**Predicting student's performance in a Serious Game with Machine Learning using Convolution Neural Network**

Dongwook An

The University of Texas at Austin

**Introduction**

**Predictive Learning Analytics**

The recent availability of large volumes of log data generated from Learning Management Systems (LMS) or Serious Games (SG) has led to extensive research on Learning Analytics (LA). Learning analytics referes to “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (Long & Siemens, 2011, p. 34). Specifically, the development of data mining as well as machine learning techniques has enabled researchers to develop more sophisticated model to predict student performance in LMS or SG. Predictive Learning Analytics (PLA) refers to the analysis and modeling of log data to accurately predict future outcomes, for example, whether a specific student will drop out or not (ECARANALYTICS, 2015). Many researchers believe that predictive modeling can help improve students’ performance (Siemens, 2012)

PLA is not a new concept in education, and many researchers have already been developing predictive models and applying them to solve their own problems. San Pedro, Baker, and Heffernan (2017) created a model to predict engagement level using user logs of a math tutoring system with machine learning as well as data mining techniques, while they were trying to investigate a relationship between engagement in the tutoring system and college enrollment. Mills et al. (2014) collected reading time information that how long readers spent on each page in a tutoring system and built a predictor of behavioral disengagement. Muratov, Lewis, Fourches, Tropsha, & Cox (2017) developed predictive models for student admissions using Random Forest machine learning technique showing that their model outperformed over some linear regression models for resident candidates while non-resident model were very similar. Seidel and Kutieleh (2017) built a decision tree model to predict more than 3500 first year students’ risk of attrition in a university using five years of students’ behavior data in their Learning Management System (LMS) as well as students’ demographic data. They also implemented intervention strategy for students with high predicted attrition probabilities before peak attrition time point, which was one of the main goals of PLA, to improve students’ retention and performance.

**Learning Analytics in Serious Games**

A variety of LA techniques has been applied to Serious Games Analytics such as cluster analysis, Decision Tree, Naive Bayes, Support Vector Machine, Logistic Regression, and Sequential Pattern Mining. Kerr (2015) identified students’ learning strategy through cluster analysis and then modified the educational game not only to reduce players’ irrelevant behaviors but also to improve positive perceptions of the game compared with the original setting. Paquette et al. (2016) built a predictive model of affect detection with game log data combined with ground-truth affect data via field observation in a SG. Kang, Liu, & Qu (2017) did Sequential Pattern Mining to explore students’ sequential behavior patterns in a SG and discovered that the patterns are various on different performance groups as well as different stages. Hicks et al. (2016) used Survival Analysis technique and Interaction Network Analysis for analyzing drop-out rates in a SG to find out potential areas of improvement. These efforts have been facilitating understanding students' engagement, behavioral patterns, learning process and performance levels in SG environments.

**Prediction problem in Complex Serious Games**

A serious game can be complicated enough to have diverse paths when solving a problem in it. This can make it difficult to build predictive models even though researchers make use of several machine learning techniques. One reason of it is that they could fail to construct a labeled dataset to make a predictive model or they could fail to predict performance significantly due to the game’s higher level of complexity although they could build dataset.

Numerous prior research in PLA was able to have labeled data with predictors in the form of multiple row to solve simple prediction problem in its context. For example, Huang and Fang (2013) used cumulative GPA, grades in four pre-requisite courses, and three midterm exam grades as features (predictors). Its target label (predicted value) was students’ performance at the end of the semester. And each row of the dataset represents each student’s information, as is usual in a traditional statistical dataset such as linear regression. For the dataset, many researchers have utilized frequency or sum of time log data for predictors as well. Mills et al. (2014) collected a dataset with students’ the location of quit (page number) as well as reading time information that imply how long readers spent on each page in a tutoring system. Each row of the dataset was comprised of each reader’s predictors with a target label: quitted or not quitted.

But in a complex research context or a sophisticated SG environment, dataset should have multiple rows per a user or multiple target labels per a user since the log data of a user should not be aggregated in a row. In that case, researchers can’t use popular machine learning techniques such as Decision Tree, Logistic regression, Neural Network, NaiveBayes, and Support Vector Machine since those methods need a dataset that should have each person each row as well as single target label (e.g. quitted or not quitted). For example, Min et al. (2016) investigated multiclass classification (multiple target label) problem in a serious game using long short-term memory networks (LSTMs) which is a variant of Recurrent Neural Network (RNN), a popular deep learning method for sequence observation. They found that LSTM models yielded the best performing model in classifying player goal with game logs of players.

In this paper, researchers explained a research context of multimodal (multi-label) learning analytics with multimodal data that need to have multiple row per a user as well as multiple target label in the dataset for PLA in a SG environment, and presented a predictive model with Machine Learning technique using Convolutional Neural Network. The following research questions guided this research:

1. How predictive model can be built to predict performance of users whose trajectories are vary in a complex serious game?

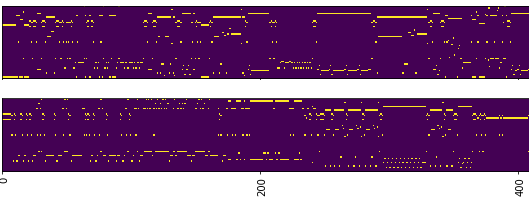
And this, in turn, leads to sub-questions such as:

1. What can be the predictors for the predictive model in a complex serious game to explain students’ performance?

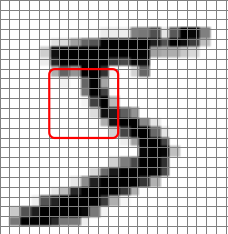
2. In case target labels are multiple, what kind of modeling technique can be used to build predictive model?

**Convolutional neural networks (CNN)**

The Convolutional Neural Network (CNN) is a class of deep learning, which was designed to process data that come in the form of multiple arrays, and has been widely used for challenging many machine-learning tasks such as text classification, image classification (Lecun, Bengio, & Hinton, 2015). The CNN captures spatial topology much better than standard neural networks since it detects distinctive features wherever they appear on the input (LeCun et al., 1998).

SG log data has spatial-temporal nature (Loh & Sheng, 2015). Figure 1 shows the spatial-temporal nature of multimodal log data in a SG environment. In the graph, two game players were randomly selected and visualized according to time (x-axis) and distinct action in different game scenes (y-axis). The action and navigation pattern of two users were quite different, but researchers in this research hypothesized that similar patterns between players could emerge at a different time period since there were frequent sequential patterns across players’ problem-solving stages in a SG environment, according to Kang, Liu, & Qu (2017)’s research. The CNN has been applied to analyze image classification such as hand-written digit image classification in Figure 2, since image pixels have a strong 2-dimension local structure in which the nearby pixels are highly correlated each other spatially or temporally (LeCun et al., 1998).

*Figure 1.* Spatial-temporal nature of interaction log data

  
*Figure 2.* CNN Image classification [Hand-written digital image] (2014, August 14). Retrieved July 13, 2018, from http://deepdish.io/public/images/PCANet\_mnist5\_patch.png

The visualized log graphs in Figure 1 also showed high correlation between nearby pixels in each graph since there were sequential action patterns in it for players to solve a problem in the SG environment. In this research, researchers worked directly with raw pixel image data like Figure 1, and the use of CNN allowed to capture the spatial-temporal patterns of log data to eventually build a predictive model.

**Method**

**Research Context**

This study focused on problem-solving activities of players in a 3D immersive SG environment called *Alien Rescue* (http://alienrescue.edb.utexas.edu), which is designed for

middle school science learning. *Alien Rescue* incorporated problem-based learning pedagogy, so that students could use various approaches as well as hypothesis testing to solve problems in it.

In *Alien Rescue*, students are asked to find new habitats for six Aliens in our Solar system who were displaced from a distant galaxy. Students navigate in a spaceship, a 3d game environment, and gather information of planets such as atmosphere, gravity, water and so forth. Then they need to compare what an alien species needs with what a planet’s data for becoming a new habitat of six aliens. Finally, students select a planet for each alien, and submit it in the game. Answers are multiple, for example, for the first alien, its new habitat can be Callisto or Charon, and for the second alien it’s Deimos, Ganymede or Callisto.

**Participants**

Participants were 535 six graders from middle schools in the Southwestern area of the United States. Initially there were 1,855 students but 1320 students were removed from dataset since they didn’t submit any solution for aliens’ new habitat. Those schools used *Alien Rescue* as part of their science curriculum in 2017, and the period of use varied from one week to over one month given their school schedule or teacher’s own preference.

**Data Sources & Preprocessing**

Gameplay data contains each player’s action type, action item, screen name on which the action is taking, and timestamp for each action (see Table 1). For example, if a student click a solar panel in Probe Design Tool screen to design a probe, then log data was {"action" : "Select: probeSource", "note" : "Solar Panels", "timestamp" : "Friday, March 10, 2017 3:06 PM", "tool" : "Probe Design", "userId" : "1QC1LbaCp3a0yS4QSpkJ7SLxQC82"} as JSON file format. The total log count was 1,517,196 from 1,855 users, but only 365,583 count log data was used for the predictive model since only 535 students submitted at least one solution. Some students didn’t submit an answer for some reason, such as absence or school schedule, individual computer malfunction or teacher’s preference using Google doc when students submit answers instead of putting the answers in the game.

Target labels of dataset was students’ submitted solution for six aliens’ new habitat. For example, a students submitted six planets as answers for each aliens and got full score if the criteria met, or a student submitted three planets for three aliens and would get one score if only one criteria met for the answers. Since answer planets are different for each aliens and students needed to send probes to different planets as hypothesis testing to get information that fits with the aliens, the labels were coded separately for each alien. For example, if a student submitted correct answer only for the first alien, then labels became 1 by 6 array, (1, 0, 0, 0, 0, 0), and if a student submitted correct answer for the first alien and the third alien, then then labels were (1, 0, 1, 0, 0, 0) for the student (see Table 2). Finally, label dataset for the predictive model was built as a form of 3-dimension array (535, 1, 6) for the 535 students.

Table 1

*An example of gameplay data*

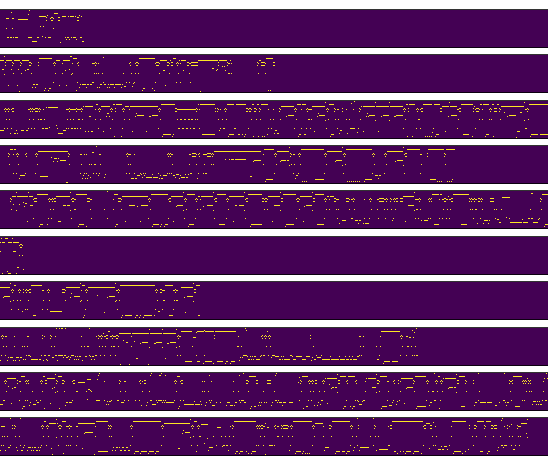
|  |  |
| --- | --- |
| Log | Description |
| "-Keu1n02ZvJ2MMeDLLga" : {  "dataLog" : {  "action" : "Select: probeSource",  "note" : "Solar Panels",  "timestamp" : "Friday, March 10, 2017 3:06 PM",  "tool" : "Probe Design",  "userId" : "1QC1LbaCp3a0yS4QSpkJ7SLxQC82"  }  } | Log id  Action type  Action item  Timestamp  Action screen  User id |

Table 2

*An example of label of dataset for each student’s submitted answer (Multi label classification)*

|  |  |  |
| --- | --- | --- |
| User id | Label (Target variable) | Total Score |
| 1QC1LbaCp3a0yS4QSpkJ7SLxQC82 | (1, 0, 0, 0, 0, 0) | 1 |
| oOjXUOiIAUPMHaWd7nVdVjWeX3t2 | (1, 1, 0, 0, 0, 0) | 2 |
| tlvqCSOAEuax3CFIXaTSyLFv8Tq2 | (0, 1, 0, 0, 1, 0) | 2 |
| mEKCOucQO0OXIEHPSc54KqcKZX33 | (0, 0, 0, 0, 0, 1) | 1 |
| OO0q44lAVhgMlwgvWqoHeDwf9Sq2 | (1, 0, 1, 1, 1, 1) | 5 |
| oOjXUOiIAUPMHaWd7nVdVjWeX3t2 | (1, 1, 1, 1, 1, 1) | 6 |

\* Total score was not used as features or labels

Predictor dataset was constructed as a form of 3-dimenstion array (535, 1180, 56). 2-D array data (1180, 56) was made for each of the 535 students from log data. First, Action field with tool field in Table 1 were dummy-coded and used as features. For example, the name of 56 dummy-coded features were: “action\_Click Back Button”, “action\_Click Destination Planet”, “action\_Click Forward Button”, “Menu,action\_Click Nav Menu”, “action\_Creat Note in: PLANET”, “action\_Lauch Probe”, “tool\_Concepts DB”, “tool\_MissionDB”, and etc. Second, all actions log of a student constructed to 2-D array data (1180, 56). For example, the first row of 1180 x 56 dataset indicates the player’s first action on a game scene, and the second row shows second action on a game scene. As 535 students have their own 2-D array data, it aggregated into 3-D array data (535, 1180, 56). Figure 3 shows the visualization of 10 students’ 2-D array data. 

*Figure 3.* Example of 10 students’ 2D array log data

Then the dataset was shuffled, split into a training set of 373 users (80%) and a test set of 94 users (20%). The training dataset was used for training the CNN, while the test dataset was used to evaluate the trained network.

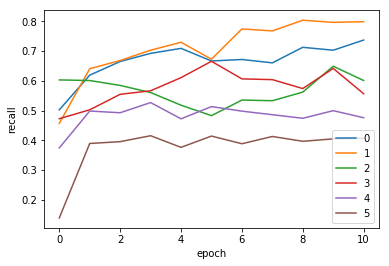
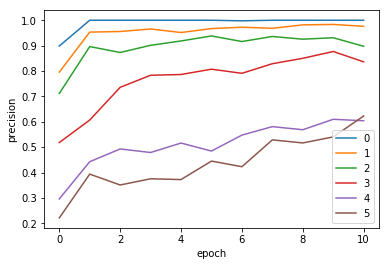
**Data Analysis**

The model is implemented and run using Keras (Chollet, 2017) with Tensorflow backend. The effectiveness of the model was evaluated by processing the test dataset. At the first time, the best test result of trained model reaches an average accurate rate of 9.235%, which means that the model can’t properly detect the student’s navigation patterns. Then researchers up-sampled the training dataset to handle imbalanced classes in the label dataset. For example, only 28 students out of 535 users had correct answer for the first alien, which means that the first column of 507 rows of the label dataset was 0. In that case, the model could predict the first label to all 0 then it could get 507/535 = 94.7% accuracy. So the researchers used up-sampling technique for every column in the label dataset to fix falsely predicting dominant class (zero) and to make balanced dataset.

**Results**

Researchers trained the network for 10 epochs. As it trains, the accuracy increased. After the first pass (epoch), it got 92.73% of training accuracy. After just 10 passes, it was achieving 99.95% multi-label classification accuracy on the training set. The best performance of .89 classification precision was achieved on the first label (the first alien). The worst performance of .59 classification precision was achieved on the fifth label (the fifth alien). Figure 4 shows the results of the CNN model precision and recall of each label. The true positive ratio, which also called precision, was less than 60% for the fifth and fourth aliens. The recall, which is defined as the number of true positives divided by the number of true positives plus the number of false negatives, was below 50% for the fifth and fourth aliens.

After developing the CNN model on the training data, researchers validated its performance on the holdout test data (20% of the whole dataset). The CNN model achieved 57.89% multi-label classification accuracy on the testing set, which is below the pure chance level of 67% prediction accuracy when predict it all zeros.



*Figure 4.* Model precision and recall of each label

**Findings and Discussion**

While the training test accuracy was achieved up to 99.95%, the test set accuracy was below pure chance accuracy. The precision and recall graph shows that the fourth and fifth label had least classification precision and recall of 50%. The second and the third label had less classification recall, which was below 60%, as shown in Figure 4.

The suggested model could not detect students’ navigation patterns properly. Labels mostly misclassified were label 5 and label 6 in the training dataset, which is likely caused by the different frequencies of in the original dataset. The label 1 and label 2 were reliably detected in the training dataset, but their accuracy were low in the test dataset, which is likely cause by the up-sampling on the training dataset. The model’s accuracy was below chance accuracy since it overfitted to the training dataset.

Although the CNN is the most powerful algorithm to capture patterns in many image classification problem, it failed to detect patterns in the game log. Future research applying another way of data pre-processing or another deep learning algorithm like Recurrent Neural Network (RNN) have possibility to attain meaningful results. But, this study demonstrated how deep learning algorithm can be incorporated into Learning Analytics in order to predict students’ performance in a Serious Game environment.

**References**

Chollet, F. (2017). *Deep learning with python*. Manning Publications Co..

ECAR-Analytics Working Group. (2015). The predictive learning analytics revolution: Leveraging learning data for student success. *ECAR working group*.

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016, April). Using game analytics to evaluate puzzle design and level progression in a serious game. *In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 440-448). ACM.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 61, 133-145.

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior*, 72, 757-770.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers and Education*, 104, 18–33. <https://doi.org/10.1016/j.compedu.2016.10.001>

Kerr, D. (2015). Using Data Mining Results to Improve Educational Video Game Design. *Journal of Educational Data Mining*, *7*(3), 1–17.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Loh, C. S., & Sheng, Y. (2015). Measuring expert performance for serious games analytics: From data to insights. In *Serious Games Analytics* (pp. 101-134). Springer, Cham.

Mills, C., Bosch, N., Graesser, A., & D’Mello, S. (2014, June). To quit or not to quit: predicting future behavioral disengagement from reading patterns. In *International Conference on Intelligent Tutoring Systems* (pp. 19–28). Springer, Cham.

Min, W., Wiggins, J., Pezzullo, L., Vail, A., Boyer, K. E., Mott, B., Frankosky, M., Wiebe, E., & Lester, J. (2016). Predicting dialogue acts for intelligent virtual agents with multimodal student interaction data. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining.* Raleigh, NC: International Educational Data Mining Society.

Muratov, E., Lewis, M., Fourches, D., Tropsha, A., & Cox, W. C. (2017). Computer-assisted decision support for student admissions based on their predicted academic performance. *American journal of pharmaceutical education*, 81(3), 46.

Paquette, L., Rowe, J., Baker, R., Mott, B., Lester, J., DeFalco, J., ... & Georgoulas, V. (2016). Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection. *International Educational Data Mining Society.*

San Pedro, M. O. Z., Baker, R. S., & Heffernan, N. T. (2017). An Integrated Look at Middle School Engagement and Learning in Digital Environments as Precursors to College Attendance. *Technology, Knowledge and Learning*, *22*(3), 243–270. <https://doi.org/10.1007/s10758-017-9318-z>

Seidel, E., & Kutieleh, S. (2017). Using predictive analytics to target and improve first year student attrition. *Australian Journal of Education*, 61(2), 200–218.

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*, 30–32. <https://doi.org/10.1145/2330601.2330605>

Siemens, G. (May 2012). Learning Analytics: Envisioning a Research Discipline and a Domain of Practice. *2nd International Conference on Learning Analytics & Knowledge*, 4–8. https://doi.org/10.1145/2330601.2330605

Stoehr, M. (2014, August 14). *Deep PCA Nets* [Hand-written digital image]. Retrieved July 13, 2018, from http://deepdish.io/public/images/PCANet\_mnist5\_patch.png