Education Tech Research Dev (2012) 60:383–398
DOI 10.1007/s11423-012-9235-8

RESEARCH ARTICLE

Using cluster analysis for data mining in educational technology research

Pavlo D. Antonenko · Serkan Toy · Dale S. Niederhauser

Published online: 21 February 2012

Abstract Cluster analysis is a group of statistical methods that has great potential for analyzing the vast amounts of web server-log data to understand student learning from hyperlinked information resources. In this methodological paper we provide an introduction to cluster analysis for educational technology researchers and illustrate its use through two examples of mining click-stream server-log data that reflects student use of online learning environments. Cluster analysis can be used to help researchers develop profiles that are grounded in learner activity—like sequence for accessing tasks and information, or time spent engaged in a given activity or examining resources—during a learning session. The examples in this paper illustrate the use of a hierarchical clustering method (Ward's clustering) and a non-hierarchical clustering method (*k*-Means clustering) to analyze characteristics of learning behavior while learners engage in a problem-solving activity in an online learning environment. A discussion of advantages and limitations of using cluster analysis as a data mining technique in educational technology research concludes the article.

Keywords Cluster analysis · Online learning environments · Learner profiles · Server logs · Data mining

Introduction

Over the past two decades we have seen a dramatic rise in internet use, transforming our society in ways that have made us increasingly dependent on using information to

P. D. Antonenko (✉)
Oklahoma State University, 210 Willard Hall, Stillwater, OK 74078, USA
e-mail: pasha.antonenko@okstate.edu

S. Toy
Children's Mercy Hospital & University of Missouri, Kansas City,
2401 Gillham Road, Kansas City, MO 64108, USA

D. S. Niederhauser
Iowa State University, N155 Lagomarcino Hall, Ames, IA 50011, USA

accomplish a wide variety of tasks in our personal and professional lives. Vast amounts of data accumulated daily in web server-logs and databases can provide useful insights into patterns associated with how individuals use informational resources. The exponential growth in the size and complexity of these data sets essentially renders direct, hands-on data analysis impossible. Analysis of these massive datasets has increasingly been augmented with indirect automatic data processing techniques like artificial neural networks, clustering and genetic algorithms, decision trees, and support vector machines. The umbrella term for these analytical methods is *data mining*–the process of applying statistical algorithms to discover patterns and correlations in large datasets—which can reveal new meaning in the data (Nisbet et al. 2009).

In education, widespread use of online learning environments (OLE's) drives the need to identify effective methods for analyzing large-scale server-log datasets as we work to improve our understanding of users' learning strategies. Cluster analysis is a group of data classification methods that can be particularly useful in this context (e.g., Barab et al. 1997; Lawless and Kulikowich 1996). Unlike more commonly used video-recording and screen-capture methods, which require time-consuming examination and coding of hours of video data, clustering algorithms can be used to group and organize server-log click-stream data to make the analysis process more manageable and efficient.

Click-stream provides a rich data source that can help educational researchers make sense of the results of learning that occurs in OLEs (Schrader and Lawless 2007). Click-stream data is a collection of learners' navigational choices (mouse clicks and keystrokes) that are automatically recorded on the host web server. These data can be analyzed to provide detailed information on learner processes like the sequence of accessing online resources and activities, the rate at which learners advance through the learning environment, and the amount of time spent examining resources. Unlike frequently used think-aloud techniques, collecting click-stream data does not interfere with the cognitive processing associated with the primary learning task because it is collected "behind the scenes" (Clark 2010; Schrader and Lawless 2007).

Analysis of click-stream data can help establish patterns of learner behavior based on their interactions with content in the learning environment. This article provides an overview of cluster analysis for educational technology researchers and provides examples of the use of two types of clustering methods for mining click-stream server data and creating profiles of learner behavior in OLEs.

Overview of cluster analysis

Cluster analysis is a group of statistical methods that has been used extensively for data mining in a number of fields including bioinformatics, industrial engineering, marketing, e-commerce, and counter-terrorism (Everitt et al. 2009). Typically used as an exploratory analysis tool, cluster analysis techniques group cases of data such that the degree of association with respect to target variables between two cases is maximal if they belong to the same group and minimal otherwise. Cluster analysis can be thought of as complementary to factor analysis: factor analysis groups *variables* across cases (e.g., individuals), clustering algorithms group *cases* based on the variables of interest. Thus, cluster analysis provides a useful technique that can be used to help researchers manage and organize large datasets—and use the established clusters in subsequent analyses.

The first step of cluster analysis involves computing proximity indices between each pair of participants relative to the variables of interest. The most straightforward and generally accepted measure of proximity for continuous and interval data is squared Euclidean distance (Everitt et al. 2009; Jain et al. 1999). Squared Euclidean distance (d^2) is the sum of the squared differences across variables. Squaring differences accentuates distance between variables and places progressively greater weight on cases that are further apart. Also, unlike Pearson's correlation, it reflects all three dimensions of multivariate data—level, scatter, and shape (Cronbach and Gleser 1953). Proximity indices are computed by statistical software as the first step of conducting cluster analysis.

Once the proximity indices are known, a clustering algorithm can be used to group similar participants into homogeneous subgroups—or clusters. Different clustering algorithms are grounded in different assumptions for grouping participants (see Table 1). The clustering process can be either hierarchical or non-hierarchical. Hierarchical procedures start with each case as a separate cluster, then sequentially combine clusters to construct a hierarchy of nested clusters reducing the number of clusters at each step until all cases are combined into one cluster. By inspecting the progression of cluster merging one can isolate clusters of cases with high similarity. Hierarchical algorithms like Ward's minimum variance clustering (Ward 1963) are useful for exploratory work when researchers do not have a preconceived idea about the likely number of clusters in the dataset.

Non-hierarchical algorithms, on the other hand, are appropriate when there is a theoretical or empirical rationale for predicting the number of clusters or when the data set is large (e.g., hundreds or thousands of cases). For non-hierarchical clustering algorithms like *k*-means clustering (MacQueen 1967), a specified *k* number of clusters are "forced" in an effort to confirm or refute initial hypothesis about the structure of the data set. Using this initial cluster information, the method calculates centroids for a set of trial clusters, then places each object in the cluster with the nearest centroid, recalculates the centroids and reallocates the objects. This process continues until there are no more changes in the cluster membership. This process is computationally more efficient than hierarchical clustering and is frequently used with large data sets (e.g., hundreds or thousands of cases) for exploratory analysis. It is common to run *k*-means clustering three or four times with three, four, or five clusters as *k*. The final number of clusters can then be confirmed via a hierarchical algorithm like two-step clustering (Table 1).

Criteria for selecting an appropriate clustering method include nature of the data (continuous vs. nominal) and size of the data matrix (number of cases and variables). Some algorithms can only be used with continuous variables while others can handle categorical variables as well, and algorithms commonly used for small data sets are typically impractical for data files with thousands of cases (see Table 1).

Table 1 Characteristics of the popular clustering algorithms

	Algorithm Type	Data Set Size	Data Type	Limitations
Ward's clustering	Hierarchical	Dozens of cases	Continuous	Tends to create many small clusters
Average linkage	Hierarchical	Dozens of cases	Continuous or nominal	Sensitive to outliers and measurement scales, so raw scores should be standardized
Two-step	Hierarchical	Thousands of cases	Continuous and/or nominal	Sensitive to order effects, so order of cases must be randomized
k-means clustering	Non-hierarchical	Hundreds of cases	Continuous or nominal	Number of clusters must be specified a priori

So, which clustering algorithm is best in any given situation? Does the method adequately define groups when they are present? To answer these questions, scholars in the area of data mining and data classification have conducted sophisticated evaluations of major clustering methods to determine the effectiveness of each method in identifying the natural structure in the data. In this line of research, the most common approach has been the Monte Carlo simulation—when the researcher creates an artificial data set with known (pre-defined) group structure and a random error, and then test the usefulness of each clustering algorithm in detecting and recovering these known groups. Such studies have been published in journals like *Psychometrika*, *Multivariate Behavioral Research*, and *Journal of Educational and Behavioral Statistics*, and the interested reader is encouraged to browse these publications (e.g., Milligan 1980). Our synopsis of about a dozen Monte Carlo studies of the past two decades is that currently most psychometricians agree that Ward's method and average linkage are the most recommended hierarchical clustering algorithms and the *k*-means method is the suggested non-hierarchical clustering algorithm (e.g., Aldenderfer and Blashfield 1984; Nisbet et al. 2009). This consensus is also reflected in the design of the Predictive Analytics SoftWare™ (PASW™, formerly SPSS™)—one of the most popular statistical analysis programs for behavioral and social scientists, which provides *k*-means, Ward's, average linkage, and two-step algorithms as the options for performing cluster analysis.

In some cases, both hierarchical and non-hierarchical techniques are used successively. For instance, a hierarchical method can be used initially with a small sample of the larger data set to get a sense of the possible number of clusters and how they merge. Then the entire data set can be analyzed using a more efficient non-hierarchical method with a predetermined number of clusters.

Finally, once participants have been grouped through an appropriate clustering algorithm, it is important to verify the initial clusters that were created because clustering algorithms produce clusters even when the data does not appear to contain natural subgroups. Thus, it is incumbent on the researcher to examine the groupings and adjust as necessary based on a systematic review of the cluster analysis. A clustering structure is deemed valid if it cannot reasonably have occurred by chance or as a result of an artifact (Jain et al. 1999). Verification of the clustering structure or cluster validity analysis typically involves the examination of fusion coefficient plots and group means across clusters as well as alignment of the number and nature of clusters with what is already known in the field about this type of data.

The examples below illustrate the application of hierarchical and non-hierarchical cluster analysis in educational technology research involving problem-solving in OLEs. Each study includes the following steps: (1) cluster identification, (2) cluster validation, and (3) cluster interpretation. The first study employed a hierarchical clustering algorithm (Ward's clustering) to group 59 undergraduate teacher education students into clusters based on the proportion of time they spent on writing tasks and visiting information resources to solve a real-life problem individually within a problem-based OLE (Toy 2008). The second study used a non-hierarchical clustering method (*k*-means clustering) to group the browsing sessions of 183 undergraduate engineering majors solving an engineering economics problem in small teams using a problem-based OLE (Niederhauser et al. 2007).

Example 1: hierarchical clustering

The use of hierarchical cluster analysis in educational technology research can be illustrated using an empirical study conducted to identify different problem-solving strategies used by individuals while solving a complex, real-life problem within an online problem-solving environment (Toy 2008). In this study, the problem-solving process was scaffolded and delivered in an OLE that tracked the sequence and proportion of time spent on information resources and writing tasks as the participants attempted to solve the problem. Subjects were 59 (32 female, 27 male) pre-service secondary education students in an introductory instructional technology course at a large Mid-western university. For this study, an ill-structured problem was created to address information literacy, plagiarism, and concerns regarding internet use in education. The participants worked on solving the problem individually.

Students completed the problem-solving activity individually in a two-hour lab session in the 18 week of a 15-week semester. Students typically completed the tasks in a given order to proceed forward (see Table 2 for actual tasks). For example, they needed to first analyze the problem, and then generate several possible solution options before reflecting on and justifying their proposed solution. This structure was based on Jonassen's ill-structured problem-solving process (1997), which provided scaffolding for the students to progress through the problem in a systematic manner. Although the environment seemingly forced students to progress in a linear fashion, they could always return to tasks to revise their previous work any time before the final submission.

The activity featured fourteen information resources, some containing relevant information, others with irrelevant information. Resources included brief background information about the main characters in the problem scenario, the educational context of the problem (statistics about the school population and technology available to the students), teacher's lesson plan for the project, the assignment in question, and portions of the questionable websites the student used.

Participants' navigational decisions (i.e., mouse clicks) were captured in the server-log. Click-stream data was transformed using Microsoft Excel™ to compute the amount of time each participant spent on each of the OLE resources and tasks. Analysis of the these data showed that participants spent an average of 63.53 min (SD = 16.13; min = 28.32; max = 100.35) working in the OLE. Overall, participants spent the majority of their time (63%) on writing tasks (see Table 3). An additional 23% of time was allocated to exploring the resources (16% on relevant resources and 7% on irrelevant resources). Participants also

Table 2 Problem-solving tasks

Task description
You recently applied for a science teacher position at East High and got invited to interview for the position. This afternoon you came in for the interview. The principal briefly met and told you about the incident Mr. Whitman had to deal with last year and is very concerned that something like this does not happen again. He has provided you with a folder including several resources with additional information and asked you to provide a written report to be used during the actual interview. To do this, read through the relevant resources, then analyze the problem, propose solutions, and reflect on your solutions.
Analyze problem: After reading the problem scenario and exploring the resources, please describe, in your own words, what you see as the problem in this scenario, examine possible causes of the problem and explain the key issues, constraints, and position of key players.
Propose solutions: Generate 3 possible solutions. Write a paragraph or two describing each solution and evaluate their merit relative to key issues and constraints.
Reflect on solution: Decide which of your proposed solutions you think will be the best way to revise Mr. Whitman's activity to avoid incidents like this. Then, reflect on possible consequences of your proposed solution and explain the relative benefit of your solution compared to the alternative ones.

spent 12% of their time reading the problem scenario and the task description. Finally, a small proportion of time (2%) was spent looking at the help section.

Assessment of problem-solving is a central issue in research on problem-based learning (e.g., Belland et al. 2009; Gijbels et al. 2005). The instrument that was used in this study to measure overall problem-solving performance was "The Holistic Critical Thinking Scoring Rubric" (Facione and Facione, 1994). This rubric was chosen because it reflects Jonassen's model of ill-structured problem-solving (1997) relative to the cognitive and metacognitive skills involved in problem-solving. Thus, participant scores served on "The Holistic Critical Thinking Scoring Rubric" served as the dependent variable to compare the effectiveness of strategies used by each cluster of problem solvers. Participant responses were divided into two groups, each of which was graded by two different graders independently using the rubric. Then, 20 percent of all the students from each group were selected randomly and cross-scored by the two trained graders. A high intraclass correlation coefficient (0.93, Donner and Koval 1980) reflected high inter-rater reliability.

Cluster identification

In this study, cluster analysis was used to identify homogeneous groups of students within the class based on three key variables of interest: proportion of time spent on (a) writing tasks (analyzing the problem, generating solutions, and reflecting on the proposed solution), (b) visiting relevant resources, and (c) visiting irrelevant resources. The empirical rationale for using these three variables is the finding that participants spent 86% of their time on these types of activities. The theoretical rationale for including time spent on writing tasks and resource exploration stems from the conceptual research on the scaffolding of problem-solving in technology-enhanced learning environments. A recent review on this topic defines problem-solving as "situated, deliberate, learner-directed, activity-oriented efforts to seek divergent solutions to authentic problems through multiple interactions amongst problem solver, tools, and other resources" (Kim and Hannafin 2011). The "efforts to seek divergent solutions" are scaffolded by separating the problem-solving process into a series of tasks that have the students analyze the problem, generate solutions, and reflect on the proposed solution (Jonassen 1997)—that is, the writing tasks used in the Toy (2008) study. The "multiple interactions amongst problem solver, tools, and other resources" are scaffolded by providing learners with relevant and irrelevant tools and resources to simulate real-world problem-solving, which typically involves discriminating between vast amounts of extraneous information and finding relevant evidence to generate possible solutions (e.g., Ryan et al. 2007). The scaffolding of problem-solving tasks (Belland et al. 2009, 2010) and resource exploration (Jeong and Hmelo-Silver 2010) is pervasive in modern technology-enhanced problem-based learning environments (e.g.

Table 3 Mean percent of time spent on specific OLE nodes

OLE node	Mean	SD	Minimum	Maximum
Problem scenario and task description	12	3	5	22
Writing tasks	63	13	30	88
Relevant resources	16	11	0	52
Irrelevant resources	7	4	0	29
Help	2	18	0	10

Table 2 in Kim and Hannafin 2011). Furthermore, learners' problem-solving performance is described typically through their performance on problem-solving tasks and the effectiveness and efficiency of their resource exploration strategies (Stevens 2007). Thus, time spent on completing the problem-solving tasks and exploring resources is a useful and measurable representation of the learners' problem-solving inquiry in OLEs.

Hierarchical clustering was selected because the data set was relatively small ($n = 59$) and the researcher had no preconceptions about the likely number of clusters. The most common and efficient hierarchical algorithm (which also happens to be the most conservative one) is Ward's minimum variance clustering (Ward 1963). It uses an analysis of variance approach to evaluate the distances between clusters. Ward's algorithm compares the proximity indices and identifies pairs of participants with the smallest distance value. The pair with the smallest distance is grouped at step 1 of the algorithm. The pair with the next smallest distance is grouped during step 2, and so forth. The algorithm continues to merge groups in ways that keep the within-group variance at a minimum. In other words, Ward's criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

Cluster validation

An important step in determining the optimal number of clusters involves examining the plot of fusion coefficients against the number of clusters (Webb 2002). Figure 1 reflects the data from the Toy (2008) study and shows the distance between clusters on the Y axis, which serves as the fusion coefficient for each stage of the grouping process. The flattening of the curve and creation of an "elbow" in the pattern of fusion coefficients suggests that a four-cluster solution is appropriate. The variance between clusters in the six-cluster solution (there was no five-cluster solution) was not meaningful enough to suggest further partitioning. In the next phase of the analysis, the four-cluster solution was further examined to determine if the clusters were theoretically meaningful.

In the final verification stage clusters are reviewed to see if they provide a conceptually meaningful representation of the data. For example, Toy (2008) disaggregated *Writing Tasks*, *Relevant Resources*, and *Irrelevant Resources* means based on the four-cluster solution (see Table 4). Analysis of the means for the four groups confirmed the groups differed based on the time they spent on writing tasks and resource exploration.

Additionally, comparing to some external variables (i.e., variables that are not used as clustering variables) can be used to enhance the validation of the clusters (Milligan and Cooper 1987 as cited in Lawless and Kulikowich 1996). In this case, the problem-solving performance means were added to determine whether the cluster membership differed relative to this very important variable. A one-way analysis of variance (ANOVA) was conducted with clusters as the between subject factor, and problem-solving performance as dependent variables. Results showed that main effect for problem-solving performance reached significance (F [3, 55] = 3.90; $p < 0.05$; MSE = 47.65). Further post hoc analysis (Tukey HSD) was conducted to determine how clusters differed in problem-solving performance. It was found that individuals in Cluster 4 had significantly higher problem-solving performance scores than those in Cluster 1 (Table 4). Even though students in Cluster 2 scored considerably higher than those in Cluster 1, this contrast probably did not reach significance—perhaps due to the small number of students ($n = 7$) in Cluster 2.

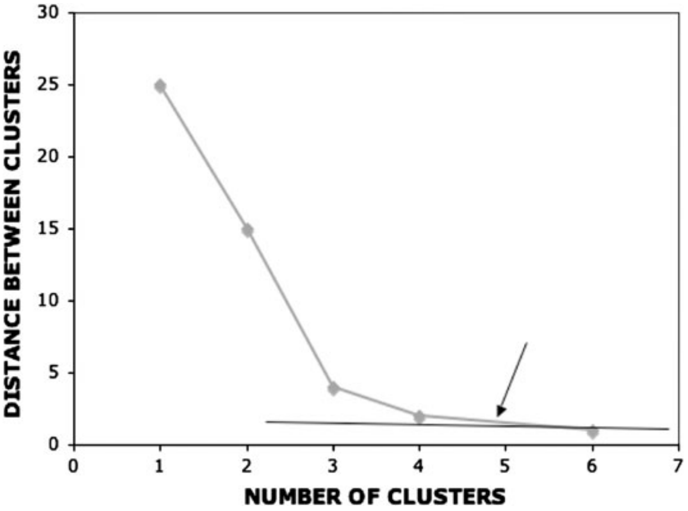

Fig. 1 Fusion coefficients plotted against number of groups generated

Table 4 Group means and standard deviations arrayed by clusters in the Toy (2008) study

	Cluster 1 (n = 13)		Cluster 2 (n = 7)		Cluster 3 (n = 19)		Cluster 4 (n = 20)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Writing tasks	0.53	0.04	0.40	0.08	0.64	0.03	0.77	0.04
Relevant resources	0.22	0.05	0.41	0.07	0.14	0.03	0.06	0.03
Irrelevant resources	0.09	0.06	0.07	0.04	0.08	0.03	0.04	0.02
Problem-solving performance	16.84	3.64	20.71	3.72	19.16	3.32	20.90	3.48

Cluster interpretation

To better understand and interpret different problem-solving strategies as reflected by the clusters, Toy (2008) first compared clusters based on time allocation in terms of (1) task focus and (2) resource use in solving the problem. Then these clusters were compared based on their problem-solving performances to see if clusters differed in terms of the effectiveness of the strategies used to solve the problem-solving tasks. Three ANOVAs were conducted to examine between-cluster differences in the task focus of students' problem-solving processes and the nature of their resource use. For all three analyses, cluster membership was used as the between subjects factor (1, 2, 3, and 4) and proportions of time spent on (1) completing writing tasks, (2) visiting relevant resources, and (3) visiting irrelevant resources (i.e., clustering variables) served as dependent measures. Based on the results of these analyses reported in detail in Toy (2008), it was possible to describe Cluster 1 as *non-discriminating investigators*, Cluster 2—*discriminating investigators*, Cluster 3—*non-discriminating writers*, and Cluster 4—*writers* (these students made very little use of resources).

In order to understand the patterns of resource use, sequential analysis of the click-stream data sorted by cluster membership was used. Each navigational choice was transformed into a number and color-coded cell in a spreadsheet. A complete session of the problem-solving navigational activity for each individual was represented as a row including all of the numbered and color-coded cells in the sequence in which they were accessed. This analysis allowed the researchers to further describe the clusters in the context of their problem-solving performance. Students in Cluster 1 were the least effective problem solvers who did not seem to discriminate relevant resources from irrelevant ones, and tended to revisit these irrelevant resources engaging in what can be called ineffective resource cycling (*extensive ineffective cycling investigators*). Students in Cluster 2 tended to consult mostly relevant resources revisiting them to confirm their understanding of the problem space (*extensive effective cycling investigators*). Although students in Cluster 3 (*minimal ineffective cycling writers*) were not particularly effective in separating relevant resources from irrelevant ones, they received satisfactory scores on the problem-solving task, which indicates that they eventually were able to identify relevant information, unlike those in Cluster 1. Students in Cluster 4 were effective problem solvers perhaps because they chose a convenient path through the problem space. They visited resources in the order that they appeared in the OLE from top to bottom and then focused their attention mainly on completing writing tasks (*minimal effective cycling writers*).

In this study cluster analysis has proven useful in examining click-stream server-log data to group individual problem solvers into meaningful clusters. Ward's clustering helped identify and characterize four distinct problem-solving strategies individuals used to solve a complex, real-life problem in an OLE. Analyzing these clusters based on resource use and task focus, Toy (2008) was able to define the characteristics of problem-solving strategies demonstrated by students making up each cluster. Post hoc analyses helped identify and describe less and more successful problem-solving strategies.

Example 2: non-hierarchical clustering

Another example that can illustrate the application of cluster analysis in educational technology research is a study that used a non-hierarchical clustering algorithm to identify solution strategies of high- and low-performing teams of undergraduate students engaged in solving a complex story problem in engineering economics (Niederhauser et al. 2007). Unlike the previous example, this study focused on collaborative problem-solving and used browsing sessions rather than individuals as the main unit of analysis.

The problem-solving process was scaffolded for the learner by sequencing problem-solving tasks (i.e., Decision Criteria, Relevant Concepts, Assumptions, Solution, and Uncertainty Analysis) using an online problem-based learning environment called Engineering Learning Portal (Kumsaikaew et al. 2006; Ryan et al. 2007). The OLE was developed to present ill-structured problems in a collaborative online environment. Participants could determine the sequence for completing tasks, and accessing resources, scoring rubric, and tools. The server hosting the OLE recorded a log of the nodes (i.e., resources, tasks, rubric, and tools) that were accessed by students as they attempted to solve the problem. The problem was a complex story problem (Jonassen 2000) that addressed selecting the best mortgage option given a variety of constraints (e.g., interest rate, term, family situation, etc.) Resources included both relevant and irrelevant information.

Undergraduate students in an engineering economics class ($n = 183$) worked on the problem in teams of three–four, and could access the OLE from different computers and work concurrently, gather together at a computer station and work through the problem collaboratively, or individual students could work independently. Each of the five subtasks—Decision Criteria, Relevant Concepts, Assumptions, Solution, and Uncertainty Analysis—contributed to the total score on the problem. Based on solution scores, researchers selected 20 high performing (HP) teams, and 20 low performing (LP) teams. Data from the remaining eight groups were dropped from the analysis.

Cluster identification

In this study, click-stream server-log data was transformed into a two-dimensional matrix containing variables (i.e., node titles) as columns and individual browsing sessions for either HP teams or LP teams as rows (Table 5). Each cell was populated with either a 0 or 1 depending on whether a specific OLE node was accessed during a given browsing session.

K-means clustering (MacQueen 1967) was chosen to partition server-log data into groups of sessions (clusters) that are close to each other based on a measure of squared Euclidean distance between the case and the cluster center used to classify the case (Ng and Han 1994). Thus, a cluster was defined in this study as a set of problem-solving sessions that contained similar lists of task and resource nodes accessed by the student teams.

K-means clustering is described as an algorithm that is more efficient than hierarchical methods because it does not require computation of all possible distances. Instead, *k*-means clustering repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during analysis. On the other hand, agglomerative clustering algorithms like Ward's analysis add cases only to existing clusters, forever capturing cases within a certain cluster, with a widening circle of neighbors (Norušis 2005). Thus, although methods like Ward's analysis optimize within-cluster homogeneity at each stage of grouping, they do not ensure optimal homogeneity of the final clusters—once cases have been merged, they cannot be separated at later stages of grouping. *K*-means clustering avoids this potential bias.

Determining the number of clusters in a data set, a quantity labeled *k*, as in the *k*-means algorithm, is a frequent problem in data clustering (Everitt et al. 2009). In *k*-means clustering, *k* is the number of clusters that have to be pre-specified by the researcher based on prior knowledge about the underlying structure of the data. Alternatively, the researcher can start out with a random partitioning and test the validity of several different cluster options. The random nature of the latter approach avoids the bias present when pre-determining clusters based on existing assumptions. Therefore, in this study *k*-means cluster analysis was rerun on the HP and LP datasets of problem-solving sessions four

Table 5 Example of click-stream data transformed for clustering

Browsing session ID	Problem description	Relevant concepts	Uncertainty analysis	Resource: ARM features	Resource: Banker
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	0	0	0
4	1	1	1	1	1
5	1	1	1	1	1

times—first, with a pre-specified two-cluster solution, second, with a three-cluster solution, third, with a four-cluster solution, and finally, with a five-cluster solution.

Cluster verification

The most appropriate number of clusters in the HP and LP datasets was verified by computing the sum of squared distances between the cases in each cluster and their cluster centroid for each of the four-cluster solutions. This verification strategy suggested that a four-cluster solution was the best representation for the data at hand because the sum of squared distances for the four-cluster solution was the smallest for HP and LP teams.

The number of clusters was also verified by performing two competing clustering algorithms (e.g., Everitt et al. 2009) on the HP data set and LP data set—two-step clustering and the average linkage method. Two-step clustering is applied with large data sets because it first divides the set into pre-clusters and then the pre-clusters are merged using hierarchical clustering. Average linkage minimizes the biases of the single linkage and complete linkage methods by computing the average similarity of a case with all the cases currently in a cluster. Then the next linkage is formed from the pair of candidates with the lowest average similarity (Sokal and Michener 1958). The results of two-step clustering confirmed that the four-cluster solution was the most appropriate one for the problem-solving sessions of both LP and HP teams. The average linkage method produced a four-cluster solution for the HP teams and a five-cluster solution for the LP teams. Analysis of the cluster membership for each of the five clusters showed, however, that the fifth cluster contained only nine cases (the next lowest cluster number of cases per cluster was 32) and visual inspection of the browsing patterns showed that activities represented in these nine browsing sessions (and their sequences) were similar to the activities performed during the 32 sessions captured in cluster 4. Thus, the researchers concluded that the cross-validation strategy of applying two competing algorithms on the same data sets (Everitt et al. 2009) confirmed the cluster structure defined by the *k*-means procedure.

Cluster interpretation

K-means cluster analysis yielded a meaningful four-cluster solution for HP groups and a meaningful four-cluster solution for LP groups (Fig. 2). Three of the clusters were remarkably similar for both groups and demonstrated a progressive level of engagement in the problem-solving process.

Understanding the problem was the most pervasive activity for both groups. During these sessions the Problem Description screen was the only information accessed when the participants likely discussed and reflected on the problem. *Examining tasks* involved participants accessing primarily the five task screens. This pattern of activity was more common for the LP groups (75/268 sessions, 28%), than the HP groups (59/306 sessions, 19%). It is likely that during these sessions the students discussed task instructions. As with the first two clusters, the Problem Description screen and task screens were accessed during *working sessions*, but during this cluster of sessions participants also used tools (e.g., calculator) and explored information resources. The HP groups tended to engage in this type of activity more frequently (65/306 sessions, 21%) than did the LP groups (38/268 sessions, 14%).

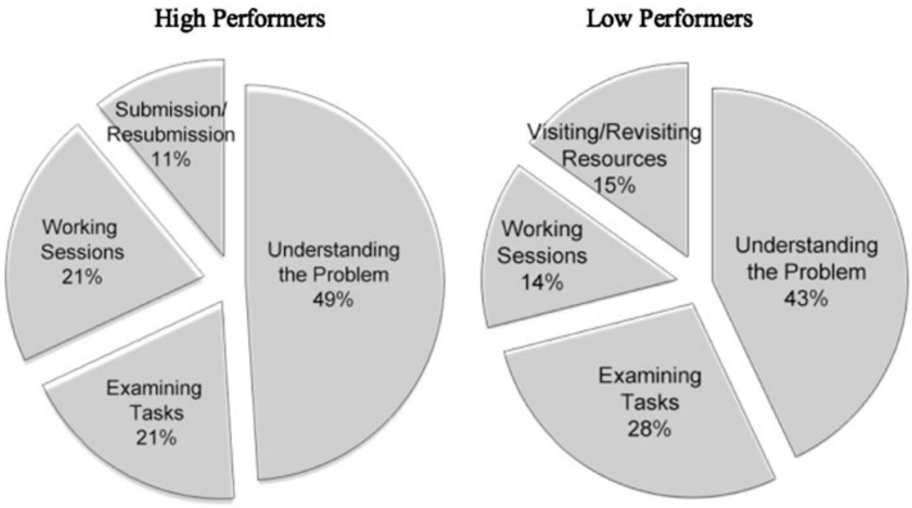

Fig. 2 HP and LP cluster comparison in the Niederhauser et al. (2007) study

The remaining two clusters were unique to HP and LP groups. During the *visiting/revisiting resources* cluster of sessions participants tended to visit (or revisit) information resources. Two of the four resources they visited most frequently did not contain information that was relevant to solving the problem. Only the LP groups exhibited this cluster pattern (40/268 sessions, 15%). *Submission/Resubmission* sessions involved participants reviewing the task screens; and submitting (or resubmitting) their responses. It is likely that participants had reflected on their solutions, and revisited the OLE to correct or update their work. Only the HP groups exhibited this cluster pattern (33/306 sessions, 11%).

In addition, visual inspection of the sequential, color-coded click-stream matrix of navigational decisions revealed that HP team members tended to start working early and continued working steadily until the deadline—logging into the OLE once every 2 or 3 of days. Students in LP teams began working on the problem closer to the deadline, and had more intensive periods of activity, frequently working late at night. This pattern likely limited the possibility for LP group participants to regulate their learning and apply metacognitive skills to reflect on their analysis of the problem and viable solutions.

Finally, ANOVA was used to examine the relative contribution of the variables (i.e., OLE nodes) to the formation of clusters. Univariate *F* tests were computed for each clustering variable as part of performing *k*-means clustering using Statistical Package for Social Sciences™. In this analysis, the *F* ratio is the ratio of cluster variance to error variance, with large *F* ratios identifying variables that are important for separating clusters. Specifically, the researchers were interested in examining which of the resources (relevant or irrelevant) were important in differentiating the clusters for HP and LP teams (Table 6).

As Table 6 demonstrates, three of four key resources contained information relevant to solving the problem were also the most important resources for separating HP groups' clusters. Conversely, the most important variables for separating clusters within the LP groups' data set were mostly irrelevant. Only one resource contained relevant information. Furthermore, two of the variables that were key resources for HP groups (and for solving the problem) were not important in differentiating LP group clusters.

Table 6 Key resources that contributed to the definition of the clusters in the Niederhauser et al. (2007) study

High-performers (HP teams)	Low-performers (LP teams)
Spouse ($F_{3,\ 302} = 174.43$) R	Banker ($F_{3,\ 264} = 307.36$) I
ARM features ($F_{3,\ 302} = 131.27$) R	Realtor ($F_{3,\ 264} = 242.49$) I
Background ($F_{3,\ 302} = 120.59$) R	Tax specialist ($F_{3,\ 264} = 176.41$) I
Realtor ($F_{3,\ 302} = 108.65$) I	Spouse($F_{3,\ 264} = 166.38$) R

R stands for relevant, *I* irrelevant

In this study, *k*-means clustering proved to be a valuable analytical method allowing researchers to explore a relatively large data set and determine that even though both HP and LP groups tended to engage in *examining the tasks* and in *working sessions*, LP teams were more likely to focus on the former, while HP teams—on the latter. Thus, HP groups spent more of their sessions actually reviewing the resources that provided information that could help them solve the problem. Clustering results also showed that LP student groups tended to use mostly irrelevant information—three of the four key resource pages that LP groups browsed most frequently did not contain any relevant information. In contrast, HP groups focused mostly on relevant information. Three of the four resources that were key in separating HP groups' clusters of browsing sessions were pertinent to solving the problem.

Caveats

While the examples described in this article demonstrate the benefits of using cluster analysis in educational technology research, like most statistical procedures cluster analysis has its limitations. The two major weaknesses of this group of methods are: (a) clustering algorithms will sometimes find structure in a dataset, even where none exists; and (b) results are sensitive to the algorithm used. It is not uncommon to obtain completely different results depending on the method chosen. These limitations highlight the importance of selecting the most appropriate algorithm based on the type of variables to be analyzed and size of the data file, and conducting cluster validity analyses using fusion coefficient plots and other clustering algorithms.

Another important issue has to do with the case proximity measure employed in most clustering algorithms—Euclidean distance, which is heavily affected by variables with large size or dispersion differences. If cases are being compared across variables that have very different variances, the Euclidean distances will likely be inaccurate. As such, it is important to standardize scores before proceeding with the analysis. Standardizing scores is especially important if variables were measured on different scales.

The results of cluster analysis can also be affected by the way in which the variables are ordered and the analysis is not always stable when cases are dropped. This occurs because selection of a case depends upon the similarity of one case to the cluster. Dropping one case or changing the ordering of variables can drastically affect the course in which the analysis progresses. Thus, selection of only the most important variables and input of these variables in a theoretically meaningful sequence as well as careful interpretation of clusters and profiling characteristics based on the purpose of the study and existing empirical evidence is essential in determining the number of clusters in the final solution.

Since there are so many different hierarchical and non-hierarchical clustering methods (not to mention proximity indices), it is natural to ask how stable any particular result is.

Ways of approaching this issue may be to compare the results yielded by different clustering algorithms and then average these results (Fielding 2007) and comparing clustering results to external variables—that is, variables that are not used for clustering.

Finally, because there are few guidelines about key clustering issues such as strategies to verify the final number of clusters, researchers should practice methodological diversity with multiple samples. Results of cluster analysis should be triangulated and tested in a variety of ways. For example, clustering results based on participants' navigational choices in an OLE can be analyzed and interpreted in light of the learning performance, prior knowledge in the content area, and metacognitive skills.

Conclusions

The use of cluster analysis to parse click-stream data and identify characteristics of more and less successful learning strategies in OLE's helps us glean insights into the nature of the cognitive and metacognitive processes that underlie knowledge acquisition and problem-solving, and provides prescriptive information to refine instructional practices. For example, the results obtained using *k*-means clustering and Ward's clustering in the studies described here suggest differences between the problem-solving and information-foraging activities of the high-performing and low-performing groups or individuals (Niederhauser et al. 2007; Toy 2008).

Cluster analysis in educational data mining offers the benefit of using the click-stream data that is collected automatically, without human intervention, by web servers and recorded in web activity logs. Cluster analysis avoids the limitations associated with examining and interpreting self-reported data and think-aloud protocols, which may interfere with the learner's primary cognitive activity (Clark 2010; Feldon 2007; Leighton 2004). Unlike methods relying on self-reports, analysis of click stream data can tap learning behaviors associated with automatized (i.e., unconscious) cognitive and metacognitive processes. Cluster analysis of click-stream data can be used as a compensatory method together with the more traditional self-report data collection and analysis techniques to provide researchers with a richer source of relevant data.

The key advantage of the clustering approach to data analysis is that it provides a fairly unambiguous profile of learning behavior, given a number of variables like the use of relevant versus irrelevant information resources, time spent on completing learning tasks, and number of task (re)submissions. Cluster analysis can also provide insights regarding what variables are most important in separating the clusters. Finally, since clustering techniques have their limitations, careful selection of the most suitable clustering algorithm and cluster verification analysis are essential to obtain valid and reliable results. When used appropriately, cluster analysis is a viable method of data mining to analyze learning behaviors in OLE's.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills: Sage Press.

Barab, S. A., Bowdish, B. E., & Lawless, K. A. (1997). Hypermedia navigation: profiles of hypermedia users. *Educational Technology Research and Development, 45*(3), 23–42.

Belland, B. R., French, B., & Ertmer, P. A. (2009). Validity and problem-based learning research: a review of instruments used to assess intended learning outcomes. *Interdisciplinary Journal of Problem-based Learning, 3*(1), 59–89.

Belland, B. R., Glazewski, K. D., & Richardson, J. C. (2010). Problem-based learning and argumentation: testing a scaffolding framework to support middle school students' creation of evidence-based arguments. *Instructional Science,*. doi:10.1007/s11251-010-9148-z.

Clark, R. E. (2010). Cognitive and neuroscience research on learning and instruction: recent insights about the impact of non-conscious knowledge on problem solving, higher order thinking skills and interactive cyber-learning environments. *Presented at the International Conference on Education Research (ICER)*, Seoul.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456–473.

Donner, A., & Koval, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics, 36*(1), 19–25.

Everitt, B. S., Landau, S., & Leese, M. (2009). *Cluster analysis* (4th ed.) London: Arnold.

Facione, P. A., & Facione, N. C. (1994). *Holistic critical thinking scoring rubric*. Millbrae: California Academic Press.

Feldon, D. F. (2007). Implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review, 19*(2), 91–110.

Fielding, A. H. (2007). *Cluster and classification techniques for the biosciences*. Cambridge: Cambridge University Press.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: a meta-analysis from the angle of assessment. *Review of Educational Research, 75*(1), 27–61.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a survey. *ACM Computing Surveys, 31*, 264–323.

Jeong, H., & Hmelo-Silver, C. E. (2010). Productive use of learning resources in an online problem-based learning environment. *Computers in Human Behavior, 26*, 84–99.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development, 45*(1), 65–94.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85.

Kim, M. C., & Hannafin, M. J. (2011). Scaffolding problem solving in technology-enhanced learning environments (TELEs): bridging research and theory with practice. *Computers & Education, 56*, 403–417.

Kumsaikaew, P., Jackman, J., & Dark, V. J. (2006). Task relevant information in engineering problem solving. *Journal of Engineering Education, 95*, 227–239.

Lawless, K., & Kulikowich, J. (1996). Understanding hypertext navigation through cluster analysis. *Journal of Educational Computing Research, 14*(4), 385–399.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: the collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.) *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkley: University of California Press.

Milligan, G. W. (1980). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16*, 379–407.

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: clustering methods. *Applied Psychological Measurement, 11*, 329–354.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.) *Proceedings of the Twentieth International Conference on Very Large Databases* (pp. 144–155). Santiago: Morgan Kaufmann.

Niederhauser, D. S., Antonenko, P., Ryan, S., Jackman, J., Ogilvie, C., Marathe, R., & Kumsaikaew, P. (2007). *Solution strategies of more and less successful problem solvers in an online problem-based learning environment.* Presented at the annual conference of the American Educational Research Association. Chicago, IL.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* London: Academic Press.

Norušis, M. (2005). *SPSS 13.0 statistical procedures companion*. Englewood Cliffs:Prentice Hall.

Ryan, S., Jackman, J., Kumsaikaew, P., Dark, V., & Olafsson, S. (2007). Use of information in collaborative problem solving. In D. H. Jonassen (Ed.), *Learning to solve complex, scientific problems* (pp. 187–204). Mahwah, NJ: Lawrence Erlbaum Associates.

Schrader, P. G., & Lawless, K. A. (2007). Dribble files: methodologies to evaluate learning and performance in complex environments. *Performance Improvement, 46*(1), 40–48.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin, 38*, 1409–1438.

Stevens, R. H. (2007). Quantifying student's scientific problem solving efficiency and effectiveness. *Cognition and Learning, 5*, 325–337.

Toy, S. (2008). *Online ill-structured problem-solving strategies and their influence on problem-solving performance.* Unpublished doctoral dissertation, Iowa State University, Ames, IA.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association, 58*(301), 236–244.

Webb, A. (2002). *Statistical pattern recognition*. Hoboken: John Wiley.

Pavlo D. Antonenko is an Assistant Professor of Educational Technology in the School of Educational Studies at Oklahoma State University.

Serkan Toy is the Director of Evaluation and Program Development at Children's Mercy Hospital, Kansas City.

Dale S. Niederhauser is an Associate Professor in the Department of Curriculum and Instruction at Iowa State University.