Article

Journal of Educational Computing Research
0(0) 1–20

DOI: 10.1177/0735633117753364
journals.sagepub.com/home/jec

Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course: An Exploratory Study

Youngjin Lee[1]

Abstract

This study investigated whether clustering can identify different groups of students enrolled in a massive open online course (MOOC). This study applied self-organizing map and hierarchical clustering algorithms to the log files of a physics MOOC capturing how students solved weekly homework and quiz problems to identify clusters of students showing similar problem-solving patterns. The usefulness of the identified clusters was verified by examining various characteristics of students such as number of problems students attempted to solve, weekly and daily problem completion percentages, and whether they earned a course certificate. The findings of this study suggest that the clustering technique utilizing self-organizing map and hierarchical clustering algorithms in tandem can be a useful exploratory data analysis tool that can help MOOC instructors identify similar students based on a large number of variables and examine their characteristics from multiple perspectives.

Keywords

massive open online course, Educational Data Mining, log file analysis, self-organizing map, clustering

[1]University of Kansas, Lawrence, KS, USA

Corresponding Author:
Youngjin Lee, University of Kansas, 1122 W. Campus Rd., Lawrence, KS 66045-3101, USA.
Email: yjlee@ku.edu

Introduction

Since the first large-scale massive open online course (MOOC) was launched by Coursera in 2011, the number of people enrolled in MOOCs are increasing every year. According to Class Central (2015), more than 35 million students signed up for at least one MOOC in 2015. One important characteristic of an MOOC is that students can be enrolled in the course anytime, and they are not required to complete all learning activities available in the course. Due to this openness and flexibility, the ways in which students are engaged in learning in MOOCs are quite different, compared to traditional learning environments such as face-to-face or closed online courses in which all students are expected to complete the same set of learning activities during a fixed time period. Most notably, a large number of students enrolled in MOOCs do not complete the course. Previous studies have found that the completion rate of MOOC students is much lower than that of face-to-face or closed online courses (Breslow et al., 2013; Ho et al., 2014; Jordan, 2014, 2015).

Another important characteristic of MOOCs, as a computer-based learning environment, is that they can capture in the log files what students are doing without interrupting their learning processes. Since information recorded in the log files allows for reconstruction of how students were using various learning resources, log file analysis opens up a new avenue for understanding how students learn in the computer-based learning environment. By carefully analyzing log files, it is possible to quantitatively study the learning behavior of students and the usefulness of computer-based learning contents. As a result, Educational Data Mining (EDM) is emerging as a new, exciting field of study.

According to Baker (2010), research in EDM falls into five general categories: prediction, clustering, relationship building, discovery with a model, and distillation of data for human judgment. While the first three categories are universally acknowledged by all data mining researchers, the fourth and the fifth categories are the ones investigated primarily by EDM researchers. The goal of prediction is to develop a quantitative model that can infer a single aspect of the data (e.g., whether MOOC students will earn a course certificate) from other aspects of the data (e.g., how MOOC students solved weekly homework problems). In clustering, the goal is to identify a set of clusters or groups of data points showing similar characteristics (e.g., finding MOOC students showing similar problem-solving patterns over the semester). Relationship mining tries to discover frequent patterns among (usually a large number of) variables in the data. In EDM, relationship mining can be used to identify the sequence of courses or pedagogical strategies that can improve the learning outcome of students. In discovery with a model, an EDM model developed via prediction or clustering is then used as a component of another analysis such as prediction or relationship mining. Distillation of data for human judgment is an approach that aims to extract information from a large amount of data in order to help human users to make a better inference about the phenomenon of interest when

it is beyond the scope of fully automated data mining methods. Using these approaches, EDM researchers are investigating various educational issues and problems such as providing feedback for instructors and students, predicting learning performance of students, developing cognitive models of how students learn specific knowledge and skills, detecting undesirable behaviors of students, or grouping students according to their characteristics (Peña-Ayala, 2014; Romero & Ventura, 2010).

This exploratory study has two objectives. The first objective is to investigate whether a clustering technique applied to *fine*-grained learner behaviors can identify different groups of students enrolled in an MOOC. While previous studies summarized in the literature review below aggregated learner behaviors over the entire *semester*, this study performed a clustering analysis employing self-organizing map (SOM) and hierarchical clustering algorithms on the *daily* and *weekly* problem-solving performance of students. Using SOM and hierarchical clustering algorithms together allows for easier exploration of complex, multidimensional log file data, which can lead to better understanding of problem-solving patterns of students in the identified clusters. The second objective is to examine how the problem-solving patterns of students in the identified clusters are related to the completion rate of the MOOC. The rest of the article is organized as follows. The Literature Review section summarizes previous research on completion rate of MOOCs, and clustering and SOM analysis conducted on the log files of computer-based learning environments. The Method section describes the log files of an MOOC analyzed, and SOM and hierarchical clustering algorithms used in this study. The Results section describes the problem-solving patterns of students in the identified clusters of students, followed by discussions and limitations of the study.

Literature Review

Completion Rate of MOOCs

Breslow et al. (2013) investigated the completion rate of 154,763 students who signed up for a physics MOOC in Spring 2012. They found that 15% of registrants attempted to complete the first homework assignment, 6% of them passed the midterm exam, and only 5% of them were able to complete the course and earned the course certificate. Ho et al. (2014) examined the completion rate of students who were enrolled in seventeen HavardX and MITx MOOCs that were offered between Fall 2012 and Summer 2013. They found that about 5% of registrants were able to earn the course certificate, 4% of them explored half or more of course materials without certifications, 56% of them viewed less than half of course materials without certifications, and 35% of them were never engaged in learning. Jordan (2015) surveyed 129 MOOCs having a varying length of study and a different grading scheme to examine the completion rate

of MOOC registrants. She found that the completion rate of MOOCs varied from 0.7% to 52.1% with median of 12.6%, depending on the length of course (shorter ones having higher completion rates), start date (more recent courses having higher completion rates), and assessment type (courses using auto grading only having higher completion rates). In response to MOOC critics who are concerned with low completion rates, Reich (2014) argues that it is important to take into account the intention of students when we examine the completion rate of MOOCs. His analysis of MOOC data from nine HarvardX courses, which had over 290,000 registrants and 79,000 survey responses, showed that 58% of registrants intended to earn a certificate, 25% of them to audit, 3% of them to browse, and 14% of them were unsure about their intention. He found that 22% of registrants who intended to complete a course earned the certificate, whereas only 6% of registrants who intended to browse a course earned the certificate.

Clustering Analysis on MOOCs

Researchers conducted a clustering analysis on the clickstream MOOC data in order to identify groups of students showing similar learning behaviors. Kizilcec, Piech, and Schneider (2013) examined the patterns of engagement of students enrolled in three computer science MOOCs offered on the Coursera platform. They found the four distinct patterns of engagement, completing, auditing, disengaging, and sampling, when they applied a clustering technique to the clickstream data capturing how MOOC students watched video lectures and solved summative assessment problems. Khalil and Ebner (2017) compared the engagement patterns of university students to those of nonuniversity students enrolled in the MOOC by conducting a clustering analysis on the frequency of reading, writing, playing videos, and quiz attempts. Liu, Brown, Lynch, Barnes, Baker, Bergner and McNamara (2016) used a hierarchical clustering algorithm to investigate the relationship between engagement patterns of MOOC students and their background such as nationality. Ferguson and Clow (2016) investigated whether the four engagement patterns identified in Kizilcec et al. (2013)'s study can be replicated in the MOOCs that are based on the socioconstructivist pedagogy emphasizing discussions and formative assessment problems, rather than video lectures and summative assessment problems. Rodrigues et al. (2016) used hierarchical and k-means clustering algorithms to examine how students used discussion forums in the MOOC. Ezen-Can, Boyer, Kellogg, and Booth (2015) applied a k-medoids clustering algorithm to group similar posts in the discussion forum of an MOOC and compared the resulting clusters to the annotations created by human experts. Li, Kidzinski, Jermann, and Dillenbourg (2015) applied clustering to user interactions to examine how students watched MOOC videos (e.g., pausing, forward and backward seeking, and speed changing), and the relationship between user interactions and perceived difficulty of videos, video revisiting behaviors, and student performance on the course

assignments. Balint, Teodorescu, Colvin, Choi, and Pritchard (2017) applied a spectral clustering technique to various performance-based measurements (e.g., fraction of correct answers to assessment problems, IRT skill parameter, etc.) to group students based on their ability and examined how students in each ability group used various learning resources, such as text, worked examples, video with and without human presence, computer simulation, discussion board, calendar, and syllabus, available in the MOOC.

SOM in EDM Research

Literature on applying SOM to EDM is sparse, compared to information retrieval and traditional data mining. Only a handful of published studies employing SOM in the context of EDM exist. Merelo-Gervos et al. (2004) utilized SOM to create a community map visualizing clusters of community members having similar interests determined by the URLs of web-based resources they used. Durfee, Schneberger, and Amoroso (2007) applied factor analysis and SOM to examine the relationship between student characteristics, such as demographics, computing skills, expertise in using computer software, and self-efficacy, and their adoption and use of computer-based training and assessment software. Lee (2007) used SOM, *k*-means clustering, and principal component analysis (PCA) to assess the mastery level of students who are learning in an integrated online learning environment. He found that applying PCA to the data preprocessed with SOM and *k*-means clustering algorithms performed much better than the conventional PCA-only approach. Recently, Ahmad, Ishak, Alias, and Mohamad (2015) used SOM to analyze the learning activities of 19 students taking a computer science course. The results of their analysis suggest that SOM can identify clusters of students showing similar learning behaviors in terms of the websites and the course materials they visited and downloaded.

Method

Data Set

This study conducted a clustering analysis utilizing SOM and hierarchical clustering algorithms on the log files that captured how 4,337 students solved weekly homework and quiz problems while they were enrolled in edX 8.MReVx Mechanics Review offered by MIT in Summer 2014 semester (hereafter, 8.MReVx). 8.MReVx is designed to provide a comprehensive overview of Newtonian mechanics and greater expertise in problem-solving. It provides various learning resources, such as e-texts, videos, discussion boards, wiki, checkpoints, weekly homework problems, quizzes, midterm exam, and final exam, to help students learn Newtonian mechanics concepts. In 8.MReVx, the achievement of students was determined by checkpoints (8%), homework

problems (34%), quizzes (36%), midterm exam (7%), and final exam (16%). Students who scored more than 60% of the maximum possible points received a course certificate. Checkpoints are easy formative assessment problems embedded in the e-text of the MOOC, whereas homework and quiz are more difficult summative assessment problems assigned each week during the 12-week long semester. Students were given 7 days to complete checkpoints, homework, and quiz problems that were due on Sunday at midnight every week. For further exploration of the course structure and the problems students solved in 8.MReVx, visit the archived course at https://courses.edx.org/courses/MITx/8.MReVx/2T2014/course/.

Of these learning activities, this study focused on solving weekly homework and quiz problems because of the following two reasons. First, these summative assessment problems could have incentivized students to exert more efforts because they were more important than checkpoint problems in getting a course certificate. Second, midterm and final exam scores were excluded because this study was conducted as an exploratory data analysis that aimed to identify variables that may be able to predict the midterm and the final exam scores of students. More specifically, this study focused on (a) how many weekly homework and quiz problems students tried to solve throughout the semester (N_{total}); (b) the daily problem completion percentage ($DPCP$) 6, 5, 4, 3, 2, and 1 day(s) before and on due date; and (c) the weekly problem completion percentage ($WPCP$) at Weeks 1, 2, 3, 4, 5, 6, and 7. $WPCP$s beyond Week 7 were not included in the analysis because many students solved few problems once they had accumulated enough points for the course certificate around Week 7. These 15 variables (N_{total}, $DPCP_6$ – $DPCP_{\text{due}}$, $WPCP_1$ – $WPCP_7$) were used to represent the problem-solving pattern of students in the data set analyzed in this study.

Finding Clusters of Students Using SOM and Hierarchical Clustering Algorithms

SOM is an artificial neural network that is designed to map a multidimensional data (e.g., N_{total}, $DPCP_6$ – $DPCP_{\text{due}}$, $WPCP_1$ – $WPCP_7$) to an X-Y plane (Haykin, 1999). SOM is different from other clustering algorithms in the respect that it places similar data points (e.g., MOOC students showing similar learning behaviors) close together in the X-Y plane, allowing for easy visualization and exploration of complex data. This study used an open source R package called kohonen (Wehrens & Buydens, 2007) to create an SOM of students enrolled in 8.MReVx, based on how they solved weekly homework and quiz problems.

After students showing similar problem-solving patterns are placed nearby on the SOM, a hierarchical clustering algorithm can be applied to produce clusters of students. Selecting a number of clusters is notoriously problematic without a priori domain knowledge. As a number of clusters in the SOM increases,

students assigned in each cluster become more homogenous in terms of their problem-solving patterns. However, the SOM with too many clusters would have a less practical importance because it is difficult to interpret the meaning of resulting SOM clusters. In order to determine an optimal number of clusters in the SOM, elbow method and Calinski–Harabasz (CH) index were utilized in this study. Elbow method is the most frequently used heuristics in determining an optimal number of clusters or factors in the clustering and factor analysis. It looks for an *elbow* point in the plot of total within sum of squares (WSS) versus number of clusters (k) (Antonenko, Toy, & Niederhauser, 2012). CH index is a ratio of between-cluster variance to total within-cluster variance, which is maximized at an optimal number of clusters (Caliński & Harabasz, 1974). The plot of WSS versus number of clusters (k) suggests an elbow point at $k = 4$, which corresponds to the number of clusters at which CH index is maximized (see Figure 1(b)). Based on the agreement between elbow method and CH index, an SOM with four clusters was created as shown in Figure 1(a).

Results

Fraction of Certificate Earners

This study found that only 434 students, out of 4,337 students who attempted to solve at least one weekly homework or quiz problems, were able to get a course certificate at the end of the semester. This result is in line with previous studies reporting a completion percentage lower than 10% (Breslow et al., 2013; Jordan, 2014, 2015). In addition, we found that certificate and noncertificate earners were concentrated in specific SOM clusters. Moreover, 99.94% of students in Cluster 1 did not get a course certificate, whereas 91.24% and 94.08% of students in Cluster 3 and Cluster 4 got a certificate. Number of noncertificate earners in Cluster 1 ($N = 3{,}687$) accounts for 94% of all noncertificate earners in 8.MReVx. Unlike the other clusters that were quite homogeneous in terms of student composition, Cluster 2 was found to be a mixture of certificate and noncertificate earners; it consists of 148 students who earned a course certificate and 194 students who did not (see Figure 2).

When a hierarchical clustering algorithm was directly applied to the original 15 variables, rather than SOM-transformed variables, it is much more difficult to visualize the resulting clusters. Figure 3(a) is a cluster dendrogram created from the original 15 variables capturing how randomly selected 434 students solved weekly homework and quiz problems. Although the cluster dendrogram includes only 10% of the data (it would be nearly impossible to visualize a dendrogram from the full data), it is already much more difficult to see the clusters, compared to the SOM created from the full data shown in Figure 1(a). More importantly, the clusters obtained from the original 15 variables

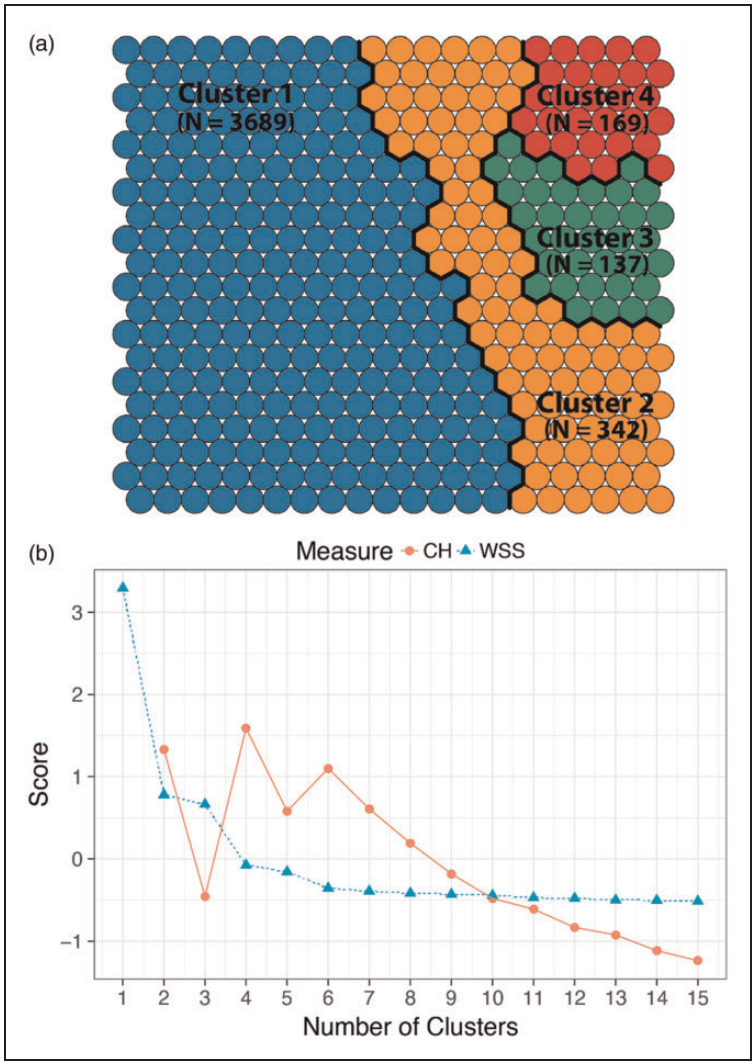

Figure 1. (a) SOM of problem-solving patterns of students enrolled in 8.MReVx. Numbers in the parentheses indicate the number of students in each cluster in the SOM. (b) *WSS* and CH index vs. number of clusters (*k*). To help make *WSS* and CH index comparable, standardized scores were used. WSS = within sum of squares; CH = Calinski-Harabasz.

appear to be less meaningful. Figure 3(b) shows the percentage of students who did and did not earn a course certificate in each cluster created from the original 15 variables. In this case, Cluster 1 and Cluster 2 exclusively contain students who did not get a course certificate, but Cluster 3 and Cluster 4 have a mixture

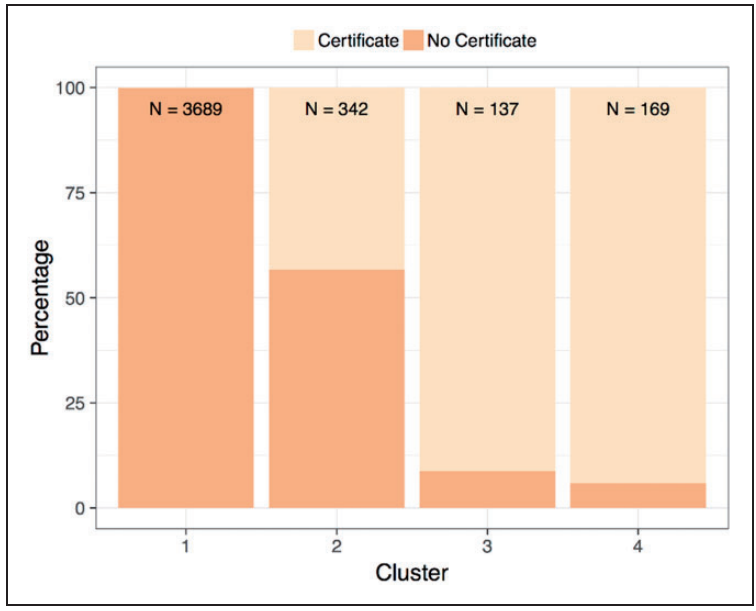

Figure 2. Percentage of students who did and did not earn a course certificate. Numbers indicate the size of each cluster.

of certificate and noncertificate earners. The hierarchical clustering without SOM was not able to identify a cluster of students who earned a course certificate. Since the SOM-based clustering yielded more informative clusters, which is in line with what Vesanto and Alhoniemi (2000) had found in their research, the subsequent analyses were focused on the SOM-based clustering approach.

In order to better understand the students in the SOM-based Cluster 2, their success in getting a course certificate was fit to a logistic regression model with total number of problems (N_{total}), daily problem completion percentages ($DPCP_6$ – $DPCP_{\text{du}}$), and weekly problem completion percentages ($WPCP_1$ – $WPCP_7$) as predictor variables. Table 1 summarizes the logistic regression coefficients, their standard errors, and p values. Not surprisingly, certificate earners in Cluster 2 tried to solve more problems than their peers who did not get a course certificate. When all the other predictor variables were held constant at their mean values, solving one more weekly homework or quiz problem (N_{total}) increases the log odds for getting a course certificate by 0.006 ($p = .005$). Similarly, certificate earners in Cluster 2 showed a much higher problem completion percentage on due date, compared to Cluster 2 students who did not get a course certificate; 1% increase in the problem completion percentage on due date ($DPCP_{\text{due}}$) increases the log odds for getting a course certificate by 0.384 ($p < .0001$). Also, the logistic regression analysis suggests that students in Cluster

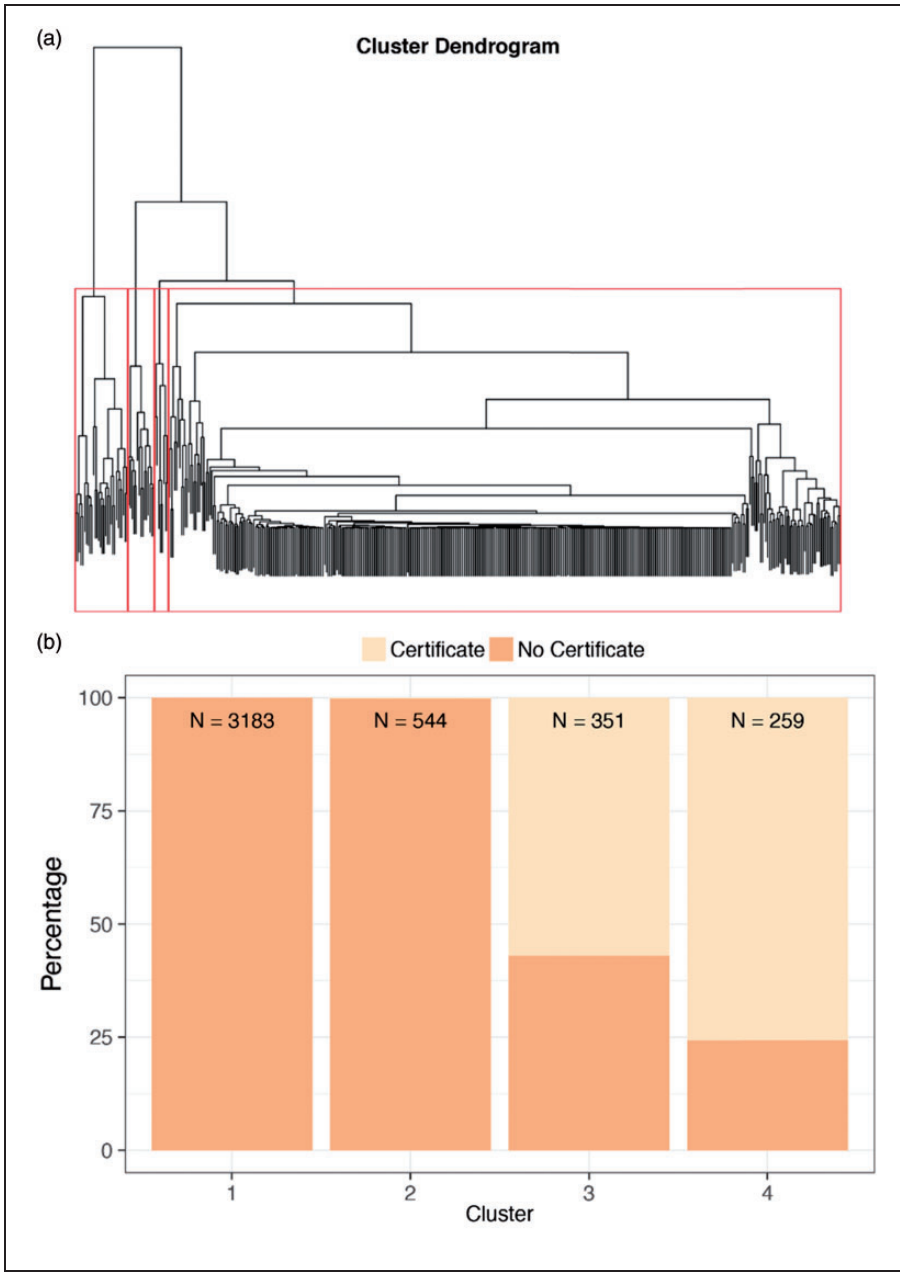

Figure 3. (a) Cluster dendrogram created from a randomly selected 10% of original data. Red rectangles indicate the four clusters identified. (b) Percentage of students who did and did not earn a course certificate. Numbers indicate the size of each cluster created from the original 15 variables.

Table 1. Result of Logistic Regression Analysis on the Problem-Solving Patterns of Students in Cluster 2.

Variable	Estimate	Standard error	*p*
N_{total}	0.006	0.002	**.005**
$DPCP_6$	0.157	0.107	.143
$DPCP_5$	0.425	0.142	**.003**
$DPCP_4$	0.242	0.112	**.031**
$DPCP_3$	0.112	0.086	.156
$DPCP_2$	−0.135	0.072	.061
$DPCP_1$	−0.008	0.043	.859
$DPCP_{due}$	0.384	0.082	**<.0001**
$WPCP_1$	−0.031	0.012	**.011**
$WPCP_2$	−0.076	0.016	**<.0001**
$WPCP_3$	0.001	0.011	.896
$WPCP_4$	−0.001	0.011	.320
$WPCP_5$	−0.001	0.010	.898
$WPCP_6$	−0.007	0.010	.498
$WPCP_7$	0.005	0.007	.489

Bold values denote the significant *p*- values, highlighting the predictors.

2 who solved more homework and quiz problems early have a higher chance to get a course certificate; 1 point increase in the problem completion percentage five ($DPCP_5$) and 4 days before due date ($DPCP_4$) are associated with the increased log odds for getting a course certificate by 0.425 and 0.242, respectively. Interestingly, the log odds for getting a course certificate decrease slightly when the weekly problem completion percentages at Week 1 ($WPCP_1$) or 2 ($WPCP_2$) increases. It may indicate that students in Cluster 2 who did not get a course certificate tried little bit harder than their successful peers when the course contents were easier early in the semester. However, these effects are not as strong as the other significant predictors, such as $DPCP_5$, $DPCP_4$, and $DPCP_{due}$, as indicated by the magnitude of the regression coefficients.

Number of Problems Students Attempted to Solve

As shown in Figure 4, students in Cluster 1 tried to solve much fewer problems than students in the other clusters. The median of number of problems students in Cluster 1 attempted to solve was just 17. Considering the fact that 99.94% of students in Cluster 1 did not earn a course certificate, it is not a surprising result. On the other hand, students in Cluster 3 and Cluster 4 tried to solve a comparable number of problems. The medians of number of problems students in Cluster 3 and

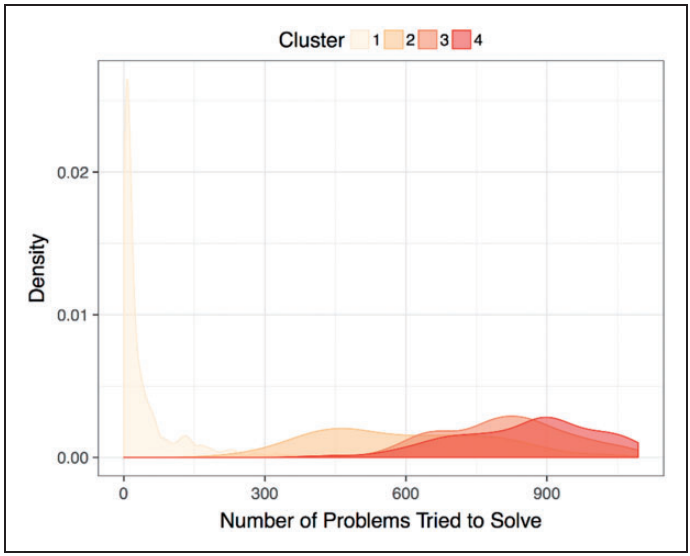

Figure 4. Density plots of number of problems students in each cluster tried to solve throughout the semester.

Cluster 4 tried to solve were 810 and 883, respectively. Students in Cluster 2 did not try to solve as many problems as students in Cluster 3 and Cluster 4. However, these students tried to solve much more problems (median = 576.5) than students in Cluster 1 (median = 17), indicating that these students exerted much more efforts to earn a course certificate than students in Cluster 1 who appeared to have quickly dropped out of the course early in the semester.

When students in Cluster 2 are divided into two groups, those who did and did not get a course certificate, their difference becomes much clearer (see Figure 5). As expected, Cluster 2 students who got a course certificate solved much more problems (median = 751.5) than students who did not (median = 475.0). Interestingly, Cluster 2 students who did not get a course certificate solved much more weekly homework and quiz problems, compared to students in Cluster 1. In other words, although these students may look similar in terms of their success in getting a course certificate, noncertificate earners in Cluster 2 exerted much more efforts (especially during the first 3 weeks of the semester as explained later) than students in Cluster 1 who dropped out of the course early.

Daily Problem Completion Percentages

Since students were given 7 days to complete the weekly homework and quiz problems, we examined how their problem completion percentage changes during the 1-week assignment cycle. Students in Cluster 3 and Cluster 4 are similar

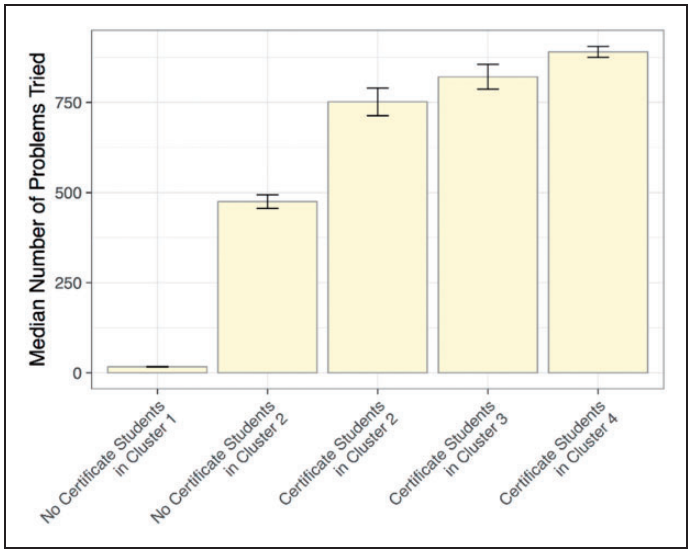

Figure 5. Median of number of problems students who did and did not earn a course certificate tried to solve throughout the semester. The error bars indicate 95% confidence intervals computed from bootstrapping.

in the respect that majority of them were able to successfully complete the course; 91.24% and 94.08% of students in Cluster 3 and Cluster 4 got a course certificate at the end of the semester (see Figure 2). However, the difference lies in when they started working on the homework and quiz problems during the week. On average, students in Cluster 4 were able to complete about 50% of their weekly homework and quiz problems 6 days before the due date, which is, on average, what students in Cluster 3 were able to achieve by the due date (see Figure 6).

Students in Cluster 3 seem to have allocated their efforts evenly throughout the week, resulting in a constant slop in their problem completion percentage plot as shown in Figure 6. On the other hand, students in Cluster 2 seemed to have exerted more efforts as the due date came closer, which is shown as a sharp increase in the completion percentage near the due date in Figure 6. However, these students did not have enough time to complete as many problems as students in Cluster 3 and Cluster 4 probably because they started working on their problems too late during the week. As a result, on average, students in Cluster 2 ended up with completing less than 40% of the assigned homework and quiz problems by the due date.

Weekly Problem Completion Percentages

Figure 7 shows how the median of homework and quiz completion percentages were changing during the first 7 weeks of the semester. Students in Cluster 1 were

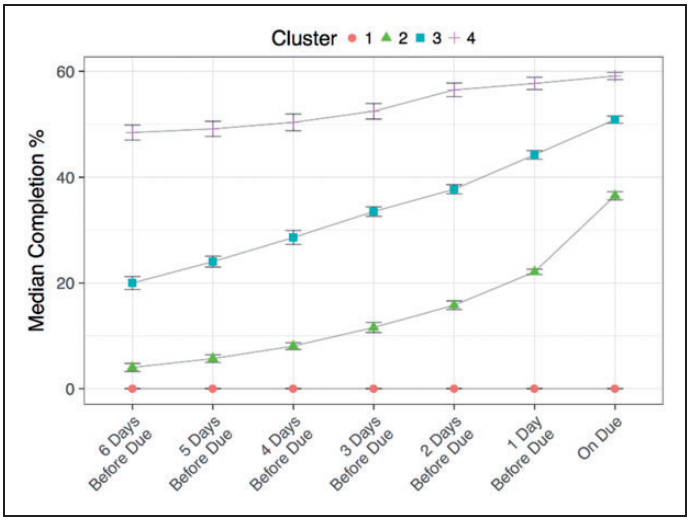

Figure 6. Median of problem completion percentages during the 1-week assignment cycle. The error bars are 95% confidence intervals computed from bootstrapping.

not engaged in problem-solving at all. These students must have been dropped out of the course very early in the semester, and it is not surprising to find that 99.94% of them did not get a course certificate. As discussed earlier, students in Cluster 3 and Cluster 4 were quite similar in terms of the fraction of certificate earners (see Figure 2) although students in Cluster 4 started working on the homework and quiz problems very early each week (see Figure 6). Similarly, their weekly problem completion profiles are similar except that students in Cluster 4 were able to complete more homework and quiz problems than students in Cluster 3 probably because they were able to spend more time working on the problems (see Figure 7).

Students in Cluster 2 show an interesting pattern in their weekly problem completion percentage over time. During the first 3 weeks, students in Cluster 2 were able to complete a good number of weekly homework and quiz problems, compared to students in Cluster 3 and Cluster 4 who were able to get a course certificate. However, their median completion percentage started to drop at Week 4, and many students in this cluster seemed to have given up completely at Week 7 as shown in Figure 7.

The heat maps of weekly problem completion percentages visualize how students in Cluster 2 were changing during this time period. At Week 1, many students in Cluster 2 were as hot as students in Cluster 3 and Cluster 4, indicating that these students were able to complete as many homework and quiz problems as certificate earners in Cluster 3 and Cluster 4. At Week 5, the top portion of the heat map for Cluster 2 got cooler, indicating that these students

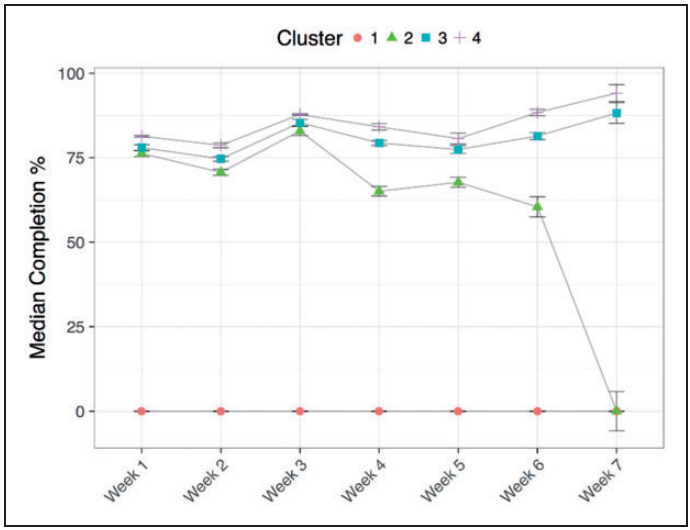

Figure 7. Median of problem completion percentages from Week 1 to Week 7. The error bars are 95% confidence intervals computed from bootstrapping.

were getting behind. Finally, the heat map of the Week 7's problem completion percentage clearly shows that more than 50% of students in Cluster 2 became completely cold, indicating that they did not solve any problems, just like non-certificate earners in Cluster 1 (see Figure 8).

Discussion

Based on the facts that 94% of students who did not earn a course certificate belong to Cluster 1, 99.94% of students in Cluster 1 were not able to get a course certificate, and they solved very few problems, Cluster 1 seemed to be equivalent to Kizilcec et al.'s (2013) disengaging students who must have dropped out of the course early. On the other hand, Cluster 3 and Cluster 4 seemed to include students who were actively engaged in learning. Students in these clusters solved more than 50% of the weekly homework and quiz problems, and more than 93% of them were able to successfully complete the course. Students in Cluster 4 were *early* starters because they finished a large portion of their weekly homework and quiz problems very early each week, whereas students in Cluster 3 seemed to have exerted their efforts evenly throughout the 1-week assignment cycle. Unlike the other clusters, Cluster 2 contained students who did and did not get a course certificate. An interesting finding about the students in Cluster 2 who did not get a course certificate is that they exerted a substantial amount of effort during the early part of the semester. These students completed as many weekly homework and quiz problems as certificate earners during the first 3

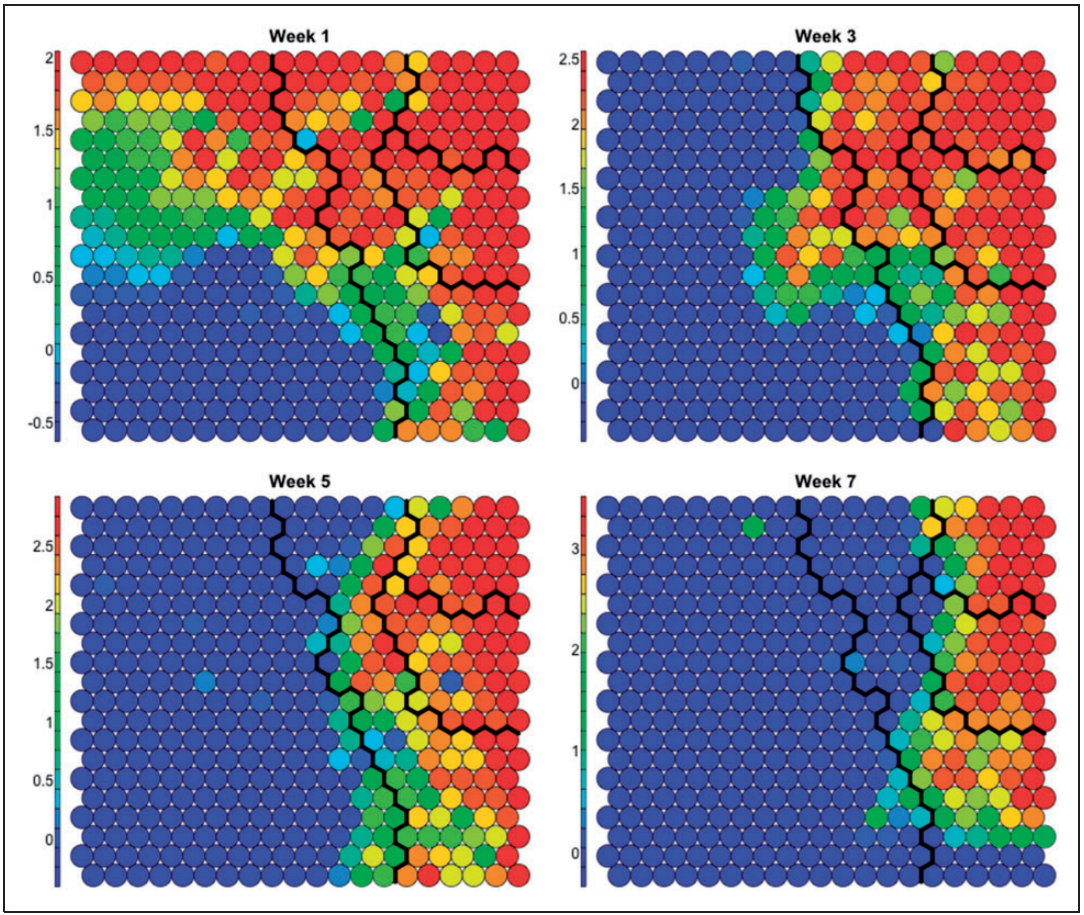

Figure 8. Heat maps showing the problem completion percentages of students at Week 1, 3, 5, and 7. The black lines mark the border of clusters.

weeks but appeared to have given up as the semester was progressing as shown in the heat maps of the SOM reported in this study.

Thanks to the advancement in information technology, it is easy to collect information about how students use a computer-based learning environment such as MOOCs. However, it is not easy to use this multidimensional information when we try to understand how students learn in the computer-based learning environment in part because it is extremely difficult to make sense of multifaceted data. By placing similar multidimensional data close to one another in an X-Y plane, SOM allows us to easily visualize clusters of students who share similar characteristics measured on multiple dimensions. Therefore, MOOC instructors can use SOM as a profiling tool that can enable them to easily identify groups of students having similar characteristics and examine their academic performance.

From the perspective of learning, students in Cluster 2 would be most important because many of them did not get a course certificate even though they exerted significant efforts during the first 3 weeks of the semester. These students might have been able to successfully complete the course if they received appropriate supports and guidance in time (probably around Week 4). SOM may be able to help instructors find the students who need help the most, which might be a first step toward addressing the low completion rate issue in MOOCs.

Identifying meaningful predictor variables is one of the most important and difficult tasks in the quantitative data analysis examining clickstream data obtained from computer-based learning environments because the quality of predictor variables has much more impact on the predictive power of the model than the complexity of the algorithm employed in the analysis. As a result, many predictive modeling projects start with an exploratory data analysis trying to identify good predictor variables for the phenomenon being modeled. SOM can be an effective exploratory data analysis method since it can not only cluster students showing similar characteristics but also provide an easy way to visualize the characteristics of the students in each cluster.

Limitations of Study and Future Works

The focus of this study was on the pattern of student engagement in solving weekly summative assessment problems in MOOCs. Considering the fact that solving problems that have one correct answer is not a primary learning activity in certain knowledge domains (e.g., social studies, literature), the findings from this study may not be generalized to the MOOCs emphasizing different types of learning activities and pedagogies (e.g., discussions with peers). It would be interesting to conduct a similar clustering analysis on the clickstream data from MOOCs employing socioconstructivist pedagogies. Also, this study examined the program completion rate of students in the clusters identified by SOM and hierarchical clustering algorithms. Although the program completion rate is an important characteristic of students, especially from the perspective of

instructors and MOOC providers, there are other important characteristics that can be studied. It would be meaningful to examine other characteristics of students, such as gender, level of prior education, nationality, ethnicity, socioeconomic status, and other performance-based measurements, in the identified clusters.

In this study, unscaled variables were used in building the SOM and the subsequent clusters of students taking 8.MReVx. Since some variables, such as number of problems students tried to solve, are skewed, appropriate preprocessing, such as log or Box-Cox transformations, could have been performed. Also, although it is possible that there is an interaction between variables (e.g., weekly vs. daily problem completion percentages), it was not considered in this study. Examination of interactions between preprocessed variables may lead to identifying more homogeneous clusters of students.

As a future work, we plan to build a predictive model that can estimate the performance of students in the midterm and final exams from how students solved weekly homework and quiz problems, in addition to other predictor variables summarizing their learning behaviors in MOOCs, because weekly problem-solving patterns seem to successfully separate students who earned a course certificate from those who did not. Finally, this study created one SOM from the entire clickstream data from the 12-week long semester. It would be interesting to create SOMs from smaller chunks of the clickstream data (e.g., Week 1–Week 3, Week 4–Week 6, Week 7–Week 9, and Week 10–Week 12) and compare how they are similar or different in terms of the learning behaviors of students.

Acknowledgment

The author thanks Professor David E. Prichard at Massachusetts Institute of Technology for providing 8.MReVx log files analyzed in this study.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Ahmad, N. B., Ishak, M. K., Alias, U. F., & Mohamad, N. (2015). An approach for e-learning data analytics using SOM clustering. *International Journal of Advances in Soft Computing & Its Application*, *7*(3), 94–112.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383–398.

Baker, R. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 112–118). Oxford, UK: Elsevier Science.

Balint, T. A., Teodorescu, R., Colvin, K., Choi, Y.-J., & Pritchard, D. (2017). Physics instructional resource usage by high-, medium-, and low-skilled MOOC students. *The Physics Teacher*, *55*(4), 222–225.

Breslow, L., Pritchard, D. E., DeBoer, J. D., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research Practice in Assessment*, *8*, 13–25.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Class Central. (2015). By the numbers: MOOCS in 2015 – Class Central's MOOC report. Retrieved from https://www.class-central.com/report/moocs-2015-stats/

Durfee, A., Schneberger, S., & Amoroso, D. L. (2007). Evaluating students computer-based learning using a visual data mining approach. *Journal of Informatics Education Research*, *9*, 1–28.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*. Retrieved from http://doi.acm.org/10.1145/2723576.2723589

Ferguson, R., & Clow, D. (2016). Consistent commitment: Patterns of engagement across time in Massive Open Online Courses (MOOCs). *Journal of Learning Analytics*, *2*(3), 55–80.

Haykin, S. (1999). *Neural Networks* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. *SSRN Electronic Journal*. doi:10.2139/ssrn.2381263

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online course. *The International Review of Research in Open and Distributed Learning*, *15*(1), 133–160.

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *The International Review of Research in Open and Distributed Learning*, *16*(3). doi:10.19173/irrodl.v16i3.2112

Khalil, M., & Ebner, M. (2017). Clustering patterns of engagement in Massive Open Online Courses (MOOCs): The use of learning analytics to reveal student categories. *Journal of Computing in Higher Education*, *29*(1), 114–132.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*. Retrieved from http://doi.acm.org/10.1145/2460296.2460330

Lee, C.-S. (2007). Diagnostic, predictive and compositional modeling with data mining in integrated learning environments. *Computers & Education*, *49*(3), 562–580.

Li, N., Kidzinski, L., Jermann, P., & Dillenbourg, P. (2015). MOOC video interaction patterns: What do they tell us?. *Lecture Notes in Computer Science*, *9307*(6), 197–210.

Liu, Z., Brown, R., Lynch, C., Barnes, T., Baker, R., Bergner, Y., & McNamara, D. (2016). MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis.

Proceedings of the 9th International Conference on Educational Data Mining. Retrieved from https://pdfs.semanticscholar.org/3c0a/68193049732f890b0632ed72d6e19d5119e8.pdf?_ga=2.84644377.1215507837.1515768297-87580748.1508175188

Merelo-Gervos, J. J., Perieto, B., Prierto, A., Romero, G., Valdivieso, P. C., & Tricas, F. (2004). Clustering web-based communities using self-organizing maps. *Proceedings of IADIS Conference on Web Based Communities*. Retrieved from https://pdfs.semanticscholar.org/f58b/1caa38e5673d6d533da16ba567034e1c27a7.pdf?_ga=2.139507507.1215507837.1515768297-87580748.1508175188

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(P1), 1432–1462.

Reich, J. R. (2014). MOOC completion and retention in the context of student intent [EDUCAUSE Review Online]. Retrieved from http://www.educause.edu/ero/article/mooc-completion-and-retention-context-student-intent

Rodrigues, R. L., Gomes, A. S., Souza, F. F., Ramos, J. L. C., Silva, J. C. S., & Maciel, A. M. A. (2016). Discovering level of participation in MOOCs through clusters analysis. *Proceedings of the IEEE 16th International Conference on Advanced Learning Technologies*. doi: 10.1109/ICALT.2016.45

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600.

Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organizing maps in R: The Kohonen package. *Journal of Statistical Software*, *21*(5), 1–19.

Author Biography

Youngjin Lee is an associate professor of Educational Technology at the University of Kansas. His current research is focusing on using Educational Data Mining and learning analytics to better understand how people learn in computer-based learning environments such as massive open online courses.