Machine Learning for Behavioral Data: Report for Milestone 4

Declan Byrne, Oliver Dudler, David Paul Mark

I. Introduction

In this report we write about our project for Milestone 4 of the course Machine Learning for Behavorial Data. We use the given dataset provided by Calcularis.

Wheel-spinning is a term used in educational studies and refers to students who fail to master a skill, even when given multiple opportunities to do so. As we wanted to get more familiar with this phenomenon, we choose to analyze the learning behaviour of students in the Calcularis dataset and tried to detect wheel-spinning using a machine learning classifier. How well can we detect wheel-spinning students given the available data?

Formally defining wheel-spinning is not an easy task. There are multiple definitions used in literature [1][2]. We choose to go with the following one: If after a certain number of practice opportunities (PO), a student was not able to correctly solve a task three or more times in a row, the student is said to be wheel-spinning. If the student was able to solve the task correctly thrice in a row, the student has mastered the skill. The maximum number of POs before a student is considered wheel-spinning is called the cutoff or threshold value. A skill-student pair is said to be indeterminate if the student has not mastered the skill but also did not solve at least as many events for the skill as defined by the cutoff. Our goal is to detect wheel-spinning students as early and as accurately as possible. We use a time-series analysis to predict if the student will wheel-spin on a specific skill in the future.

II. Data Processing

Before being able to extract our relevant features, we had to clean the dataset. We removed duplicate events and sub-tasks, got rid of events which did not have any corresponding sub-tasks and we also filtered out all events which were not of type 'task'. This is because we are only interested in events corresponding to actual games and not things like zoo visits etc.

A. Features

Most papers dedicated to detecting wheel-spinning use similar sets of features to do so. We choose to use a sub-set of features as used in the paper by Yue Gong and Joseph E. Beck [1]. A lot of features used in said paper relied on a hint-system within the application or the response time of students. Both of those are not present in the Calcularis dataset. We used the following features, where each feature applies to a student-skill pair, as wheel-spinning/mastery is defined on a per-skill basis:

- Correct response count: Number of events correctly solved on the first attempt
- Correct response percentage
- Correct response in a row count: Current streak of correctly solved events
- Correct response in a row percentage
- Time on current skill: Total time a student invested for this skill
- Skill ID

From the 28 features mentioned in the paper [1], these were the only ones we were able to extract from the Calcularis dataset.

For the cutoff, we chose a value of 10. On the one hand, this value was frequently used in literature, on the other hand, the cutoff should be chosen dependent on the given dataset. We tried out multiple values for the cutoff before deciding to use 10. It was a good trade-off between not throwing away too much data by choosing a too-low cutoff, and artificially boosting the percentage of indeterminate student-skill pairs by having a too-high cutoff. Using the cutoff of 10, only about 3.66% of the data was lost/removed.

III. Methodology

In order to evaluate how well we can detect wheel-spinning, we wanted to use multiple machine-learning models and compare performances on our defined features. Given that the data processing took longer than expected, we chose to only consider a model using Random Forests and then evaluate more models in the future as a creative extension of the project. For each PO, from the first one up to the cutoff, we attempted to predict whether a student is wheel-spinning or not on the n-th PO given the data of the first $n-1$ POs. Using a grid search on the number of estimators and the depth of the decision trees of the random forest, we searched for the best parameters. We set 10% of users aside as a validation set, and then performed 3-fold cross-validation on the remaining data, splitting on the user level, i.e. the users in the validation set, training set and test set were all disjoint. For training our model, we got rid of all the indeterminate student-skill pairs as these data points have no ground truth for their labels. Alternative approaches would have been to consider these cases as mastery (optimistic wheel-spinning) or as wheel-spinning (pessimistic wheel-spinning) which would lead to a lower or higher percentage of students who are wheel-spinning,

respectively. An important note is that about 27.5% of student-skill pairs are indeterminate, therefore we removed a non-negligible fraction of the dataset. Reducing the dataset size is generally bad for training machine learning models, however, it is unavoidable here and choosing a different cutoff could lead to other problems as detailed above. A lot of students simply did not solve many tasks for the same skill.

IV. Results

In this section we examine the performance of the model with the best average accuracy in cross-validation, on the validation set. This model was a RandomForest model with 100 estimators (Decision trees) and a maximum depth of 15. Also, as the function to measure the quality of a split we used entropy. Figure 1 shows the confusion matrix of our model when given data for the first 10 POs and predicts whether students are wheel-spinning or not. The entries are shown as percentages. We can deduce that the precision of our model was in this case around 77.7% and the recall was 68.4%. When comparing to the paper's result [1], we achieved roughly the same precision but a better recall, i.e., we were able to detect more wheel-spinning cases ratio-wise. This might be explained by the fact that we have significantly more wheel-spinning students (roughly 37%) than in the paper (around 20%).

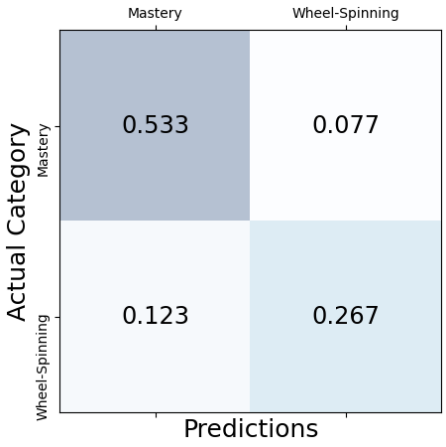

Figure 1. Confusion matrix of the Random Forest Model

Figure 3 shows the precision and recall over the time series data, i.e., where we try to predict wheel-spinning based on the first i PO's where $i \in [1, 10]$. We observe that the curves appear to stabilize around practice opportunity 4. This might be counter-intuitive since more data should lead to better predictions, but this could be explained by the fact that we have a strong drop-off in data as POs grows as seen in Figure 2, therefore the model doesn't have much data for higher POs.

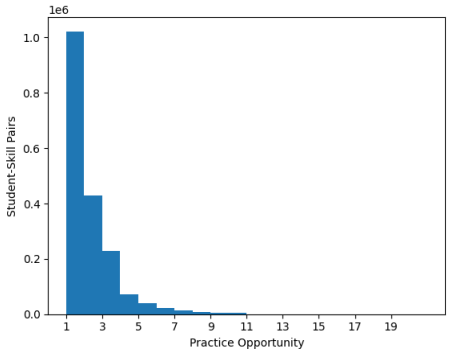

Figure 2. Distribution of Events over Practice Opportunity

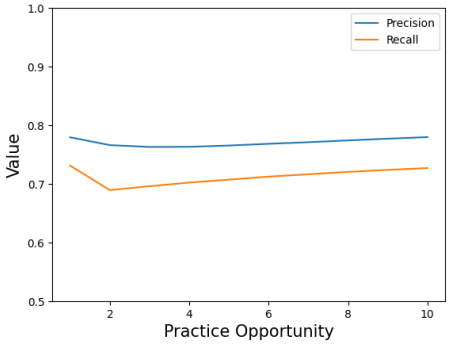

Figure 3. Precision and Recall for the Time Series Data

Figure 4 shows the accuracy and AUC scores over the time series data, similar to above. We see a slight upwards trend, however, after PO 4 the values remain approximately the same. This is likely explained by the strong drop-off mentioned above.

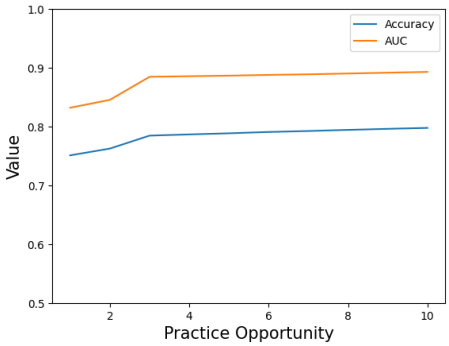

Figure 4. Accuracy and AUC for the Time Series Data

V. Conclusion

We were not able to test out as many models and parameters as we initially wanted to. However, even with the limited number of features, we achieved a somewhat

reasonable model, which was able to predict whether students are wheel-spinning or not up to a certain degree. The performance of our model might be improved if we had the necessary data to extract other features described in literature on this topic, as well as the exploration of different models such as a regression model.

VI. Remarks

There are a lot more models that could be considered and also more involved testing could give better results. Also, there are alternative ways of defining wheel-spinning (for example by using Bayesian Knowledge Tracing). Another possible extension would be to consider what type of interventions one could make when detecting that a student is wheel-spinning, as considered here [3]. As data processing took longer than expected, these extensions were not considered for Milestone 4, but remain valid possible extensions for Milestone 5.

References

[1] Y. Gong and J. E. Beck, "Towards detecting wheel-spinning: Future failure in mastery learning," 2015.

[2] C. Zhang, Y. Huang, J. Wang, D. Lu, W. Fang, J. Stamper, S. Fancsali, K. Holstein, and V. Aleven, "Early detection of wheel spinning: Comparison across tutors, models, features, and operationalizations." *International Educational Data Mining Society*, 2019.

[3] T. Mu, A. Jetten, and E. Brunskill, "Towards suggesting actionable interventions for wheel-spinning students." *International Educational Data Mining Society*, 2020.