A Course Agnostic Approach to Predicting Student Success from VLE Log Data Using Recurrent Neural Networks

Owen Corrigan and Alan F. Smeaton

Insight Centre for Data Analytics, Dublin City University,
Glasnevin, Dublin 9, Ireland
owen.corrigan@insight-centre.org

Abstract. We describe a method of improving the accuracy of a learning analytics system through the application of a Recurrent Neural Network over all students in a University, regardless of course. Our target is to discover how well a student will do in a class given their interaction with a virtual learning environment. We show how this method performs well when we want to predict how well students will do, even if we do not have a model trained based on their specific course.

Keywords: learning analytics, student intervention, machine learning

1 Introduction

Our goal is to improve student outcomes by predicting how well they will do in exams, by the middle of a semester. One approach to this is to take student data from systems they interact with, and feed it into a machine learning algorithm to identify students who are struggling [4]. We examine students interactions with a Virtual Learning Environment (VLE).

We faced two major difficulties when we designed our system. The first was a lack of data to build the models with. Typically when training a machine learning we need a large amount of training samples. Our data set included over 255,659 exam sittings, where we had both the student who took the exam and associated Moodle logs with that person. However we discovered that class sizes were not that large, with the largest one containing 2,879 students across 5 years and the average amount of students per module was 127. Classifiers trained on individual classes had poor results, particularly those where we had few examples of failures. We dealt with this issue by only keeping courses in our intervention program if they met some heuristics. For example, the classification accuracy was sufficiently high (e.g 0.6 minimum ROC), the class size must have at least 100 students per year, and a maximum of 85% pass rate. However, this only left us with a handful of the largest classes for which we could run our intervention system on.

The second was that it is impossible to build a model for a new course, due to a lack of training data. If a new course starts, we cannot include it in our alerting system, until one year of data has been collected.

In this paper we propose method of getting a larger data set for training, by training our classifier over all students in all 255,659 modules. This solves the first problem by building a model which can serve as a baseline model, across all students, which can be applied to offset the poor performance of a model trained on a small class. This also approach also allows us to predict student success for modules where we do not even have any training data for the course that they are in.

As a side effect of this, we are able to model the data using a recurrent neural network. This is useful, as they can perform very well on time series data, but can only be trained on a large corpus of data. We show they easily outperform random forests, which is what we found previously to perform the best on the data.

2 Related Work

The signals project [2] was an early pioneer of the concept of predicting student success, and feeding that information back to students and faculty in order to improve student outcomes. In their project they divided how well a student was doing in three tiers - red, amber and green. This is a very broad range of values to give and our project aims to improve the granularity of these predictions, making them more useful for staff and administrators.

There are many other examples of systems which make predictions and use them to make [8], [5], [3] and [7]. In [1] Agudo et. al. examine whether it is possible to predict student performance in VLE environments, face to face environments and online only environments. In [6] Okubo et. al use a Recurrent Neural Network to predict student grades. However, this is over a single course with 108 students, and so suffers from the same generalizability issues that we encountered earlier.

In a novel paper [9] Zorrilla et al. attempt to solve a similar problem by training many ready-made models. They then train a meta-classifier to take a new dataset and classify it based on which model is most similar to it. We believe that our method improves on this, as it does not require extracting meta-features from sets of data.

3 Regression Analysis

In our solution we extract very simple features from our data namely the number of times a user accessed Moodle in a given week.

Since these data points come in a stream, one natural solution to this algorithm is to use Recurrent Neural Networks, in it's most popular architecture variation — Long Short Term Memory (LSTM). A RNN is a type of Artificial Neural Network in which neurons can also connect back to themselves. This allows them to be trained on sequences, and to learn to remember important features of the series. They have been applied successfully to time series predictions, as well as a wide variety of other tasks such as handwriting simulation and speech synthesis. Because our problem involves predictions based on time series

and making predictions at multiple steps in time, we used RNN's as we believed they would be a natural fit. We also evaluate our results using a Random Forest, as this was what we found worked best previously.

4 Experiment Details

In Table 1 we see the results of running our regressor across several regression algorithms and parameters.

Table 1. Regression Results

Classifier	mean squared error	r-squared	p-value
Dummy Regressor	203.81967	0	0.99999
LSTM	200.64860	0.13382	5.36e-238
Random Forest Regressor	210.50645	0.08072	2.30e-87
LSTM - Dropout 0.2	201.0785	0.13547	6.81e-244
LSTM - Dropout 0.5	200.1333	0.13360	3.08e-237

The models we ran in our experiments were:

- A "Dummy" Regressor which always returns the mean exam results. Our classifiers should be at least as good as the Dummy Regressor;
- An Random Forest regressor. This had 1,000 estimators as a hyper parameter;
- A simple LSTM. All of the LSTM's were run for 300 epochs over the whole data set;
- 2 Versions of an LSTM with different values for "Dropout". This is a technique to reduce over-fitting. During training, this will set a random set of hidden nodes to be ignored. This forces the system to build in redundancy, which reduces the ability of the network to learn features based on noise.

From the table we can see that LSTM far outperforms the Random Forest regressor, explaining 13.3% of the variance of the model, as opposed to 8.1%. This result is particularly good, as when we tried to fit a regressor to course-level features, the results were not significant at the $p<0.05$ level. We can also see that setting dropout to 0.2 improves the performance of the LSTM marginally.

5 Conclusions and Future Work

In this paper we have shown that it is possible to build accurate models for predicting student exam outcomes over all courses. Doing so provides surprisingly good results which can be used as part of a student intervention system. We have shown that RNN's perform very well at this task.

In future work we will explore combining this information with more complex features such as activity types, times of day accessed, etc. We also believe there is scope for a method which gets the best of both worlds — using a course-agnostic approach when there is isn't a lot of information available about a course, and adapting the predictions to a particular course when there is more data.

Acknowledgements This research was supported by Science Foundation Ireland under grant number SFI/12/RC/2289, and by Dublin City University.

References

[1] Ángel F Agudo-Peregrina et al. "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning". In: *Computers in human behavior* 31 (2014), pp. 542–550.

[2] Kimberly E Arnold and Matthew D Pistilli. "Course signals at Purdue: Using learning analytics to increase student success". In: *Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM. 2012, pp. 267–270.

[3] Qijie Vicky Cai, Carrie L Lewis, and Jude Higdon. "Developing an early-alert system to promote student visits to tutor center". In: *The Learning Assistance Review* 20.1 (2015), p. 61.

[4] Owen Corrigan et al. "Using Educational Analytics to Improve Test Performance". In: *Design for Teaching and Learning in a Networked World*. Springer, 2015, pp. 42–55.

[5] Emma Howard, Maria Meehan, and Andrew Parnell. *Developing Accurate Early Warning Systems Via Data Analytics*. 2016. eprint: arXiv:1612.05735.

[6] F. Okubo et al. "A Neural Network Approach for Students' Performance Prediction". In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. LAK '17. Vancouver, British Columbia, Canada: ACM, 2017, pp. 598–599. ISBN: 978-1-4503-4870-6. DOI: 10.1145/3027385.3029479.

[7] Yeonjeong Park and Il-Hyun Jo. "Development of the Learning Analytics Dashboard to Support Students' Learning Performance." In: *J. UCS* 21.1 (2015), pp. 110–133.

[8] Mohammed Saqr, Uno Fors, and Matti Tedre. "How learning analytics can early predict under-achieving students in a blended medical education course". In: *Medical Teacher* 39.7 (2017), pp. 757–767. DOI: 10.1080/0142159X.2017.1309376.

[9] Marta Zorrilla and Diego Garcia-Saiz. "Meta-learning: Can It Be Suitable to Automatise the KDD Process for the Educational Domain?" In: *Rough Sets and Intelligent Systems Paradigms: Second International Conference, RSEISP 2014*. Springer International Publishing, 2014, pp. 285–292. ISBN: 978-3-319-08729-0. DOI: 10.1007/978-3-319-08729-0_28.