feature should be analyzed in a more detailed way. The most relevant weights of the items for participants' goals setting are shown in Table 10 and Table 11.

goals setting	K_Profile1	K_Profile2	K_Profile3	K_Profile4
standards establishment				
very often true	0,105	0,354	0,271	0,238
always true	0,829	0,011	0,229	0,013
fairly often true	0,010	0,188	0,125	0,313
short- and long-term goals establishment				
very often true	0,114	0,357	0,25	0,251
always true	0,867	0,029	0,333	0,057
high standards maintenance				
very often true	0,019	0,347	0,271	0,251
always true	0,933	0,025	0,313	0,048
fairly often true	0,010	0,188	0,125	0,269
temporal planification management				
very often true	0,086	0,343	0,292	0,251
always true	0,876	0,036	0,333	0,026
fairly often true	0,010	0,191	0,083	0,286
confidence in work quality assurance				
very often true	0,067	0,357	0,271	0,251
always true	0,810	0,166	0,417	0,167
fairly often true	0,019	0,181	0,063	0,256

Table 10 Participants' goals setting for K-Means

goals setting	S_Profile1	S_Profile2	S_Profile3	S_Profile4
standards establishment				
very often true	0,357	0,263	0,267	0,147
short- and long-term goals establishment				
very often true	0,357	0,277	0,325	0,111
high standards maintenance				
very often true	0,333	0,270	0,242	0,134
temporal planification management				
very often true	0,286	0,252	0,300	0,157
confidence in work quality assurance				
very often true	0,357	0,266	0,217	0,171
always true	0,381	0,284	0,350	0,129

Table 11 Participants' goals setting for SOM

The perception of the participants regarding the three types of interaction were very positive (extremely and very important) in both approaches. In K-Means, learner-learner interaction was the less important interaction because of its low weights for positive participants' attitudes (extremely important in K_Profile1 and moderately important in the other three profiles). Learner-content interaction was the most important interaction because of its highest weights for participants' attitudes (extremely important in K_Profile1 and very important in the other three profiles). Learner-teacher interaction followed the same trend that learner-content interaction, although with slightly lower weights (extremely important in K_Profile1 and very important in the other three profiles). In SOM, learner-learner interaction followed the same trend that in K-Means, it was the less important one (very important in S_Profile3 and moderately important in the other three profiles). Learner-content interaction in SOM, as in K-Means, was depicted with the highest weights except for S_Profile4 (very important in three of the four profiles although the weights were more similar between S_Profile1 and S_Profile2). Finally, participants' attitudes for learner-teacher interaction did not show such a regular behavior as the ones described above, although it was highlighted that participants who belonged to S_Profile1 considered very important this type of interaction with the highest weight. Undoubtedly, the three interactions played their role in each and every one of the profiles, even on those where the weight was low and, for sure, an in-depth analysis should be accomplished. The most prominent weights of the items for the three types of interactions are shown in Table 12 and Table 13.

Types of interactions	K_Profile1	K_Profile2	K_Profile3	K_Profile4
learner-learner interaction				
very important	0,267	0,264	0,229	0,22
extremely important	0,295	0,051	0,083	0,031
moderately important	0,162	0,339	0,292	0,330
neutral	0,124	0,177	0,25	0,238
learner-content interaction				
very important	0,219	0,523	0,479	0,454
extremely important	0,705	0,267	0,375	0,352
learner-teacher interaction				
very important	0,219	0,379	0,396	0,432
extremely important	0,59	0,199	0,25	0,189
moderately important	0,095	0,314	0,271	0,26

Table 12 Types of interactions for K-Means

Types of interactions	S_Profile1	S_Profile2	S_Profile3	S_Profile4
learner-learner interaction				
very important	0,190	0,252	0,317	0,115
moderately important	0,310	0,317	0,267	0,147
learner-content interaction				
very important	0,476	0,435	0,425	0,249
extremely important	0,452	0,406	0,383	0,194
learner-teacher interaction				
very important	0,500	0,356	0,383	0,166
extremely important	0,190	0,306	0,250	0,147
moderately important	0,262	0,245	0,258	0,166

Table 13 Types of interactions for SOM

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 approachers (K-Means and SOM) and finally determines which one is closer to a real picture of MOOC participants.

Conclusions

In this study it was chosen two types of algorithms from two different approaches, a partitional clustering algorithm and an artificial neural network. K-Means and SOM were performed in order to find out, with the application of selected quality measures, the (sub)-optimal clustering for both of them. 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 the personal development and could represent a way of discovering the intrinsic structures within the data sample and, consequently, designers and other policy-makers might 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 or SOM) without an additional analysis where the role of MOOC experts is more than relevant.

A more realistic understanding of the profiles of the people is a step forward for many disciplines that call for a more in-depth knowledge of their customers and Open Education is no exception, as it also might be positively impacted by a deeper knowledge of the heterogeneity of profiles that can be found in MOOC format. Therefore, future work in the short to medium term involves a deeper research of clustering techniques, specially both evaluation and interpretation stages, with the involvement of the whole data collection of MOOCKnowledge project.

Bibliography

[1] Liyanagunawardena, T.R., 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 Distance Learning , 14* (3), 202-227.

[2] UNESCO. (2012). 2012 Paris OER Declaration. *2012 World Open Educational Resources (OER) Congress UNESCO.* Paris.

[3] Commission of the European Communities (2001). *Making a European Area of Lifelong Learning a Reality.* COM(2001) 678 final, Brussels.

[4] Santos-Hermosa, G., Ferran-Ferrer, N., & Abadal, E. (2012). Recursos educativos abiertos: repositorios y uso. *El profesional de la información , 21* (2), 136-145.

[5] 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.

[6] Farias, R., Durán, E.B., & Figueroa, S.G. (2008). Las Técnicas de Clustering en la Personalización de Sistemas de e-Learning. In *XIV Congreso Argentino de Ciencias de la Computación (CACIC).*

[7] 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.

[8] Forgy, E.W. (1965). Clustering analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics, 21*, pp. 768-769.

[9] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariante observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* (Vol. 1, No. 14, pp. 281-297). Berkeley: University of California Press.

[10] Hartigan, J.A., & Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28* (1), 100-108.

[11] Daniel, J., Cano, E.V., & Cervera, M.G. (2015). El futuro de los MOOC: ¿aprendizaje adaptado o modelo de negocio? *RUSC. Universities and Knowledge Society Journal, 12* (1), 64-74.

[12] Sinha, T. (2013). Supporting MOOC instruction with social network analysis. *Summer Internship from June-August 2013.*

[13] Yousef, A. M.F., Chatti, M. A., Wosnitza, M., & Schroeder, U. (2015). Análisis de clúster de perspectivas de participantes en MOOC. *Monográfico: Los MOOC: ¿una transformación radical o una moda pasajera?*, *12* (1), 74-91.

[14] Kizilcec, R.F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK'13)*, (pp. 170-179). ACM.

[15] Lynn, S. (2014a). Self-Organising Maps for Customer Segmentation. Theory and worked examples using census and customer data sets. (D. R. Group, Ed.)

[16] Lynn, Shane (2014) Self-Organising Maps for Customer Segmentation using R. R-bloggers.

[17] Fang, S.Y., Malt, B.C., Ameel, E., & Li, P. (2013, November). A computational model of semantic convergence in bilingual lexicon. In *Talk Presented at the 43rd Annual Meeting of the Society for Computers in Psychology (SCiP)*.

[18] Wehrens, R., & Buydens, L. M. (2007). Self- and Super-organizing Maps in R: The Kohonen Package. *Journal of Statistical Software , 21* (5), 1-19.
[19] Ravikumar, S., & Shanmugam, A. (2012). Comparison of SOM Algorithm and K-Means Clustering Algorithm in Image Segmentation. *International Journal of Computer Applications (0975-8887) , 46* (22), 21-25.
[20] Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S., & Zhang, M.Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica 12* , 241-262.
[21] Baçao, F., Lobo, V., & Painho, M. (2005). Self-Organizing Maps as Substitutes for K-Means Clustering. In *Computational Science-ICCS 2005,* (pp. 476-483).: Springer Berlin Heidelberg.
[22] Jain, A.K., Murty, M.N., & Flynn, P.J. (1999). Data Clustering: A Review. *ACM Computing Surveys (CSUR), 31* (3), 264-323.
[23] Hertz, J., Krogh, A., & Palmer, R.G. (1991) Introduction to the Theory of Neural Computation*. Reading: Addison-Wesley*.
[24] Deligiorgi, D., Philippopoulos, K., & Kouroupetroglou, G. (2014). An Assessment of Self-Organizing Maps and k-means Clustering Approaches for Atmospheric Circulation Classification. In *Proceedings of the 2014 International Conference on Environmental. Recent Advances in Environmental Science and Geoscience*, (pp. 17-23).
[25] Kohonen, T. (1989). *Self-Organization and Associate Memory* (3rd edition ed.). Springer. New York.
[26] Kind, M.C. & Brunner, R.J. (2013) SOMz: photometric redshift PDFs with self organizing maps and random atlas*. Monthly Notices of the Royal Astronomical Society,* 438(4), 3409-3421.
[27] Balakrishnan, P.S., Cooper, M.C., Jacob, V.S., & Lewis, P.A. (1994). A study of the classification capabilities of neural networks using unsupervised learning: A comparison with k-means clustering. *Psychometria, 59* (4), 509-525.
[28] Waller, N.G., Kaiser, H.A., Illian, J.B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika, 63* (1), 5-22.
[29] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-54.
[30] Brachman, R. J., & Anand, T. (1994, July). The Process of Knowledge Discovery in Databases: A First Sketch. *AAAI Technical Report WS-94-03.* AT&T Bell Laboratories. Atlanta.
[31] Jahan, S. (2015). *Human Development report 2015. Work for Human Development.* United Nations Development Programme (UNDP), New York.
[32] Moore, M.G. (1989). Editorial: Three Types of Interaction.
[33] 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.
[34] Cárdenas Montes, M. (2013). Clustering: Clasificación no supervisada. Gráficas estadísticas y minería de datos con python.
[35] Team, R.C. (2015). R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. *Document freely available on the internet at: http://www.r-project.org*.