Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education
New Zealand: Asia-Pacific Society for Computers in Education

Predicting Student Carefulness within an Educational Game for Physics using Support Vector Machines

Michelle P. BANAWAN[a,b], Ma. Mercedes T. RODRIGO[b]
& Juan Miguel L. ANDRES[c]
[a]*Ateneo de Davao University, Davao City, Philippines*
[b]*Ateneo de Manila University, Quezon City, Philippines*
[c]*University of Pennsylvania, Philadelphia, USA*
*mpbanawan@addu.edu.ph

Abstract. Student carefulness is defined as being attentive, mindful or focused on the task at hand. In this paper, we create a predictive model for student carefulness within an educational game called Physics Playground (PP). We used game logs and manually-labeled gameplay clips of 54 students from the Philippines to develop three support vector regression models that predict carefulness using: (1) predictors of the game developers, (2) predictors from social science research, and (3) the combination of these predictors. After preprocessing and feature selection, the support vector regression models were able to significantly predict student carefulness. This research' empirical findings suggest that carefulness in Physics Playground can best be predicted by expanding the model of the game developers and including predictors that have been previously researched in the broader social science literature.

Keywords: carefulness · machine learning · support vector machines · regression· Physics Playground

1. Introduction

Carefulness is a characteristic of action that involves giving cautious attention, being thorough, alert, attentive, heedful or mindful. Social science and education researchers have extensively studied student carefulness. They have linked carefulness to improved problem-solving skills and higher-order thinking (Whimbey, 1980). On the other hand, students who are not careful have been found to be prone to impulsive, hurried, or incomplete problem solving (Kirby, Winston, and Santiesteban, 2005) and to making mistakes even after a skill has already been mastered (San Pedro, Baker, and Rodrigo, 2011). When a student is being careful, he/she is most likely to have self-discipline and avoid trivial errors (Gong, Beck, and Heffernan, 2009).

We study carefulness within one such learning environment, a game for Newtonian Physics called Physics Playground (PP). The creators of Physics Playground developed a model for student carefulness within PP. These predictors include: number of objects drawn, number of object limits reached in the game, average time in an attempt, average time before making an action in the first attempt, and average time between actions. However, they have not empirically validated this model. Further, the model does not include indicators of novelty, mastery, reflectivity, and other constructs that social science literature has cited as related to carefulness.

2. Research Goal

The goal of this study is to create a predictive model for carefulness by empirically validating the carefulness predictors of PP developers and expanding that model to include additional features from social science constructs that the Physics Playground developers may have not considered.

3. Physics Playground

Physics Playground (Shute et al., 2013) is a two-dimensional educational game that was designed to help secondary school students understand the application of Newton's three laws of motion. The students are required to bring the green ball to a red balloon by drawing agents like the ramp, pendulum, lever and springboard. Shute et al. modeled carefulness as an "unobservable" or a construct and was mapped to a number of "observables" or indicators.

4. Social Science Research Findings on Carefulness

The broader social science literature has found a number of cognitive and non-cognitive constructs that are related to carefulness. Examples of these will be briefly discussed in this section.

Novelty is an example of these constructs. It is defined in the Merriam-Webster dictionary as the quality or state of being new. Prior work (Ostafin and Kassman, 2012) has found that novelty is related to a student's problem solving behavior such that when a student is given new problems where he/she can be more creative and give non-routinary solutions or answers, he/she seemed to exercise more care in solving the problem. Within PP, student actions that might be indicative of novelty include attempts on new problems and use of new solutions to previous problems attempted. Another construct that has been frequently related to carefulness is mastery. Mastery is the possession of a skill or knowledge. It occurs when the content and objectives of instruction are learned (Bloom, 1968). There has been a number of studies that investigated the relationship that exist between carefulness and mastery, e.g. the student's level of proficiency in solving problems and degree of carefulness (Tsiriga, Virvou, 2002), student's intellectual performance and mindful engagement in games and computerized learning platforms (Salomon, Perkins, and Globerson, 1991), etc. Within PP, a student action or event that might be indicative of mastery would be his correct solutions to problems, i.e gold and silver badges, as PP awards these badges to students every time that they are able to solve a problem. Reflectivity or having the quality of being reflective is described as thinking carefully. It is another construct that has been investigated by researchers interested in student carefulness. It is likened to mindfulness or attentiveness and was seen to help students in problem solving and in academic performance. Its characteristics include being aware and attentive to the present and immediate experience and involves actions like re-reading the problem, backtracking, understanding the problem better, reviewing, rechecking and ensuring that everything has been considered. Related work show findings that reveal relationships between students' habitual action and lower levels of reflection (Lim, 2011), and an investigation of the nature of reflectivity and its relationship to learning goals (Lin, et al., 1999). Within PP, a student action that might be indicative of reflectivity would be when the student is taking time in solving the problems correctly.

5. Support Vector Machines for Prediction

Support vector regression has been used for small data sets of educational data to find "support vectors" that allow the widest margin between classes. It tolerates outliers and collinearity in data better than linear regression (Ashlay, Chan and Ikeda, 2006) and addresses the pre-requisite assumptions of normality and linearity for regression models. It is known to be able to get good generalization even with a limited number of learning patterns (Basak, Pal, and Patranabis, 2007). Support vector machines transforms the original data points in the training data to a higher-dimensional feature space and a linear regression computation is performed in this high dimensional feature space (Guajardo, 2006). Support vector machines aim to minimize the classification error of both training data and unseen data and has been known to outperform conventional classifiers, most especially when the number of training data is small (Abe, 2005).

6. Data Collection and Processing

Primary data was gathered enlisting the participation of 180 high school students from 3 universities in the Philippines. These students were given 120 minutes to play the game, during which the game logged all player interactions. PP automatically recorded player events/actions. These events have been captured in the logs resulting to an entry in the logs for (roughly) every second of gameplay. These raw logs were parsed to produce a comma separated value (csv) file which we later used in further summarization and preprocessing. A single attempt (on a single problem) is a summary of information of all the rows between the Level Start action to the Level End action from the interaction logs. On average, an attempt consisted of 4,373 actions (rows) with the shortest attempt only having 11 actions (rows) and the longest attempt having 109,091 actions (rows).

Parallel to the automated capture of the interaction logs, student usage/gameplay is also recorded such that they can be viewed as videos using PP's Replay Viewer utility. We divided these videos into clips that correspond to attempts. Unfortunately, not all attempts had equivalent gameplay videos. There were only 2,640 playable clips, all in all. A stratified sampling of these clips (1,990) was selected, video-casted to mp4 format and a utility was developed for the ground truth labeling. Physics experts/professors have been consulted to come-up with the criteria for labeling the clips. The coders, then, referred to these criteria during the coding process. The clips were then coded via consensus as to 1-least careful, 2-somewhat careful, and 3- very careful. Clips that cannot be coded as such were given a code of 0 for undetermined. Ground-truth coding used the experts' criteria and were not based on any of the features of Shute et al. nor candidate features of mastery, novelty or reflectivity. Finally, the attempts were further summarized and aggregated to student level details to build models based on Shute, et al.'s (2013) predictors.

7. Findings

7.1 Feature Engineering

We characterized each student with the following features:

1. number of object limits reached in problem [ObjectLimits]
2. number of objects drawn per level [ObjectsDrawn]
3. average time in seconds spent drawing per level [TimeDrawing]
4. average time in seconds between actions [TimeActions]
5. average time in seconds before making an action on the first attempt [Time1stAttempt]
6. number of gold badges earned [CountOfGold]
7. number of silver badges earned [CountOfSilver]
8. average difference between the time spent on problems and the median time for the problem/level [timediff]
9. count of unique solutions [UniqueSolutions]

Features 1-5 were taken from Shute, et al. (2013). Features 6-9 were candidate predictors indicative of other constructs that have been found to be related to carefulness. For this paper, we are trying to capture the unique attempts by looking at the problems that the students answered for the first time. However, even if such attempt is not the first attempt yet the student used a solution that is different from the previous solution used (e.g. using a springboard when the previous solution was a pendulum), then the attempt is still considered as novel. Initially two derived features were extracted for this purpose: (1) unique problems attempted and (2) unique solutions to non-unique problems, both having equal weights. These two features were later combined into a single feature, i.e. **unique solutions** as both refer to the same characteristic. The formula used was ***unique solutions = (unique problems attempted + unique solutions to non-unique problems) / number of attempts***, such that a perfect score of 1 means that the student had all unique solutions or attempted all problems only once.

Further, in PP, a student is able to earn badges every time he/she solves a problem. A gold badge is earned if the student has drawn three, or less, objects in coming up with the solution

and a silver badge is earned if the solution entailed drawing more (greater than three) objects. It is when a student draws more objects than necessary that he/she reaches the set object limits for the problem (feature #1). Hence, we also included the **count of gold badges** and **count of silver badges** as candidate predictors. We also felt that the time spent on solving the problems should also be relative to the overall time-based performance on a specific problem/level, i.e. median time for the problem/level. We, then, engineered another time-based feature - **timediff,** to be able to see if an attempt took longer than the median time or took less than the median time for the specific problem/level. A high **timediff** value means that the student is slower than other students, and a low **timediff** value means that the student is faster.

7.2 Descriptive Summary of the Features

Given 9 features of 54 labeled student instances, descriptive statistics are computed (table 1).
Table 1: Descriptive Statistics of the Features

Feature	Minimum	Maximum	Mean	Standard Deviation
ObjectLimits	0	55	5.61	9.51
ObjectsDrawn	349	2,201	832.13	350.18
TimeDrawing	1,930.10	7,975.70	4,117.66	1,126.82
TimeActions	88,280.40	326,090.63	169,168.73	49,815.30
Time1stAttempt	9,853.55	41,804.53	19,781.12	8,235.60
timediff	-73,883.50	219,703.03	18,457.63	53,319.30
CountOfSilver	7	34	17.80	6.93
CountOfGold	0	14	5.62	3.38
UniqueSolutions	15	76	31.59	11.86
Carefulness	1.26	2.71	2.09	0.28

The average carefulness label of the students ranged from 1.26 (1.0 – least careful) to 2.71 (3 – very careful), with an average of 2.09 (standard deviation of 0.28) which shows that, at average, students were somewhat careful during gameplay.

7.3 Feature Selection

Before building the models, we forward-selected the most efficient set of attributes per dataset, i.e. dataset with Shute et al.'s features only, dataset with candidate features from social science literature and the dataset combing all the features The remaining feature subsets are shown in table 2.
Table 2: Feature subsets of the three datasets after Forward Selection

Model	Features
Dataset with Shute, et al.'s features	ObjectsDrawn Time1stAttempt
Dataset with candidate features from Social Science literature	UniqueSolutions CountOfGold CountOfSilver
Dataset with combined features	UniqueSolutions CountOfGold CountOfSilver Time1stAttempt

Table 2 shows that only the number of objects drawn (ObjectsDrawn) and average time before making an action on the 1st attempt (Time1stAttempt) were selected as features out of the 5 proposed predictors of Shute, et al. For the candidate features (from Social Science literature), three out of four features were selected, only the difference between the time spent on the attempt and the median time for that level (timediff) was not selected. For the combined dataset, the

features selected were: the count of unique solutions (UniqueSolutions), number of gold and silver badges (CountOfGold and CountOfSilver), and the average time spent before making an action on the first attempt.

7.4 Carefulness SV Regression Models

With the outliers removed and the features selected through forward selection, the support vector machine algorithm was used resulting to the three SV Regression models in table 3:

Table 3: SV Regression Models Weight Vector

Model	Feature	Weight
Model 1	ObjectsDrawn	- 0.062
	Time1stAttempt	- 0.008
Model 2	UniqueSolutions	-0.196
	CountOfGold	0.127
	CountOfSilver	0.117
Model 3	UniqueSolutions	-0.195
	CountOfGold	0.137
	CountOfSilver	0.118
	Time1stAttempt	0.025

From Model 1, only the number of objects drawn (***ObjectsDrawn***) and average time before making an action on the 1st attempt (***Time1stAttempt***) came out as the significant predictors. Both predictors are negatively weighted. For this model, student carefulness can be attributed to and predicted by both ***ObjectsDrawn*** and ***Time1stAttempt***.

For Model 2, ***CountOfGold*** and ***CountOfSilver*** are positively weighted significant predictors, and ***UniqueSolutions*** is a negatively weighted significant predictor.

For Model 3, ***CountOfGold***, ***CountOfSilver*** and ***Time1stAttempt*** are positively weighted significant predictors and ***UniqueSolutions*** remained to be a negatively weighted significant predictor, consistent with Model 2. Further, the novelty of the problems attempted were not predictive of carefulness, as we initially suspected, and were predictive of non-carefulness instead. It is interesting to note that contrary to Model 1, the sign of ***Time1stAttempt*** in this expanded model has changed from negative to positive, which implies that the time that the student takes before making an action on the first attempt when combined with the badges earned predicts carefulness. Unlike in Model 1 where the time that the student takes before making an action on the first attempt predicts non-carefulness when combined with the number of objects drawn.

To evaluate the models, 10-fold cross validation was used resulting to the following average performance vector (table 4):

Table 4: SV Regression Models Performance Vector (at $p < 0.05$)

Model	RMSE	Correlation (R)	SquaredCorrelation
Model 1	0.244 +/- 0.142	0.345 +/- 0.362	0.250 +/- 0.332
Model 2	0.162 +/- 0.057	0.688 +/- 0.247	0.534 +/- 0.224
Model 3	0.159 +/- 0.065	0.652 +/- 0.309	0.521 +/- 0.340

The RMSE of all three models are relatively low (0.244, 0.162, and 0.159), taking into consideration that the carefulness label has a range of 1.46. Models 2 and 3 have lower RMSE values than model 1. The RMSE values, which as we know share the unit of the label, and the r^2 values obtained, indicate good fitness of the predictive models, most especially models 2 and 3.

8. Conclusions

We were able to empirically validate the carefulness predictors of Shute, et al. and expand their model by adding significant predictors studied in social science research.The model derived from Shute, et al.'s features significantly predicts carefulness ($r^{2=}$ .250 +/- 0.332; $p < 0.05$) with the number of objects drawn (ObjectsDrawn) and average time before making an action on the 1st

attempt (Time1stAttempt) as the most significant predictors. From the candidate predictors' model, carefulness has been significantly predicted (r^2 =.534 +/- 0.224; $p < 0.05$) by the counts of the badges (CountOfGold and CountOfSilver) and the UniqueSolutions. Consequently, this set of predictors describes carefulness as attributable to the students' ability to solve the problems by earning gold or silver badges. If we define the badges together with the time spent as indicators of reflectivity, then, we have reason to say that reflectivity is a determinant of carefulness, i.e. reflective students tend to be careful. Further, looking at the resulting combined features model ($r^{2=}$ 0.521 +/- 0.340; $p < 0.05$), which has a better fit than Shute, et al.'s model (Model 1), we can say that the predictors of Model 2 together with the average time before making an action on the 1st attempt (Time1stAttempt) significantly predicts carefulness. This finding corroborates work that carefulness is related to mastery and reflectivity.

9. Summary of Contributions and Future Work

One contribution of this work is the empirical validation of PP developers' model of carefulness that revealed that only the number of objects drawn and the time spent before making an action on the first attempt are significant predictors of student carefulness. Another contribution is the expansion of this model with the addition of reflectivity and mastery as significant predictors of student carefulness, confirming previous findings in social science research.

As a future work, to improve and ensure student carefulness, guidelines can be formulated on the design of educational games encouraging reflectivity among students and incorporating appropriate interventions for students with high level of mastery.

Acknowledgements

We thank the officials at the Ateneo de Davao University, University of San Carlos, University of the Cordilleras Laboratory High School and Bakakeng National High School for making this study possible. We would also like to thank Physics Playground developers and researchers, Dr. Valerie Shute, Dr. Matthew Ventura, and their colleagues at the Florida State University for collaborating with us.

References

Ashlay, K., Chan, T. W., & Ikeda, M. (2006). Intelligent Tutoring Systems. Springer-Verlag Berlin/Heidelberg.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. Neural Information Processing-Letters and Reviews, 11(10), 203-224.

Guajardo, J., Weber, R., & Miranda, J. (2006). A forecasting methodology using support vector regression and dynamic feature selection. Journal of Information & Knowledge Management, 5(04), 329-335.

Lim, L. A. Y. L. (2011). A comparison of students' reflective thinking across different years in a problem-based learning environment. Instructional Science, 39(2), 171-188.

Lin, X., Hmelo, C., Kinzer, C. K., & Secules, T. J. (1999). Designing technology to support reflection. Educational Technology Research and Development, 47(3), 43-62.

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. Educational researcher, 20(3), 2-9.

San Pedro, M. O. C. Z., d Baker, R. S., & Rodrigo, M. M. T. (2011, June). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In International Conference on Artificial Intelligence in Education (pp. 304-311). Springer Ber Heidelberg.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*(6), 423-430.

Tsiriga, V., & Virvou, M. (2002, October). Initializing the student model using stereotypes and machine learning. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics.

Whimbey, A. (1980). Students can learn to be better problem solvers. Educational Leadership, 37(7), 560-565.