Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning

Bondareva[1] D., Conati[1] C., Feyzi-Behnagh[2] R., Harley[2] J., Azevedo[2] R., Bouchet[2] F.

[1] University of British Columbia, [2] McGill University
{bondaria,conati}@cs.ubc.ca, {reza.feyzibehnagh,
jason.harley}@mail.mcgill.ca, {roger.azevedo,
francois.bouchet}@mcgill.ca

Abstract. In this paper, we explore the potential of gaze data as a source of information to predict learning as students interact with MetaTutor, an ITS that scaffolds self-regulated learning. Using data from 47 college students, we show that a classifier using a variety of gaze features achieves considerable accuracy in predicting student learning after seeing gaze data from the complete interaction. We also show promising results on the classifier ability to detect learning in real-time during interaction.

Keywords: student modeling, eye-tracking, self-regulated learning

1 Introduction

Student modeling is known to be a difficult problem because there is often a large gap between students behaviors observable by an Intelligent Tutoring Systems (ITS) and the students' states and processes that the ITS needs to model in order to provide personalized instruction. One approach that is being explored to address this problem is to investigate the use of sensors that can help reduce the gap between the student's relevant states and what an ITS can observe about them].

This paper contributes to this body of research by exploring the value of eye-tracking data (also referred to as *gaze data* from now on) in assessing student learning during interactions with MetaTutor, a multi-agent ITS that scaffolds self-regulated learning (SRL) while students study material on the human circulatory system [2]. This research is part of a larger endeavor to understand and model the relations among affect, cognition and meta-cognition in learning with MetaTutor, by leveraging multi-channel data sources including think-aloud protocols, eye-tracking, human-agent dialogue, log-file, embedded quizzes, galvanic skin response, and face recognition. We decided to start by focusing on gaze data, because there is already evidence that it can provide useful information on all the student modeling dimensions we are interested in: cognitive [e.g. 3–5], metacognitive [6] and affective [7, 8]. We start by investigating if and how gaze data can be used to predict learning in MetaTutor because tracking whether a student is learning is important for a tutoring agent to decide when to provide personalized instruction.

adfa, p. 1, 2011.

The main contribution of this paper are results showing that gaze data can indeed be a useful source of information to predict student learning with MetaTutor. This result is especially important because it does not exist in isolation. Similar research using a different type of learning environment (an interactive simulation to support learning by exploration), also found that gaze data was a good predictor of student learning [4]. Therefore, the results reported here contribute to confirm the importance of gaze data as a predictor of learning across different types of learning environments, that can be leveraged for providing real-time personalized support.

In the rest of the paper, Section 2 summarizes related work. Section 3 describes MetaTutor, and the study that generated the data used in this paper. Section 4 describes how we trained classifiers on eye-tracking data to predict student learning. Section 5 reports the classification results, followed by conclusions and future work.

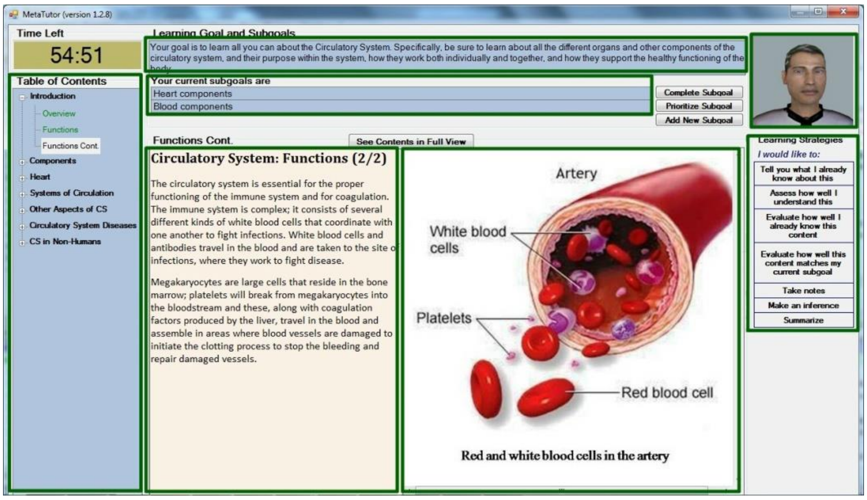
2 Related Work

Eye-tracking has been the focus of increasing interest in student modeling, as a way to track user's states and processes at the cognitive, meta-cognitive and affective level. At the cognitive level, in addition to [4], discussed above, Gluck and Anderson [5] used gaze data to assess student problem-solving behaviors within an ITS for algebra, including attention shifts, problem disambiguation and processing of error messages. Sibert et al. [9] explored gaze tracking to assess reading performance in a system for automated reading remediation that provides support if a user gaze patterns indicate difficulties in reading a word. D'Mello et al. [3] show that tracking a student's attention toward a Pedagogical Agent in a dialogue-based ITS and generating prompts to guide this attention, improves student learning. At the meta-cognitive level, [6] shows that using gaze data improves a student model's ability to track students' self-explanation behaviors (i.e. generating explanations to one-self to improve one's understanding), and consequent learning. At the affective level, Qu and Johnson [7] leveraged gaze data to assess student motivation in an ITS for teaching engineering skills. Muldner et al. [8] looked at pupil dilation to detect relevant student affective and meta-cognitive states during the interaction with an ITS that supports analogical problem solving.

In the context of modeling students' SRL processes, so far researchers have mainly relied on mining action logs. For instance, Kinnebrew and Biswas [10], used sequence mining on action logs to identify effective and ineffective behaviors in students interacting with Betty's Brain, an ITS for scaffolding SRL via teachable agents. Bouchet et al. [11] performed similar work with MetaTutor, the ITS used in this paper. Sabourin et al. [12], mined both actions and students self-reports on their affective states for the early prediction of SRL processes during interaction with Crystal Island, a narrative-based and inquiry-oriented serious game for science.

3 Meta Tutor Study

MetaTutor is an adaptive hypermedia learning environment which includes 38 pages of text and diagrams, organized by a table of contents displayed in the left pane of the environment (see Figure 1[1]) [2]. Text and diagrams are displayed separately in the two central panels of the interface. In addition to providing structured access to relevant content, MetaTutor also includes a variety of components designed to scaffold learners' use of SRL processes and their learning of science topics, such as the human circulatory system. Four pedagogical agents (PAs) are displayed in turn in the upper right-hand corner of the environment. Each agent provides spoken prompts and feedback on various SRL processes. For example, one PA assists the student in establishing two learning sub-goals related to the overall learning goal for the session (see top horizontal panel in Figure 1, with sub-goal panel right below). The shading of the sub-goal bars in the corresponding panel shows the student's current progress towards completing that sub-goal as the interaction proceeds.



Your goal is to learn all you can about the Circulatory System. Specifically, be sure to learn about all the different organs and other components of the circulatory system, and their purpose within the system, how they work both individually and together, and how they support the healthy functioning of the body.

Circulatory System: Functions (2/2)

The circulatory system is essential for the proper functioning of the immune system and for coagulation. The immune system is complex; it consists of several different kinds of white blood cells that coordinate with one another to fight infections. White blood cells and antibodies travel in the blood and are taken to the site o infections, where they work to fight disease.

Megakaryocytes are large cells that reside in the bone marrow; platelets will break from megakaryocytes into the bloodstream and these, along with coagulation factors produced by the liver, travel in the blood and assemble in areas where blood vessels are damaged to initiate the clotting process to stop the bleeding and repair damaged vessels.

Fig. 11. Sample MetaTutor interface

Other SLR processes supported by the PAs include taking notes, writing summaries of the viewed content, evaluating one's current understanding, etc., and they can be initiated via the learning strategy palette displayed in the right interface pane.

A study was conducted in 2012 with the goal of collecting multi-channel data to examine the role of cognitive, metacognitive, and affective processes during learning with MetaTutor [13]. The study included two conditions: one (adaptive) in which the Meta-Tutor's PAs provided prompts and feedback adapted to each student's performance; another (non-adaptive) in which prompts and feedback were generic. The

[1] The boxed areas in the figure indicate Areas of Interest used for eye-tracking, as described in Section 4.

study consisted of two sessions. In the first, participants (university students who were randomly assigned to the two study conditions) completed a pre-test on the circulatory system and demographics questionnaires. The second session started with the calibration of apparatuses, including a Tobii T60 eye-tracker[2]. Next, each participant watched video tutorials on SRL processes and related interface functionalities, and was then asked to set two sub-goals for the session. After that, the participant interacted with MetaTutor for one hour, followed by a post-test. In this paper, we focus on exploring whether the gaze data collected in the study can be leveraged to predict student learning, as measured by the study pre- and post-tests The next section describes how we built gaze-based classifiers to achieve this goal.

4 Classification experiments

For the current work, we used 64 participants with eye-tracking data collected in the study described above. For the subsequent analysis, we focused on data related to students interacting with MetaTutor, excluding parts of the interaction during which participants were watching video tutorials.

The Tobii T60 eye-tracker used in the study is embedded in a LCD screen and thus it is non-intrusive, because it does not constrain participants' movements. While this is a great asset, the down side is that the collected data can be noisy and needs validation. One source of noise is due to participants looking away from the screen, which the eye-tracker interprets as invalid data. These look-away events happen when there are pauses in the session or when students use one of the tools provided by MetaTutor to submit typed text to the system (e.g., while writing summaries on the material seen so far)[3]. We created scripts to parse the study action-log files for these events and remove the corresponding segments from gaze data.

A second source of noise is due to actual eye-tracking errors that generate invalid gaze samples. Participants with gaze data that include too many invalid samples need to be discarded because the missing data makes it difficult to draw reliable inferences from these participants' attention patterns. To account for this source of noise, we adopted the data validation process discussed in [4], which essentially discards participants that have less than 80% valid samples overall, as reported by the eye-tracker (after removing known look-away events). The validation process resulted in discarding 16 users, leaving a total of 48 for the actual classification study.

4.1 Gaze features

An eye-tracker captures gaze information in terms of *fixations* (i.e., maintaining gaze at one point on the screen) and *saccades* (i.e., a quick movement of gaze from one fixation point to another). Gaze patterns are further defined by measures that represent

[2] Precision/accuracy for X are 0.4-0.5°/0.18-0.36°, for Y are 0.4-0.6°/0.18-0.30°. the smallest trackable size of Area of Interest is 30 by 30 pixels.

[3] These activities can be reliability tracked using action logs, and will be included as part of our future work.

gaze direction, including *absolute path angles* (i.e., the angle between a saccade and the horizontal) and *relative path angles* (i.e., the angle between two consecutive saccades). Following the approach suggested in [14], and followed in [4], we computed a large variety of features based on raw gaze data. These are divided into two types. The first type was generated by applying summary statistics such as mean and standard deviation (SD) to the above measures, taken independently of the specific interface layout. This process generated 10 features representing general gaze trends that do not take into account the nature of the interaction with MetaTutor (see Table 1, "no-AOI" column, where AOI stands for Area of Interest). The second type consists of features that do incorporate interface-specific information in terms of salient areas, or AOIs, of the MetaTutor's interface. We defined seven of these AOIs (labeled with rectangles in Figure 1): Text Content, Image Content, Goal, Subgoals, Learning Strategies Pallete, Agent and Table of Contents.

Table 1. Description of gaze-based features

No-AOI Features	AOI-based Features
Rate and Number of Fixations	Fixation rate in AOI
Mean and SD of Fixation Duration	Proportion of fixation time and fixation number in AOI
Mean and SD of Saccade length	Duration of longest fixation
Mean and SD of Relative Path Angles	Proportion of transitions from every other AOI to the current one (7 different features)
Mean and SD of Abs Path Angles	

For each AOI, we calculated the following features: rate of fixations, proportion of time and number of fixations, and duration of longest fixation. We also included the proportion of transitions from every other AOI to the current one. Proportional measures were used to assess the relative magnitude of attention devoted to each AOI over the course of a complete interaction. In total, there are 77 AOI-based features (summarized in the second column of Table 1). In the classification experiments described next, we trained separate classifiers on each of the two feature sets described above, as well as on a third feature set obtained by combining the two, referred to as the *Full* feature set from now on. Our goal is to ascertain the relative importance of AOI dependent and AOI independent features in predicting student learning.

4.2 Training classifiers on gaze data

A large number of features can lead to over-fitting when only relatively small datasets are available for training. To avoid this issue, we reduced the number of features by performing wrapper feature selection [15]. This approach is based on searching subsets of the available features to find one that gives the classifier with the highest accuracy, where the search is greedy if the initial set of features is large (as is the case for our *Full* and *AOI-based* feature sets). To further reduce the likelihood of over-fitting, the feature selection process was cross-validated. For each of the original feature sets, the final set of features was obtained by discarding all features that appeared in less than 10% of the cross validation folds.

Classification labels were generated by dividing students into High Learners (HL) or Low Learners (LL) based on a median split of their learning performance, measured as proportional learning gains (PLG), namely the ratio of the differences between post and pre-test scores, and between maximum post-test score and pre-test. One outlier was excluded, resulting in a dataset of 47 participants. It should be noted that, in this dataset, we found no significant differences between users from the adaptive and non-adaptive study conditions described in section 3[4] (t(45) = -0.77, p = 0.45, Cohen's d = 0.23). Thus, for the purpose of our analysis, it makes sense to collapse the two groups. Performing a median split on this dataset resulted in 23 LL (Mean PLG = 0.93, SD = 36.05), and 24 HL (Mean PLG = 67.01, SD = 16.48). Given these labels, we used the WEKA data mining toolkit to train a variety of classifiers with feature selection on our three feature sets: *Full*, *AOI-based* and *no-AOI*. The next section summarizes our results.

5 Results

5.1 Classification accuracy

Table 2. Accuracy and Kappa [5]scores for different classifiers and feature sets

Full Feature set	**Accuracy (%)**			**Kappa**
	Overall	**LL**	**HL**	
Simple Logistic Regression	78.3	70.43	85.83	0.56
Multinomial Logistic Regression	61.28	66.52	56.25	0.23
Naïve Bayes	71.7	51.3	91.25	0.43
Random Forest	64.48	67.83	61.67	0.29
Multilayer Perceptron	66.59	60.86	72.08	0.33
AOI-based Feature set	**Overall**	**LL**	**HL**	**Kappa**
Simple Logistic Regression	64.47	51.3	77.08	0.28
Multinomial Logistic Regression	54.47	51.3	57.5	0.09
Naïve Bayes	69.57	56.52	82.08	0.39
Random Forest	68.08	72.61	63.75	0.36
Multilayer Perceptron	56.59	51.3	61.67	0.13
No-AOI Feature set	**Overall**	**LL**	**HL**	**Kappa**
Simple Logistic Regression	52.55	60.43	45	0.05
Multinomial Logistic Regression	58.3	60.43	56.25	0.17
Naïve Bayes	52.34	45.65	58.75	0.04
Random Forest	48.93	48.69	49.17	-0.02
Multilayer Perceptron	55.96	54.78	57.08	0.12

Comment [D1]: Reviewer 2: what va
of kappa is considered large or significan

Comment [D2]: Do we need to add reference to (Landis and Koch, 1977) fo
Kappa interpretation

<0.20 – poor
0.21 – 0.40 – fair
0.41 – 0.60 – moderate
0.61 – 0.80 – good
0.80 – 0.81 – very good

[4] There was also no significant difference in PLGs between the two conditions in the original group.

[5] As per [16] kappa: <0.2 is poor; 0.21-0,4 is fair; 0.41-0.6 is moderate; >0.61 is good.

All the results reported here are based on 10-fold cross-validation, with 10 runs per fold, and pertain to the 5 best performing classifiers among the ones we tested (Simple Logistic Regression, Multinomial Logistic Regression, Naïve Bayes, Random Forest and Multilayer Perceptron). Table 2 reports, for each feature set (Full, AOI-based and No-AOI): overall accuracy (percentage of data points correctly classified), accuracy on each class (LL and HL), and kappa scores (another commonly used measure of accuracy that accounts for agreement due to chance)[17].

To ascertain the impact that different feature sets have on classification performance, we performed two, two-way ANOVA with feature set (3 levels) and classifiers (5 levels) as factors on both overall accuracy and kappa-scores. The two analyses generated analogous results, thus here we discuss only results on overall accuracy, because they are easier to interpret in terms of practical classification performance.

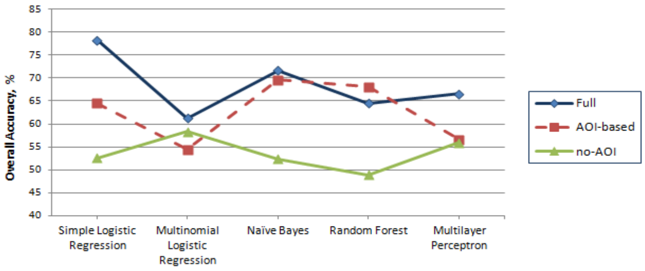


Fig. 2. Overall accuracy of the 5 classifiers over the 3 features sets

Figure 2 shows the mean of overall accuracy for each combination of classifier and feature set. There are significant main effects of both classifier (F(4, 36) = 9.01, p<0.001, $\eta_p^2 = 0.50$) and feature set, (F(2, 18) = 112.55, p<0.001, $\eta_p^2 = 0.93$), further qualified by a significant interaction between factors, F(8, 72) = 16.63, p<0.001, η_p^2=0.65), showing that classifier type influences the relative accuracy that can be achieved with each feature set. We performed planned contrast analysis (with corresponding Bonferroni adjustments) to gain a better understanding of the relative value of AOI-dependent and AOI independent features. This analysis shows that, in general, the performance of the classifiers that were trained on the Full feature set is significantly better than those trained on AOI-based features (t(72) = 6.21, *p*<0.001, Cohen's d = 1.46). The latter classifiers, in turn, perform better than those trained on no-AOI (t(72) = 9.53, *p*<0.001, Cohen's d = 2.24). In particular, the highest overall accuracy is achieved by Simple Logistic Regression on the Full dataset (78.3%, kappa = 0.56), which also shows good balance in class accuracy (70.4% on LL and 85.8% on HL as shown in Table 2).

We see this result as strong evidence of the value of eye-tracking data as a source of rich information in student modeling, because it shows that gaze information can be a good predictor of student learning, even before taking into account other student interaction behaviors (e.g., interface actions). Furthermore, this result seems to generalize across at least some learning environments that are different in nature, because similar accuracies were found in [4], where the authors looked at how gaze data predicts learning with an interactive simulation to support exploratory learning.

Simple Logistic regression on the Full dataset performs significantly better (t(72)=4.12, *p*<0.001, Cohen's d = 0.97) than the best performing classifier on AOI-only features, namely Naïve Bayes (69.6% accuracy, kappa = 0.39). This classifier is also quite unbalanced in terms of class accuracy (56.5% for LL, and 82% for HL), indicating that AOI-independent features have considerable added value when combined with AOI-dependent ones, although on their own they do not perform that well. It is interesting to see that the importance of having a combination of AOI-dependent and AOI-independent features is confirmed by the results of feature selection. For the Simple Logistic Regression classifier, which showed the best overall accuracy on the Full feature set, 14 features were selected: 4 AOI-independent features (mean and standard deviation of fixation duration, rate of fixations and mean of relative path angles), and 10 AOI-dependent ones. These include:

- 7 features describing proportion of transitions between AOIs: (i) from Table Of Contents (ToC), Learning Strategies Palette and Text Content to Subgoals; (ii) from ToC to Overall Learning Goal; (iii) from ToC and Image Content to Learning Strategies Palette; (iv) from Text Content to ToC.
- Longest fixation in Overall Learning Goal;
- Proportion of time and number of fixations spent in Subgoals.

It is worth noting that seven out of the ten AOI-based features are related to Overall Learning Goal and Subgoals AOIs, suggesting that attention to these elements is indeed important for assessing learning with MetaTutor. The next most frequent AOI to appear in this set, with two related features, is the Learning Strategies Palette, also supporting the importance of this element in gauging learning with MetaTutor. A notable absence is related to any feature involving the Agent AOI. As described in section 3, the MetaTutor agents provide spoken feedback and prompts during interaction. The fact that attention to the Agent AOI does not seem to play a role in our classification results may be due either to the fact that learners do not need to always look at an agent to process its audio prompts and feedback or, if they do, to the fact that agents' prompts and feedback do not impact learning enough to help detect it (an explanation supported by the lack of difference in learning between the adaptive and non-adaptive conditions in the original MetaTutor study).

5.2 Accuracy over time

The results in the previous section show that gaze data can be a rather powerful source of information to predict student learning, when data from the complete interaction with MetaTutor is available. Here we explore whether it can also be a source of information for detecting a student's learning performance *during* interaction with MetaTutor, to support real-time personalized help and feedback when needed. To address this question, we simulated online system conditions by incrementally feeding gaze data from the Full feature set to the best performing classifier from the previous section (Logistic Regression), and calculated overall and class accuracy (cross-validated) at regular intervals of 2 minutes.

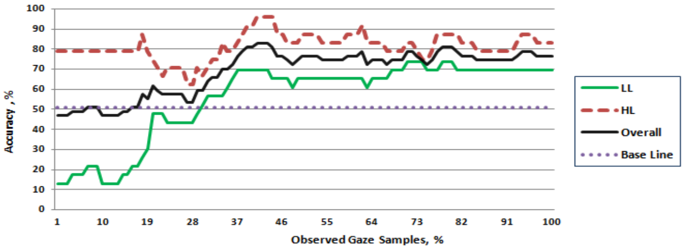


Fig. 3. Accuracy over time (Simple Logistic Regression, Full feature set)

Figure 3 shows the result of this process, i.e., the accuracy over time (overall and for each class) of the Logistic Regression classifier on the Full dataset. The classification accuracy starts growing above a baseline that predicts the most likely class (HL) based on a simple median split (51% overall accuracy), after seeing about 28% of the data (28.70 minutes from the beginning of the session). After seeing about 37% of the data (36.61 minutes), overall accuracy stabilizes above 72%, with some small fluctuations. The average accuracy over the session was 68.83%. We argue that these results provide strong support for using eye-tracking data as a source of on-line prediction of student learning, because they are obtained for an interactive system without even considering interface actions. We expect that combining features based on gaze data and features based on interface actions (e.g., taking notes, writing summaries, number of content pages visited, number of sub-goals completed) will boost accuracy over time, a finding that has already been observed in [18], where this approach was used on the interactive simulation discussed in [4].

6 Conclusions and future work

We presented research on understanding the value of gaze data to predict student learning during interaction with MetaTutor, an ITS that supports the acquisition of SRL processes. Our results show that gaze data alone achieves 78% classification accuracy on student learning after seeing all data from an interaction, and reaches 72% accuracy after seeing 37% of the data. These results replicate findings obtained by previous research using a different type of learning environment, and confirm the value of using gaze data as a source of information that ITSs can leverage to assess student learning and react accordingly. Our next step will be to combine gaze data with other multi-channel data sources (e.g., interaction logs, facial expressions of emotions), to see how this increases classification accuracy. We also plan to repeat this analysis to predict student states at the affective level (e.g. curiosity, boredom).

References

2. Azevedo, R., Behnagh, R., Duffy, M., Harley, J., Trevors, G.: Metacognition and self-regulated learning in student-centered leaning environments. Theoretical foundations of student-centered learning environments (2nd ed.). 171–197 (2012).
3. D'Mello, S., Olney, A., Williams, C., Hays, P.: Gaze tutor: A gaze-reactive intelligent tutoring system. Int. J. Hum.-Comput. Stud. 70, 377–398 (2012).
4. Kardan, S., Conati, C.: Exploring gaze data for determining user learning with an interactive simulation. In: Proc. of UMAP, 20th Int. Conf. on User Modeling, Adaptation, and Personalization. pp. 126–138 (2012).
5. Anderson, J.R., Gluck, K.: What role do cognitive architectures play in intelligent tutoring systems. Cognition & Instruction: Twenty-five years of progress. 227–262 (2001).
6. Conati, C., Merten, C.: Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. Knowledge-Based Systems. 20, 557–574 (2007).
7. Qu, L., Johnson, W.L.: Detecting the learner's motivational states in an interactive learning environment. In: Proc of AIED, 12th Int. Conf. on Artificial Intelligence in Education. pp. 547–554 (2005).
8. Muldner, K., Christopherson, R., Atkinson, R., Burleson, W.: Investigating the Utility of Eye-Tracking Information on Affect and Reasoning for User Modeling. In: Proc. of UMAP, 17th Int. Conf. on User Modeling, Adaptation, and Personalization. pp. 138–149 (2009).
9. Sibert, J.L., Gokturk, M., Lavine, R.A.: The reading assistant: eye gaze triggered auditory prompting for reading remediation. In: Proc. of the 13th annual ACM symposium on User interface software and technology. pp. 101–107 (2000).
10. Kinnebrew, J.S., Biswas, G.: Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. In: Proc. of EDM, 5th Int. Conf. on Educational Data Mining. pp. 57–64 (2012).
11. Bouchet, F., Azevedo, R., Kinnebrew, J.S., Biswas, G.: Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. In: Proc. of EDM, 5th Int. Conf. on Educational Data Mining. pp. 65–72 (2012).
12. Sabourin, J.L., Mott, B.W., Lester, J.C.: Early Prediction of Student Self-Regulation Strategies by Combining Multiple Models. In: Proc. of EDM, 5th Int. Conf on Educational Data Mining. pp. 156–159 (2012).
13. Azevedo, R., Landis, R., Feyzi-Behnagh, R., Duffy, M. et al.: The Effectiveness of Pedagogical Agents' Prompting and Feedback in Facilitating Co-adapted Learning with MetaTutor. In: Proc. of ITS, 11th Int. Conf. on Intelligent Tutoring Systems. pp. 212–221 (2012).
14. Goldberg, J.H., Helfman, J.I.: Comparing information graphics: a critical look at eye tracking. In: Proc. of BELIV, 3rd Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization. pp. 71–78 (2010).
15. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The J. of Machine Learning Research. 3, 1157–1182 (2003).
16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. 33, 159–174 (1977).
17. Ben-David, A.: About the relationship between ROC curves and Cohen's kappa. Engineering Applications of Artificial Intelligence. 21, 874–882 (2008).
18. Kardan, S., Conati, C.: Comparing and Combining Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In: Proc. of UMAP, 21st Int. Conf. on User Modeling, Adaptation and Personalization (2013).