Time Series Analysis of VLE Activity Data

Ewa Młynarska
Insight Centre, University College Dublin, Ireland
ewa.mlynarska@insight-centre.org

Derek Greene
Insight Centre, University College Dublin, Ireland
derek.greene@ucd.ie

Pádraig Cunningham
Insight Centre, University College Dublin, Ireland
padraig.cunningham@ucd.ie

ABSTRACT

Virtual Learning Environments (VLE), such as Moodle, are purpose-built platforms in which teachers and students interact to exchange, review, and submit learning material and information. In this paper, we examine a complex VLE dataset from a large Irish university in an attempt to characterize student behavior with respect to deadlines and grades. We demonstrate that, by clustering activity profiles represented as time series using Dynamic Time Warping, we can uncover meaningful clusters of students exhibiting similar behaviors even in a sparsely-populated system. We use these clusters to identify distinct activity patterns among students, such as Procrastinators, Strugglers, and Experts. These patterns can provide us with an insight into the behavior of students, and ultimately help institutions to exploit deployed learning platforms so as to better structure their courses.

Keywords

Learning analytics, Data mining, Moodle, Time series, VLE

1. INTRODUCTION

The availability of log data from virtual learning environments (VLEs) such as Moodle presents an opportunity to improve learning outcomes and address challenges in the third level sector. We propose representing a student's efforts as a complete time-series of activity counts. We analyse yearly anonymised Moodle activity data from 13 Computer Science courses at University College Dublin (UCD), Ireland, and seek to identify patterns and relationships between more than one attribute that might lead to a student failing a course. A major potential benefit of this would be to introduce mechanisms identifying issues in the learning system early during the semester, supporting interventions and changes in the way in which a course is delivered.

A large amount of previous research in this area relates to different activity types, which are most predictive for a single dataset [1, 3]. This makes it difficult to generalise those methods to systems where the type and volume of Moodle activity can vary significantly. In order to facilitate the performance prediction on less structured systems, we need methods incorporating multiple features to deal with the sparsity problem. As a solution, we present a method for mining student activity on sparse data via Time Series Clustering. We explore the use of Dynamic Time Warping (DTW) as an appropriate distance measure to cluster students based on their activity patterns, so as to achieve clustering indicating more structured activity patterns influencing students' grades. DTW allows two time series that are similar but out of phase to be aligned to one another. To gain a macro-level view regarding whether these patterns occur across all assignments, we subsequently perform a second level aggregate clustering on the clusters coming from each assignment. This results in seven prototypical behaviour patterns (see example in Figure 1), that we believe can lead to better understanding of the behaviour of larger groups of students in VLEs.

2. TIME SERIES ANALYSIS

To perform clustering, the Moodle activity data was transformed into a series of equispaced points in time. In our case, a time series is a three week timeline – from two weeks before a given assignment submission date until one week after the deadline. These timelines were divided into 12 hour buckets of activity counts. We applied k-means clustering using DTW as a distance measure to cluster the timelines for each assignment. For a given number of clusters k, the algorithm was repeated 10 times and the best clustering was selected (based on the fitness score explained below). Due to the fact that DTW is not a true metric, k-means is not guaranteed to converge, so we limited each run to a maximum of 50 iterations. To choose the size of the DTW time window, we ran k-means for *window sizes* $\in [0, 3]$. The results did not conclusively indicate that any single *window size* leads to a significant decrease in cluster grade variance, which is unsurprising. In cases where there are many time series exhibiting little activity, it will be difficult to differentiate between the series and so a larger window size will be more appropriate. Based on this rationale, we believe that *window size* selection should be run for each assignment separately when applying this type of analysis in practice. The fitness function helping in selection of the best clustering needs to take into consideration that two clusters of different sizes might have the same variance value; this issue can be solved by applying a penalty to smaller clusters. We also

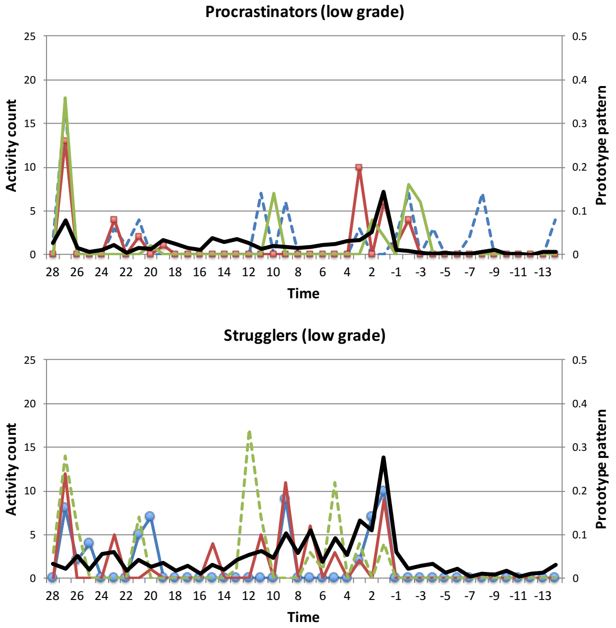

Figure 1: Two of the seven prototype activity patterns that occurred in Assignment #1. The black trend-line represents the prototype pattern. The coloured lines represent the activities of individual students. Negative numbers on the Time axis represent time after the deadline.

would like a "balanced clustering" where the variance of the cluster sizes is as small as possible. Based on these requirements, the fitness score calculation for a clustering generated by k-means consists of three steps:

1. The mean variance of the k-means clustering is calculated using the weighted average of all the clusters' variances, where the weight is based on the size of the cluster. This way the clusterings containing larger clusters with lower variances will be awarded better scores.
2. It is crucial to test the difference between a baseline clustering and actual results to define the significance of the clustering. For that purpose we run multiple random assignments of time series to calculate the expected score which could be achieved by chance for a given number of clusters.
3. To incorporate the baseline comparison in the score, the weighted average variance score from Step 1 is normalised with respect to the random assignment score from Step 2. A good clustering should achieve a low resulting score.

3. DISCUSSION

In our analysis, we took into account 52 two weeks assignments due to their longer and richer time series. We applied the time series clustering methodology described in previous section to the activity data for each of the assignments in the dataset, which are naturally split into two semesters. The Semester 1 clusterings appeared to show a number of frequently-appearing patterns across different courses. To gain a deeper insight into these patterns, we applied a second level of clustering – i.e. a clustering of the original clusters from all assignments. To support the comparison of clusters originating from different modules, the mean time series for each cluster was normalised. Based on the associated assignment scores, these normalised series were then stratified into low, medium, and high grade groups. We subsequently applied time series clustering with $k = 4$ and *window size* 1 to the normalised series in each of the stratified groups. Grade group names chosen by us were motivated by the behavioural pattern of students and some of them were inspired by previous research [2]. This second level of clustering revealed seven distinct prototypical patterns, which are present across multiple assignments and courses: *Procrastinators, Unmotivated, Strugglers, Systematic, Hard-workers, Strategists and Experts.*

The students rewarded with low grades were the second largest group of submissions after medium graded submissions having the smallest average activity per submission. The first out of 3 largest clusters was a group barely active on Moodle, performing submission activity at the deadline only (See Figure 1). As mentioned by Cerezo *et al.* [2], these could be labelled as Procrastinators. The black trend-line on the graph depicts prototype activity pattern and group of time series represents activity of students from the sample cluster. The third biggest group contains those students doing the minimum amount of work and showing larger activity towards the deadline (see Figure 1). The second academic semester courses mostly exhibit similar clusters from the first semester. The percentages indicate that for the Low Grade group, the Strugglers were most common and Procrastinators were less common.

While we did observe significant numbers of outliers, the relevant courses should be considered using a separate analysis to determine whether external factors are at play (e.g. continuous assessment rather than discrete assignments, lack of material provided on Moodle for a specific course). Finally, it is worth exploring anomalous clusters in the context of activity outside that assignment or course. We are currently in the process of extending our research to address the behavioural patterns of knowledge seekers in alternative, more complex learning environments.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

4. REFERENCES

[1] C. Brooks, C. Thompson, and S. Teasley. A time series interaction analysis method for building predictive models of learners using log data. In *Proc. 5th International Conference on Learning Analytics And Knowledge.*, ACM, 2015.

[2] R. Cerezo, M. Sanchez-Santillan, J.C. Nunez, and M.P. Paule. Different patterns of students' interaction with moodle and their relationship with achievement. In *Proc. 8th International Conference on Educational Data Mining*, 2015.

[3] L. V. Morris, C. Finnegan, and S. Wu. Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8.3:221–231, 2005.