The Usefulness of Log Based Clustering in a Complex Simulation Environment

Samad Kardan[1], Ido Roll[2], and Cristina Conati[1]

[1] Department of computer Science, University of British Columbia
[2] Centre for Teaching, Learning, and Technology, University of British Columbia
2329 West Mall, Vancouver, BC, V6T 1Z4, Canada
skardan@cs.ubc.ca

Abstract. Data mining techniques have been successfully employed on user interaction data in exploratory learning environments. In this paper we investigate using data mining techniques for analyzing student behaviors in an especially-complex exploratory environment, with over one hundred possible actions at any given point. Furthermore, the outcomes of these actions depend on their context. We propose a multi-layer action-events structure to deal with the complexity of the data and employ clustering and rule mining to examine student behaviors in terms of learning performance and effects of different degrees of scaffolding. Our findings show that using the proposed multi-layer structure for describing action-events enables the clustering algorithm to effectively identify the successful and unsuccessful students in terms of learning performance across activities in the presence or absence of external scaffolding. We also report and discuss the prominent behavior patterns of each group and investigate short term effects of scaffolding.

Keywords: Educational Data Mining, Clustering, Scaffolding.

1 Introduction

A major component of any Intelligent Tutoring System (ITS) is the learner model (see [1, 2]). The learner model is in charge of estimating the learners' proficiency and adapting the instruction accordingly. Building a learner model is especially challenging in exploratory environments and ill-defined domains in which students' responses do not have a well-defined accuracy. These environments and domains include games (e.g., Newton's Playground [3]), simulations (e.g., [4]), open-ended activities (e.g., [5, 6]), and meta-cognitive tutoring (e.g., The Help Tutor [7]), to name a few. The challenge of modeling learners becomes even more acute in complex environments, where students can engage in a variety of behaviors. One solution in these environments has been to group similar actions together. For example, in Betty's Brain [5], an environment that supports learning by drawing causal diagrams, all actions that involve editing the diagram are labeled as Edit Map. A further complication is introduced in environments which are used as platforms with a large variety of activities.

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 168–177, 2014.

The current work applies a clustering approach to learning in an open-ended physics simulation which enables complex behaviors and is used with diverse activities. Specifically, we address the following research questions:

1. How can a clustering approach be applied to complex data from an exploratory learning environment?
2. What can the data mining tell us about the relationship between student behaviors in the environment, their learning, and the given activity?

We first discuss related work on clustering and describe the learning environment. We then describe the experimental design and data handling. Last, we describe the clusters and associated rules, and discuss their meaning.

2 Related Work

In the field of Educational Data Mining, clustering has been applied to different applications for discovering groups of similar users. Relevant to our work, in problem solving activities, clustering has been used to find better parameter settings for models that assess student knowledge [8], as well as discovering student learning tactics [9]. In [10] clustering and rule mining were successfully used to investigate student behaviors in an interactive simulation. However, to date, clustering and rule mining were typically applied to data from relatively constrained environments.

The current work extends the scope of using clustering by analyzing data from a high-complexity environment, the DC Circuit Construction Kit simulation, which is part of the PhET project. PhET (phet.colorado.edu [11]) is a freely-available and widely-used suite of simulations in different science and math topics. These 120 simulations are used over 45,000,000 times a year by a community of middle-school to college students. Figure 1 shows the DC Circuit Construction Kit, one of the more popular simulations of the PhET family[1]. In this specific simulation, students explore basic properties of DC circuits by connecting wires, light bulbs, resistors, switches, and measurement instruments, on a virtual test bed.

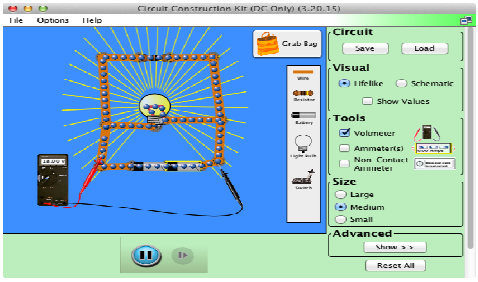

Fig. 1. The DC Circuit Construction Kit simulation

[1] http://phet.colorado.edu/en/simulation/circuit-construction-kit-dc

Several microworlds and simulations offer detailed scaffolding and explicit feedback (e.g., using cognitive tools such as hypotheses builders [4, 12]). However, PhET Simulations attempt to stay closer to an authentic inquiry environment, and thus offer neither explicit scaffolding nor explicit feedback. PhET Simulations are used as open-ended platforms for investigation. Teachers and instructors who assign the simulations to their students create their own activities, usually on paper. As a result, PhET simulations are used in a large variety of contexts and populations, using a large variety of activities. While some activities include very detailed directions for students, other activities let students explore the topic without much guidance.

3 User Study

One hundred students from first-year physics courses in a large Canadian university volunteered for a study which took place outside their normal classroom hours. The study included two activities on the topic of DC circuits, each of which took 25 minutes. The first activity focused on the effect of combining light bulbs in different arrangements. The second activity focused on the effect of combining resistors with different resistances. As PhET simulations are typically used with a large variety of activities, students were assigned to one of the two following conditions for the first activity: Low Scaffolding (LS) and High Scaffolding (HS). Students in the LS condition received only the general learning goal and a general recommendation to explore several light bulbs on the same loop, on different loops, and a combination of the two. Students in the HS condition received the same learning goal and recommendation, and in addition, were given diagrams, tables, and guiding questions. The diagrams instructed students which circuits to build; the tables asked them to document the parameters of the different circuits; and the guiding questions asked students to reflect, compare, and contrast the different circuits. The HS condition was modeled after the recommended activities for this context by the PhET project team. The study began with a short pre-test, following which students were randomly assigned to either the LS or the HS version of the light-bulb activity (see Figure 2). All students received a LS activity for the second activity on resistors. This allows us to evaluate how the same students alter their behaviors based on the given scaffolding. Last, a post-test on both activities was given, together with a survey. Three students had a perfect score on the pre-test and were removed from the analysis.

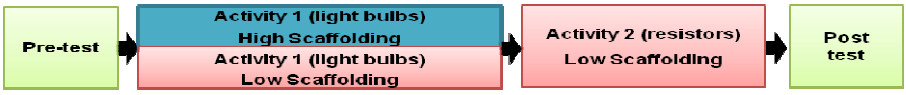

Fig. 2. The study structure

4 Analysis of User Actions

Students in the simulation work with a variety of components that include batteries, wires, light bulbs, resistors, and measurement instruments such as ammeters and

voltmeters. Overall, there are 124 different types of actions that students can perform at each moment. These actions include adding, moving, connecting, splitting, and removing components, as well as changing the attributes of components (such as resistance). Additional actions relate to the interface (such as changing views or zooming in and out), or the simulation itself (such as pausing or resetting the simulation). In addition, the outcomes of these actions depend on the state of the simulation. For example, a student will get different feedback depending on whether a testing instrument is connected to the circuit or not when s/he is changing the resistance of a resistor. This makes it quite difficult to relay on the analysis of user actions alone for the purpose of understanding the learning performance of users. In fact, clustering students according to the raw data did not support inferences about learning, as explained further below. Thus, we have constructed a multi-layer structure to capture the context of each action. In this section we introduce this structure and briefly describe the method used for behavior discovery.

In order to go beyond the raw action types recorded in the log files, we define an "action-event" as the entity that is formed by a combination of the user action and relevant contextual information. Each action-event consists of a user action (not to be confused with raw action types), the component involved in that action, the family that this specific action in the given context belongs to, and finally, outcome of the action (Figure 3). Overall, we have identified 226 action-events. Notably, these features do not create a hierarchy. For example, *joining* (action) a *wire* (component) may lead to a *current-change* (outcome) in some cases when *revising* a circuit (family), and to *no-change* (outcome) when *organizing* the circuit (family).

It is important to note that by creating this structure we have added semantic information to the data. Converting the data is done automatically by a parser which keeps track of the context (e.g., if a component is connected to the circuit) and based on over 100 conditions, assigns a value from each layer to each line of log records.

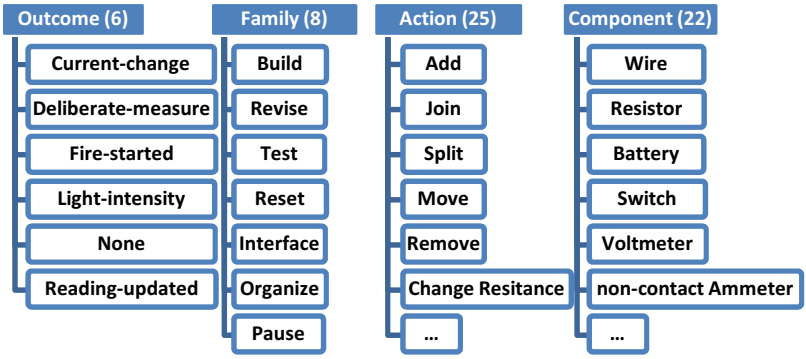

Fig. 3. The four-layer structure of the action-events

For each of the action-events we calculate three features: (i) frequency of the action-event (proportion of the number of times each action-event is used compared to total count of all action-events) denoted by _f, (ii) mean, and (iii) standard deviation of the time spent before each action-event (denoted by _m and _s respectively).

Generating features using different number of layers (e.g., Outcome only, Outcome and Family, etc.) would result in different feature-sets which contain different levels of detail about the action-events. Using only the outcome layer would generate 18 features while using all four layers of the action-event structure would result in 678 features. Interestingly, including only a subset of layers in the cluster analysis did not lead to meaningful results. Thus, we have clustered students based on all 4 layers of information (and 678 features). This highlights the importance of the semantic information that was added to the data in the preparation phase. In order to model the behaviors of the students we use the user modeling framework proposed in [10] for discovering groups of students who showed similar interaction behaviors as well as finding the representative behaviors of each group. Specifically, we look at whether the identified clusters can detect differences with regard to students' learning outcomes and the given activity (high vs. low scaffolding). The mentioned user modeling framework is used for providing support during interaction with an interactive simulation, personalized to each student's needs [10]. We will only focus on the *Behavior Discovery* phase of the framework in this paper (see [10, 13] for more details on the complete framework).

In *Behavior Discovery* user interaction data is first pre-processed into feature vectors representing each user. In our case, each vector includes the (i) frequency, (ii) mean, and (iii) standard deviation of time before each action-event (thus, 226 action-events × 3 measures per action-event = 678 features). Then, these vectors are clustered in order to identify users with similar interaction behaviors. The distinctive interaction behaviors in each cluster are identified via association rule mining [14]. This process extracts the common behavior patterns in terms of class association rules in the form of $X \rightarrow c$, where X is a set of feature-value pairs and c is the predicted class label for the data points where X applies. A confidence value is assigned to each rule calculated as the proportion of cases where X is true and class label is c over all cases where X is true. We use the Hotspot algorithm from the Weka data mining toolkit [15] for association rule mining.

In order to associate behaviors to learning performance, it is first necessary to establish how the user groups generated by clustering relate to learning. If learning performance measures are available, then we can assign a label to each cluster by comparing the average learning performance of the users in that cluster with the performance of the users in the other clusters. This is the approach we successfully adopted in [10] and will be used in this paper (see [16] for an alternative approach and related discussion). Introduction of the multi-layer action-events in this work enables us perform the clustering at different levels with different degrees of details and find the right amount of details that describes the user behaviors effectively.

5 Results and Discussion

As described in the previous section, we apply clustering on user interaction data to find groups of users in terms of how they interacted with the simulation. Similar to [10], we are interested to see if the discovered clusters of users correspond to different

levels of learning performance. However, unlike [10], employing user actions alone (i.e., either the action layer or the combination of action and component layers in the action-events structure) did not lead to meaningful results. We attribute this to the complexity of the interactions in the simulation under study here compared to the one used in [10]. Thus, we use the full 4-layer action-events structure in our analyses.

In addition to learning performance, we are interested in finding any difference in distribution of the students in the HS vs. LS conditions between the discovered clusters. Due to performing two simultaneous comparisons on the data, α for the tests (described below) is adjusted to 0.025 using Bonferroni correction. Furthermore, we will discuss the association rules describing the behaviors of users in each cluster. Our analysis first focuses on Activity 1 (A1) and Activity 2 (A2) individually and then we compare the results between the two activities.

For each activity, the optimal number of clusters is the lowest number suggested by C-index, Calinski and Harabasz [17], and Silhouettes [18] measures of clustering validity. The summary statistics of the clusters discovered for A1 and A2 are presented in Table 1 (from left the columns describe: the activity, optimal number of clusters, cluster labels (HL and LL are described later), and for each cluster: number of students, average of the standardized pre-test and post-test scores, and number of students from the HS and LS conditions). When performing clustering we faced cases in which the final clusters had only one member (singletons), therefore we had to remove the outlier user forming the singleton and repeat the clustering. This process reduced the number of students to 86 for A1 and 94 for A2.

Table 1. Summary statistics of the clusters for each activity

Activity	Number of clusters	Cluster Label	Overall Number of students	Average Pre-test Performance (SD)	Average Post-test Performance (SD)	Number of HS students	Number of LS students
A1	4	1	3	-0.9 (.1)	-1.2 (.3)	1	2
		2	3	0.7 (1.1)	1.2 (.4)	0	3
		3 (LL_1)	22	0.2 (.9)	-0.5 (1.1)	2	20
		4 (HL_1)	58	0.0 (1.1)	0.2 (.9)	42	16
A2	3	1 (LL_2)	21	-0.2 (.9)	-0.5 (.8)	11	10
		2 (HL_2)	65	0.0 (1.0)	0.2 (1.0)	36	29
		3	8	0.2 (1.2)	-0.3 (1.2)	1	7

In order to compare the learning performance of the students in each cluster we use the standardized post-test scores of the students while using pre-test scores as a covariate in our analysis (using ANCOVA). For the post-hoc analysis, the p values are again adjusted using Bonferroni correction. We apply χ^2 tests in order to see whether the distribution of students in the LS and HS conditions for the discovered clusters is different from the even distribution of the two conditions in the whole sample.

5.1 Analyzing Behaviors in Activity 1

There is a significant difference in post-test performance of the students in the four clusters ($p = .001$) with a large effect size ($\eta^2 = 0.181$) after controlling for the pre-test performance. Since the first two discovered clusters are very small ($n = 3$), we exclude

them from post-hoc analysis. For clusters 3 and 4 there is a significant difference in learning performance of students ($p = 0.006$). The students in cluster 4 are doing more than half a standard deviation better in post-test (estimated mean difference is 0.718) while there is no significant difference in pre-test scores. We will refer to the cluster 3 as Lower Learning (LL_1) and cluster 4 as Higher Learning (HL_1).

A χ^2 test on distribution of students from the LS and HS conditions shows a significant difference with the expected distribution for the four clusters discovered for A1 ($p < .001$). The same test performed only on the LL_1 and HL_1 clusters also provides similar results ($p < .001$). The majority (over 90 percent) of LL_1 students are from the LS condition. While HL_1 cluster is somewhat more balanced in terms of HS to LS ratio, it comprises over 90 percent of all students in the HS condition. The concentration of the students from the HS condition in a single cluster shows that the scaffolding provided to them encouraged them to behave similarly.

The output of association rule mining process for the LL_1 and HL_1 clusters of A1 is shown in Table 2. Rules that applied to at least 50 percent of the members of the cluster and achieved a confidence level over 0.6 were selected. Each part of the association rules is in form of a feature and a corresponding value assigned to it, for example "None.Build.join.resistor_f = Low" indicates that the (f)requency of using the resistor component when building the circuit was low.

Table 2. Selected Rules for A1 (confidence values in brackets)

A1 Cluster 3 (LL_1) 4 rules overall:
1. Reading_updated.Test.endMeasure.nonContactAmmeter_f = Low [0.625]
2. Reading_updated.Test.endMeasure.nonContactAmmeter_f = Low AND None.Build.join.seriesAmmeter_m = High [1.0]
3. Reading_updated.Test.endMeasure.nonContactAmmeter_f = Low AND None.Revise.remove.lightBulb_m = Medium [1.0]
A1 Cluster 4 (HL_1) 6 rules overall:
1. Reading_updated.Test.endMeasure.nonContactAmmeter_f = High [0.919]
2. Reading_updated.Test.endMeasure.nonContactAmmeter_f = High AND None.Build.join.resistor_f = Low [0.971]
3. Deliberate_measure.Test.startMeasure.nonContactAmmeter_f = High [0.856]

Rules 1-3 for the LL_1 cluster (Table 2) show that LL_1 students did not use one of the main measurement devices, the nonContactAmmeter, frequently enough. Rules 2 and 3 include additional conditions which are hard to interpret at this point. The HL_1 cluster includes mainly students in the HS condition. Thus, it is of no surprise that all selected rules include a frequent use of the nonContactAmmeter, which was required in order to fill out the tables successfully. Rule 2 also describes infrequent addition of a resistor. This behavior makes sense, as A1 focuses on light bulbs, and not resistors.

5.2 Analyzing Behaviors in A2

Similar to A1, there is a significant difference in post-test performance of the students in the three clusters discovered for A2 ($p = .011$) with a medium effect size ($\eta^2 = 0.096$)

after controlling for the pretest performance. The post-hoc analysis for A2 shows a significant difference in learning performance between clusters 1 and 2 (estimated mean difference in post-test is 0.646). Cluster 3 was excluded due to its small size (n =8). Similar to A1, there is no significant difference in pre-test scores between clusters 1 and 2. We will refer to the cluster 1 as Lower Learning (LL_2) and cluster 2 as Higher Learning (HL_2). Unlike A1, the χ^2 test for A2 does not show a significant difference in distribution of students to clusters by conditions. This means that the cluster analysis was not able to identify any differences among students who received different levels of scaffolding prior to the task (unlike A1, all students received the same scaffolding in A2).

Table 3. Selected Rules for A2 (confidence values in brackets)

A2 Cluster 1 (LL_2) 13 rules overall:
1. None.Build.join.lightBulb_m = Average [0.923]
2. Current_change.Revise.join.wire_f = Low AND None.Pause_f = Low AND Current_change.Revise.join.resistor_m = Low [0.778]
3. Current_change.Revise.join.wire_f = Low AND None.Pause_f = Low AND None.Test.endMeasure.nonContactAmmeter_f = Low [0.75]
A2 Cluster 2 (HL_2) 3 rules overall:
1. Current_change.Revise.join.wire_f = High [0.957]
2. None.Build.join.lightBulb_m = Low [0.853]
3. None.Build.join.lightBulb_m = Low AND Current_change.Revise.join.wire_f = High [0.978]

The selected rules extracted from LL_2 and HL_2 clusters are shown in Table 3 (with the same selection criteria used for A1). The second rule for LL_2 talks about students who do not revise circuits by adding wires, do not pause to study their outcomes, and last, when joining resistors to existing circuits, they do so rapidly. These three conditions suggest that students in the LL_2 cluster, test relatively simple circuits (without adding wires and loops to existing circuits), and do so hastily – without taking sufficient time to reflect. Rule 3 shared many of these characteristics. Students join few loops to working circuits, take only few pauses, and use one of the instrument devices, the nonContactAmmeter, only rarely. Put together, rules 2 and 3 of the LL_2 cluster match current theories of learning. To learn, students should take time to reflect, compare similar circuits, and measure the outcomes of their methods. Students in this cluster only rarely engaged in these behaviors. Notably, the rules talk about specific aspects of extending circuits and using measurement instruments (e.g., nonContactAmmeter is included, but not Voltmeter). Additional data is required to understand these characteristics of the rules.

The rules for the HL_2 cluster are at sharp contrast with the LL_2 cluster. As the first rule shows, these students often extended working circuits by adding loops. The last two rules talk about students who take little time before adding light bulbs. These rules are somewhat surprising, as the activity was about resistors and not about light bulbs. Additional data is required before these rules can be interpreted.

5.3 Comparing A1 and A2

Comparing the discovered rules for A1 and A2 helps us to understand the behaviors that are specific to an activity vs. the ones that transfer across all activities and levels of scaffolding. Additionally, such comparison can highlight the advantages and limitations of using clustering to identify learners in a complex simulation.

Overall, the rules for the four clusters show one clear trend that repeats across activities and levels of scaffolding: A frequent use of the measurement devices, and especially the nonContactAmmeter, is associated with higher learning. The converse is true, too – an infrequent use of the instrument is associated with low learning.

While the trend can be seen in three of the four clusters, it is notable that the rules themselves are dissimilar. It may be that our search space included too many similar features, so that alternative features that hold similar meanings appeared in different rule sets. An alternative explanation is that the behaviors as captured by user actions is dependent on the task, which means although users with similar learning performance tend to show similar behaviors, these behaviors vary from task to task. In this case, transferability of cluster-based user models in simulation environments may be limited. We plan to collect additional data from other simulations to evaluate the transferability of the identified behaviors. A statistical analysis of effects of changes in scaffolding levels between A1 and A2 is presented in [19].

6 Conclusions

We clustered students who worked with two activities and two levels of scaffolding in an open-ended simulation. Our results show that the clusters gave us meaningful information about learning, but only when the raw data was augmented with semantic, contextual data.

Analysis of the clusters also revealed several interesting patterns in the data. All students who received high scaffolding were clustered in the same group, suggesting that the scaffolding directed them to a certain behavioral style. Notably, students who received low levels of scaffolding were distributed across four clusters.

In addition, one main behavior was associated with better learning across activities: the frequent use of measurement devices. At the same time, while the interpretation of the rules may be similar, the actual rules are different, and thus their transferability across activities should be further studied. For example, it is not yet clear why only certain aspects of testing appear in the clusters, and not others. Furthermore, some rules remain hard to interpret. It may be that shrinking the feature list without losing semantic information may lead to more consistent rules across activities.

Acknowledgements. This work is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) grant #430-2012-0521 and by the Betty and Gordon Moore Foundation. We would like to thank the PhET project team for their assistance.

References

1. Koedinger, K.R., Corbett, A.T.: Cognitive tutors: Technology bringing learning science to the classroom. In: The Cambridge Handbook of the Learning Sciences, pp. 61–78 (2006)
2. VanLehn, K.: The Behavior of Tutoring Systems. International Journal of Artificial Intelligence in Education 16, 227–265 (2006)
3. Shute, V.J., Ventura, M., Kim, Y.J.: Assessment and Learning of Qualitative Physics in Newton's Playground. The Journal of Educational Research 106, 423–430 (2013)
4. Gobert, J.D., Pedro, M.A.S., Baker, R.S.J.d., Toto, E., Montalvo, O.: Leveraging Educational Data Mining for Real-time Performance Assessment of Scientific Inquiry Skills within Microworlds. JEDM - Journal of Educational Data Mining 4, 111–143 (2012)
5. Leelawong, K., Biswas, G.: Designing Learning by Teaching Agents: The Betty's Brain System. International Journal of Artificial Intelligence in Education 18, 181–208 (2008)
6. Roll, I., Aleven, V., Koedinger, K.R.: The Invention Lab: Using a Hybrid of Model Tracing and Constraint-Based Modeling to Offer Intelligent Support in Inquiry Environments. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 115–124. Springer, Heidelberg (2010)
7. Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R.: Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. Learning and Instruction 21, 267–280 (2011)
8. Gong, Y., Beck, J.E., Ruiz, C.: Modeling Multiple Distributions of Student Performances to Improve Predictive Accuracy. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 102–113. Springer, Heidelberg (2012)
9. Shih, B., Koedinger, K.R., Scheines, R.: Unsupervised Discovery of Student Strategies. In: Proceedings of the 3rd Intl. Conf. on Educational Data Mining, pp. 201–210 (2010)
10. Kardan, S., Conati, C.: A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces. In: Proc. of the 4th Int. Conf. on Educational Data Mining, Eindhoven, The Netherlands, pp. 159–168 (2011)
11. Wieman, C.E., Adams, W.K., Perkins, K.K.: PhET: Simulations That Enhance Learning. Science 322, 682–683 (2008)
12. De Jong, T., Van Joolingen, W.R.: Scientific discovery learning with computer simulations of conceptual domains. Review of Educational Research 68, 179–201 (1998)
13. Kardan, S.: Data mining for adding adaptive interventions to exploratory and open-ended environments. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 365–368. Springer, Heidelberg (2012)
14. Zhang, C., Zhang, S.: Association rule mining: Models and algorithms. Springer, Heidelberg (2002)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11, 10–18 (2009)
16. Kardan, S., Conati, C.: Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) UMAP 2013. LNCS, vol. 7899, pp. 215–227. Springer, Heidelberg (2013)
17. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179 (1985)
18. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)
19. Roll, I., Yee, N., Briseno, A.: Students' Adaptation and Transfer of Strategies Across Levels of Scaffolding in an Exploratory Environment. In: Proc. of the 12th Intl. Conf. on Intelligent Tutoring Systems (2014)