Mining behaviours of students in autograding submission system logs

Jessica McBroom, Bryn Jeffries, Irena Koprinska and Kalina Yacef
School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia
jmcb6755@uni.sydney.edu.au, {bryn.jeffries, irena.koprinska, kalina.yacef}@sydney.edu.au

ABSTRACT

Effective mining of data from online submission systems offers the potential to improve educational outcomes by identifying student habits and behaviours and their relationship with levels of achievement. In particular, it may assist in identifying students at risk of performing poorly, allowing for early intervention. In this paper we investigate different methods of following the development of student behaviour throughout the semester using online submission system data, and different approaches to analysing this development. We demonstrate the application of these methods to data from a junior computer science course (N=494) and discuss their usefulness in understanding the common behavioural strategies of students in this course and how these develop over time. Finally, we draw links between behaviour in weekly coding tasks and student performance in the final exam and discuss whether these methods could be applicable midway through the semester.

Keywords

Clustering student behaviour; autograding system; assessment and feedback.

1. INTRODUCTION

Autograding submission systems are valuable tools in a modern teaching environment. By automatically assessing a student's submission, feedback can be returned to the student immediately without increasing the burden of marking for the teacher. Students are empowered to repeatedly improve their submission before a final deadline. However, such systems are only likely to improve the student's learning experience if the student allocates time to use feedback for subsequent submissions.

Teachers know from observation that students adopt a range of approaches to learning exercises, especially when outside the classroom. At one extreme, an ideal student will attempt an exercise immediately, and make increasingly better submissions based upon the feedback received. At the other extreme a student may make their first attempt just prior to the submission deadline, leaving no opportunity to improve or even make a decent first attempt. These behaviours, and many in between, may be due to deeply ingrained habits or external factors such as other time commitments. Using online submission systems in our teaching provide us with the opportunity to exploit the historical data of students' attempts. In this work, we investigated techniques of identifying and following the development of student behaviour over the semester, with specific focus on the application of these techniques to a junior computer science course. We were interested in the most common behaviours of students, whether these behaviours changed over time, and relationships between these behaviours and final exam outcomes. We were also interested in how applicable these methods were midway through the semester.

This paper is structured as follows. We first give an overview of the related work on the use of autograding systems and on mining student behaviour in these systems. Section 3 explains the context in which our data was captured. Section 4 is the main part of the paper: it presents our clustering-based approach to detecting and tracking students' behaviours. We finally conclude with a discussion on these different approaches.

2. RELATED WORK

The use of autograding systems in computer science courses have been reported in [1-6], with the majority of studies focusing on analysing the effectiveness of the autograding systems as opposed to understanding student behaviours. Sherman et al. [1] introduced Bottlenose, an autograding system used in a first year programming course in C, and compared the student behaviour on the same assignments when using Bottlenose and when not using it. The results showed that the number of submissions per student per assignment was significantly higher when using the autograding system, which was attributed to students making use of the feedback to improve their programs. Enström et al. [2] developed Kattis, an automated assessment system used at KTH in Sweden for teaching programming and algorithms courses. The use of Kattis resulted in improved student motivation (increased number of submissions) and also in higher student satisfaction in the course evaluation survey. The autograding system Autolab [3] was developed at Carnegie Mellon University and used in a first year programming course in C. Its real-time scoreboard, which shows the class performance on the assessment task, was found to create a healthy competition encouraging students to improve their assignments, and do this quicker.

There has also been some recent work on mining log data from autograding systems [4-6]. Gramoli et al. [4] analysed the impact of autograding and instant feedback using the system PASTA in various computer science courses, from first to fourth year. They found that the instant feedback was beneficial not only for courses focusing on programming but also for courses that use programming as a tool to solve subject specific problems. The relation between the student performance and the chosen programming language and the time when the students start and finish their assignment submissions was also studied. Koprinska et al. [6] investigated whether students at risk of failing in a first year

programming course can be detected early in the semester, using information from three sources: the autograding system PASTA, a discussion board and assessment marks. They built a decision tree that was able to achieve 87% accuracy in predicting the exam mark from information available in the middle of the semester. It was also shown that using the information from the autograding system improved the accuracy, compared to only using the assessment marks. In [5], data from the same sources was used to define the characteristics of high, average- and low-performing students and predict their performance.

More broadly, the related work also includes mining log data from student submissions in computer science courses. Perera et al. [7] analysed behavioural data from online group collaboration logs in a software development project. The goal was to identify patterns and behaviours associated with positive and negative outcomes. Clustering was applied to find similar students and similar teams, and sequential pattern mining was used to extract sequences of frequent events. Student behavioural data from a high school computer science MOOC was analysed by Tomkins et al. [8]. They characterised the performance of high and low achieving students based on the student behaviour in the course and discussion board, and built a predictive model using support vector machines to predict if a student will pass or fail an exam, conducted after the course has finished.

In this paper we extend the previous work on mining log data from autograding systems in computer science courses. Our goal is to study the evolution of student behaviour during the semester, with a view that this could assist in early intervention in future course offerings or provide guidance for course restructuring. We propose different clustering methods and demonstrate their application in the context of a large first year computer science course. We discuss the effectiveness of these techniques for extracting and understanding behavioural patterns, and how these patterns develop over time.

3. DATA

PASTA is an autograding system for computer programming courses developed in our school [9]. Students submit their solution (programming code) to an assessment task. Then PASTA checks this solution by running a set of tests designed by the teacher and provides immediate feedback to the student about the passed and failed tests. Students can then correct their mistakes and resubmit the solution until all tests are passed. PASTA can be configured in different ways - the number of allowed attempts can be limited or unlimited, some tests can be hidden (i.e. not available for immediate feedback, only available after the deadline) and teachers can also add manual comments to complement the automatic feedback. It supports several languages (e.g. Java, C, C++, Python and Matlab) and has been used for various courses – introductory programming, data structures, algorithms, formal languages, artificial intelligence, databases and networks.

PASTA has received positive feedback from students due to the instant feedback and multiple attempts features. Its use has resulted in better student engagement, and also transparent and fair marking as the same tests are used for all students. For each student and task, the PASTA data contains: all submission attempts, the tests that were passed and failed, the time stamps and the mark obtained.

The data used in this paper comes from a junior unit of study on data structures [10], which ran in Semester 2 of 2015 with 494 students enrolled. Students were using PASTA on a weekly basis to submit exercises, over a period of 11 weeks. The exercises were made available just after the lecture related to the topic (say Hashing) and constitute the core material of the tutorials (2 hour computer-based practical sessions, with a ratio of one teacher to 20 students). Each week, one exercise was flagged for assessment and was due the following week, i.e. 12 days after release. The number of attempts allowed was unlimited.

4. ANALYSIS OF STUDENT BEHAVIOUR

There are many ways students work towards their weekly exercises and use PASTA. For instance, students may start early and submit several attempts until their submission is 100% successful; some may start late and have time to submit only once a half-done attempt; others may not submit anything at all; and so on. Our approach to follow students' behaviour on their weekly work is to first cluster behaviours on all submissions, for all students (section 4.1). Then we explore several ways of tracking students' behaviour during the semester (sections 4.2 to 4.5).

4.1 Submission clustering: typical behaviour on one submission

In order to determine the types of approaches students take when completing weekly tasks, we performed a clustering on all the data available. For each given student and week, we created a vector containing information about the student's behaviour on that week's submission. We chose features which related to student submission times as an indication of their approach to the task. We also included features relating to student marks, number of attempts and number of compile errors, which provided an indication of performance. In total there were 5434 vectors (11 weeks, 494 students), each representing a submission (possibly non-existent) by one student. Table 1 describes the features used in this initial clustering.

Table 1. Features used in initial clustering

Feature	Description
percent_early	Percentage of attempts made three days or more before the due date
percent_normal	Percentage of attempts made that were neither early nor late.
percent_late	Percentage of attempts made on the due date
num_compile_errors	Number of attempts involving compilation errors.
first_mark	Percentage of tests passed on first attempt.
last_mark	Percentage of tests passed on last attempt.
num_attempts	Number of attempts not involving compilation errors.
time_taken	Indicator for the time between the first and last submission. 0: student only made 1 submission (time between the first and last submission not relevant); 0.5: student took less than 26.45 minutes to complete their task; 1: student took more than 26.45 minutes to complete their task; -100: student did not attempt the task; (forces students who did not submit into their own cluster)
single_attempt	Specifies whether the student made no attempts ("none"), a single attempt ("yes" or multiple attempts ("no").

We note that the features, percent_early, percent_normal and percent_late are dependent. However, removing one would lead to different results depending on which feature was removed, so all were included to preserve symmetry.

We then clustered these 5434 vectors (with k-means algorithm) into six groups, with centroids are summarised in Table 2. Since these clusters would be used to perform further clustering, in which the distance between all clusters would be assumed to be equal, it was important that there were not two similar clusters, or one cluster comprised of what should be two clusters. We experimented with various numbers of clusters in the range of 4-7, and found that 6 clusters best satisfied these criteria.

Table 2. Cluster centroids of submissions clustering

Feature	Full Data	Cluster Number (Number of Vectors)					
		0 (5434)	1 (488)	2 (1017)	3 (903)	4 (719)	5 (607)
% early	0.30	0.55	1.00	0	0.39	0.00	0.00
% normal	0.22	0.19	0.00	0	0.43	0.99	0.01
% late	0.17	0.27	0.00	0	0.18	0.01	0.99
num compile	0.14	0.79	0.08	0	0.06	0.17	0.21
first mark	0.57	0.65	0.96	0	0.68	0.93	0.88
last mark	0.64	0.76	0.98	0	0.96	0.96	0.90
num attempts	0.44	0.59	0.52	0	0.94	0.54	0.52
time taken	-31*	0.78	0.07	-100*	0.74	0.17	0.18
single attempt	yes	no	yes	none	no	yes	yes

The features typical of each of the clusters allow us to interpret the general behaviour captured in these clusters. These are summarised in Table 3 and discussed in more detail below. Note that we refer to the following five grade categories from here on: High Distinction (HD), mark of 85 or above; Distinction (D), mark between 75 and 84; Credit (CR), mark between 65 and 74; Pass (P), mark between 50 and 64; Fail (F), mark below 50.

Table 3. Brief description of submissions clusters

Cluster	Typical Behaviour for the submission
0	Early start, steady improvement from CR to D.
1	Early start, strong first attempt.
2	No submission made
3	Normal start, steady improvement from CR to HD.
4	Normal start, strong first attempt.
5	Late start, strong first attempt.

Cluster 0: Attempts in this cluster were started early and progressed for a long and had a high number of compile errors in the attempts. They contained a medium number of attempts, and their improvement was moderate: attempts began with around a credit and improved to a distinction. (9% of vectors were in this cluster).

Clusters 1, 4 and 5: these represent cases where students performed well in the weekly task and began early, neither early nor late, and late respectively. Students, when in any of these three clusters, on average began with an initial and final mark of HD. However, Cluster 1 students had the highest average mark in both cases (96-98), followed by Cluster 4 (92-96), then Cluster 5 (88-90). These students usually made a medium number of attempts with a small number of compilation errors over a small amount of time. (19, 13 and 11% of instances respectively).

Cluster 2: This cluster represents cases where students did not attempt the task. (31% of cases).

Cluster 3; The high number of submissions and time taken suggests students, when in this cluster, put in the most effort. Improvement was typically large – from around a low credit (68) to an HD (96). The majority of these students' attempts were not late, and there were a low number of compilation errors. (17% of instances).

Intuitively, we would describe Clusters 0 and 3 as the behaviours that make best use of the autograding system, by making use of the feedback to achieve a significantly higher final grade.

Clusters 1, 4 and 5 are interesting because these behaviours are unlikely to benefit from being able to make multiple attempts, since early attempts are already of a high quality. It might be that students who found a task easy to complete in one week may not feel the need to invest time early in subsequent.

Figure 2 shows the general distribution of behaviours each week. We can see that many students were in Cluster 1 in the first week, probably due to the simplicity of the task, and that the number of students who did not submit at all (Cluster 2) is similar from week 2 to week 8, but increasing towards the end of the semester, especially in weeks 9, 10 and 11. This can be explained by the fact that these weeks are heavy in assignment deadlines in all the courses, including this course.

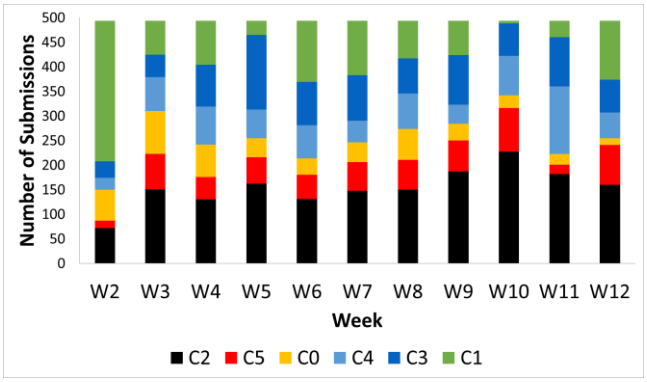

Figure 1. Number of students in each submission cluster each week. Order of clusters follows order discussed in Section 4.4.

4.2 Evolution of students with different exam grades

Figure 3 shows the relationship between the submission clusters each week and the final exam grades of students corresponding to those clusters.

We chose to study the relationship of the submission clusters with the final exam since it is the main and most comprehensive assessment component in the course. It is worth 60% of the final mark, covers all topics and is highly correlated with the final mark for the course. Here we use the same grade categories as previously: HD, D, CR, P, F. NA denotes students who did not sit the exam. There is a minimum requirement policy of scoring at least 40% in the final exam to pass the course: this means that even if students scored very high during the semester (say, 100% of 40), they would fail the course if they scored less than 40% at the final exam (say 30% of 60), even though their raw mark would be above a pass (58%).

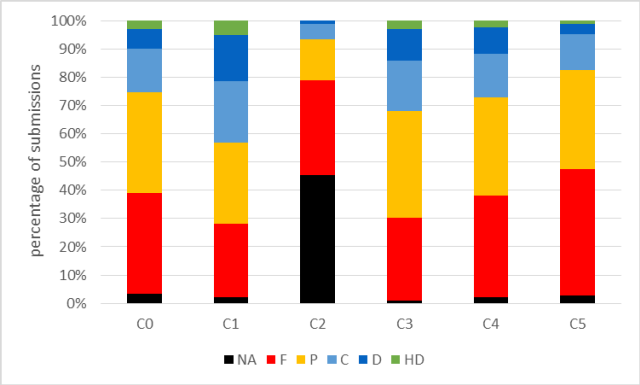

Figure 3. Percentage of submissions in each submission cluster each week with the submitting student's final exam grade

We can see that the students who obtained HD and D in the exam were often in Cluster 1 during the semester and also sometimes in Clusters 4 and 3. These clusters corresponding to the best performing students during the semester, with Cluster 1 containing the students who start early with a very high initial mark, Cluster 4 – the students who start normally with a high mark and Cluster 3 – the students who start early or normally from an average mark and work very hard to improve their submissions.

The students who obtained CR and P at the exam did not show a predominant behavioural pattern during the semester when completing the weekly tasks – they belonged to all clusters. However, more P than CR students were in Cluster 2 (the cluster of students who did not submit), for all weeks. In contrast, very few of the CR students were in Cluster 2 in the early weeks although this number increased after week 8.

A large proportion of the students who failed the exam were in Cluster 2 during the semester, but there are failing students in all behavioural clusters. The students who did not sit the exam are predominantly from Cluster 2 and, from Figure 1, their number is relatively stable from week 2 to week 12, which shows that most likely these students dropped out early in the semester.

4.3 Evolution of students from a given cluster

We can also follow the evolution of the students from a given cluster from a specific week. For example, starting with the six clusters from Week 3, we can analyse each cluster separately and investigate where the students from each cluster go in the subsequent weeks, as shown in Figure 2.

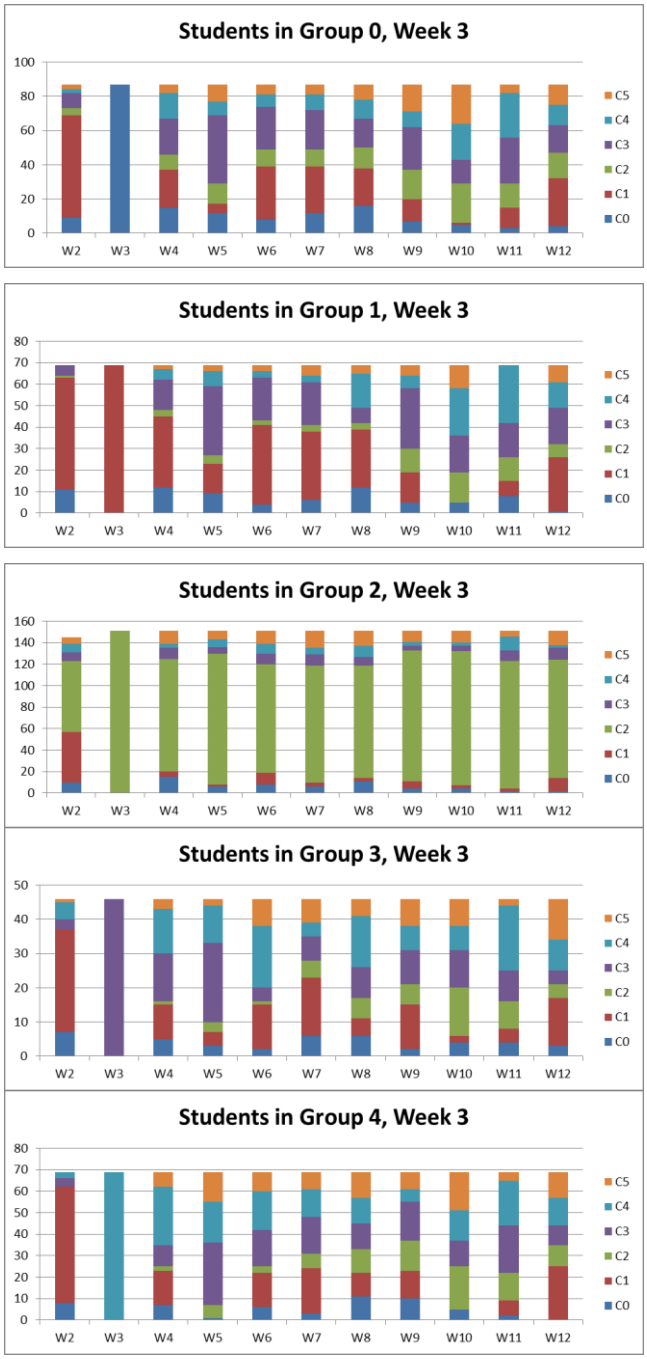

Figure 2. Analysing the six clusters from week 3 separately - percentage of students and each each cluster in subsequent weeks

The graphs show that the students from Cluster 0 in week 3 were mainly in Clusters 1 and 3 in the following weeks, i.e. they were able to achieve a higher mark on the weekly tasks compared to week 3. The students from Cluster 1 in week 3 mainly stayed in the same cluster or moved to Cluster 3, i.e. had to put more effort to maintain high marks. The students from Cluster 2 in week 3 (the non-submitting students) stayed in the same cluster with very few exceptions. The students from Clusters 3 and 4 together stayed in these clusters, and the students from Cluster 5 in week 3 moved

between Clusters 3, 5 and 2 during the semester, i.e. they were not always able to achieve high mark, possible because they started late, and also did not submit in some weeks, e.g. week 10.

We can clearly say that extracting patterns based on visual analysis of the graphs is difficult. This motivated our second clustering of behavioural data described in the next section.

4.4 Comparing the clusters in the middle and end of the semester

To better understand the stability of the clusters over time, we conducted clustering in the middle of the semester (after week 7) using the same method as described in Sec 4.1. We then compared the new clustering to the old clustering, described in Sec 4.1, to determine whether the end-of-semester clusters had already formed midway through the semester. Note that the clustering in both cases is done using all the available data at that time point, i.e. the mid-semester (early) clustering uses the data from week 2 to week 7, and the end-of-semester (end) clustering uses the data from week 2 to week 12.

In both cases, we followed the same clustering procedure – one example represents one submission. We paired each early cluster with a corresponding end cluster, seeking to maximize the overlap between the matched clusters.

More precisely, we considered the bijection, *m*, from the set of end clusters to the set of early clusters which minimized the distances between the centroids of each late cluster c_i and the paired early cluster $m(c_i)$. We then defined the accuracy of *m* on an early cluster $m(c_i)$ to be the proportion of submissions in end cluster c_i that were also in early cluster $m(c_i)$. That is,

$$accuracy(m(c_i)) = \frac{|S(m(c_i)) \cap S(c_i)|}{|S(c_i)|}$$

where *i* is an integer from 0 to 5, $S(x)$ denotes the set of submissions assigned to the cluster *x*, and $|X|$ denotes the number of elements in set *X*.

The chosen bijection gives the accuracies shown in Table 4. We can see that the accuracy of the mapping of four of the end clusters (1, 2, 3 and 5) is very high (≥90%). This is to be expected of Cluster 2 as all non-attempts are forced into their own cluster. However, this is not the case for Cluster 1, Cluster 3 and Cluster 5, and the high accuracy indicates that these clusters had already formed midway through the semester. End Cluster 4 had also emerged in week 7, as evident by relatively high accuracy of the mapping to it (76%), but had not stabilized yet. The mapping of end Cluster 0 had a low accuracy, indicating that this cluster had not yet been formed in week 7. A closer examination shows that the students in early Cluster 0 used strategies typical not only of end Cluster 0 but also of end Clusters 1 and 4, as well as end Clusters 5 and 3, to a lesser extent.

Table 4. Accuracy of each cluster in the middle of the semester (week 7) relative to the end of the semester (week 12)

End cluster (week 12)	0	1	2	3	4	5
Accuracy in week 7	13%	90%	100%	91%	76%	97%

In summary, the comparison of the end clusters from week 12 with the early clusters from week 7 shows that most of the end clusters had already formed or emerged in the middle of the semester. We can use these results to provide feedback to students in the middle of the semester and devise appropriate early intervention.

4.5 Behavioural evolution in time

The submission clustering in section 4.1 gave us clusters capturing behaviour per student per weekly task. An interesting question is how each student's behaviour evolved during the semester in regards to their weekly task. In order to explore this question, we performed an additional clustering to identify groups of students with similar submission behaviours over the weeks. The features used for this clustering try and capture the variety and frequency of behaviours (in terms of submission clusters found in 4.1). Note that features, c0-c5 count, are dependent, since the number of weeks are fixed. However, as previously, we maintain all to preserve symmetry. These features are described in Table 5. K-means clustered students into 6 groups, where the number of clusters was determined empirically. The centroids of this new clustering, which we call behavioural clustering, are shown in Table 6.

Table 5. Features used in behavioural clustering

Feature	**Description**
num_clusters	Number of submission clusters a student's submission belonged to over the semester
c0_count	Number of weeks where a student's submission belonged to behavioural cluster 0
c1_count	Number of weeks where a student's submission belonged to behavioural cluster 1
c2_count	Number of weeks where a student's submission belonged to behavioural cluster 2
c3_count	Number of weeks where a student's submission belonged to behavioural cluster 3
c4_count	Number of weeks where a student's submission belonged to behavioural cluster 4
c5_count	Number of weeks where a student's submission belonged to behavioural cluster 5

Before we describe these clusters, we also examined the relationship between final exam marks and a student's behavioural cluster. Figure 3 shows the percentage of students in each behavioural cluster receiving each of the possible exam grades: HD, D, CR, P, F and NA, where NA indicates that a student did not sit the final exam. The behavioural clusters in this figure have been ordered from lowest to highest based on the percentage of students passing the final exam in those clusters (i.e. behavioural clusters 3, 4, 1, 5, 2, then 0). We see in general that the proportion of passing students that receive higher bands increases, as well as the proportion of students who sit the final exam.

Table 6. Behavioural cluster centroids

		Behavioural Cluster Number					
Feature	**Full Data**	**0**	**1**	**2**	**3**	**4**	**5**
num_clusters	3.92	4.17	5.13	4.41	1.31	3.48	4.42
s0_count	0.99	1.00	1.51	1.49	0.14	0.66	0.83
s1_count	2.06	5.26	1.44	2.21	0.11	0.93	1.48
s2_count	3.44	0.69	2.12	0.72	10.65	7.20	0.63
s3_count	1.83	2.22	1.61	4.49	0.04	0.52	2.14
s4_count	1.46	1.42	1.43	1.31	0.03	0.41	4.52
s5_count	1.23	0.42	2.90	0.77	0.04	1.28	1.41

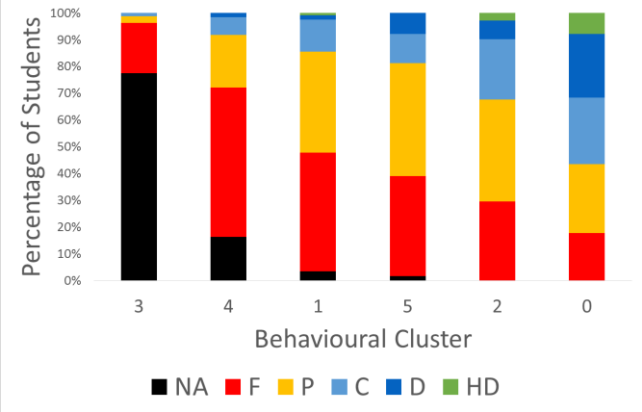

Figure 3. Exam performance of students in behavioural clusters, ordered in increasing proportion of students passing their final exam

We note that over 80% of students in Behavioural Cluster 0, which comprised 20.4% of the cohort, passed the final exam – the highest percentage of all the secondary clusters. In addition, over 50% of students in this behavioural cluster received at least a credit.

Behavioural Cluster 2 had the next highest pass rate of around 70%. The proportion of students receiving high bands in this cluster was lower than Behavioural Cluster 1, but greater than in other clusters.

Using the cluster centroids in Table 6, the weekly behaviours typical of different behavioural clusters are summarised below, in the cluster order used in Figure 3.

Behaviour Cluster 3: These students belonged to an average of 1.3 different clusters throughout the semester. 96.8% of the time they were assigned to Submission Cluster 2, indicating that they almost never completed their weekly tasks. These students may have dropped out of the course during the semester. (16.2% of students).

Behavioural Cluster 4: These students oscillated between an average of 3.5 clusters throughout the semester. 65.4% of the time, they fell into submission Cluster 2, indicating that they frequently did not complete their weekly tasks. However, these students belonged to submission Cluster 5 11.6% of the time, suggesting they sometimes started late but still performed well. From this, we see that these students are possibly quite capable, but do not put much effort into their weekly tasks.

Behavioural Cluster 1: These students were in an average of 5.1 submission clusters over the semester. Cluster 5 was the most common submission cluster, which students were in 26.3% of the time, followed by Cluster 2 (19.3%), Cluster 3 (14.6%), Cluster 0 (13.8%), Cluster 1 (13.1%) and Cluster 4 (13.0%). Thus these students often started late but did well, but also often didn't submit at all. These students sometimes worked hard and achieved high marks, sometimes worked hard without achieving high marks, sometimes began early and did very well and sometimes began neither early nor late and did well. These students displayed inconsistent behaviour over the weeks, sometimes putting in a great amount of effort and sometimes not trying at all. (24% of students).

Behavioural Cluster 5: These students belonged to an average of 4.4 different clusters over the semester. They fell into submission Cluster 4 the most often - around 41.1% of the time – followed by submission Cluster 3 (19.5%), Cluster 1 (13.5%) and Cluster 5 (12.8%). Thus these students very often started their weekly tasks neither early nor late and did well, commonly started early and worked hard until they did well, sometimes started early from a high mark and sometimes started late from a high mark. (13 % of students)

Behavioural Cluster 2: These students belonged to an average of 4.4 different submission clusters over the semester, with Cluster 3 being the most common (40.8%), then Cluster 1 (20.1%), Cluster 0 (13.6%) and Cluster 4 (11.9%). Thus, these students commonly began early with a medium mark, worked hard and achieved good marks. They also often started early from a high mark, sometimes worked hard without achieving a high mark and sometimes started neither late nor early with a high mark. These are hard-working students who often found the tasks challenging, but still did fairly well in them.

Behavioural Cluster 0: Finally, in the behavioural cluster with the highest final exam pass rate, students oscillated between an average of around 4.2 clusters in the course of the semester. They were in submission Cluster 1 47.8% of the time, Cluster 3 20.2% of the time, Cluster 4 12.9% of the time and cluster 0 9.1% of the time. This suggests these students started early with high marks around half the time. They often started early with medium marks, but worked hard until they achieved a high mark and sometimes started neither late nor early, achieving high marks. Occasionally they worked hard without achieving high marks. (20% of students). These students often did well on their first submission but, when they didn't, they worked hard to achieve high marks.

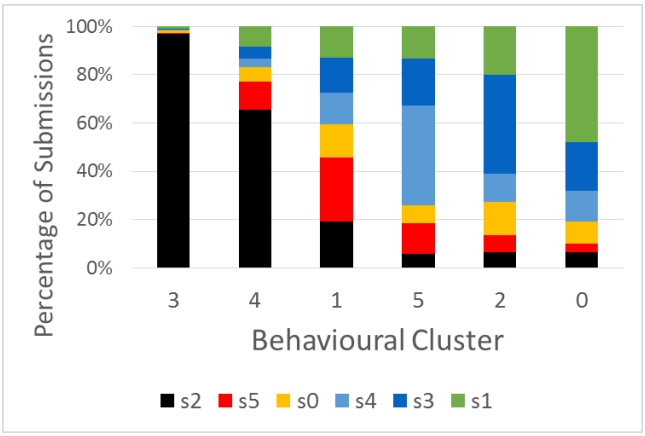

Figure 4. For each behavioural cluster, the percentage submissions in each submission cluster (s0, s1, s2, s3, s4, s5)

4.5.1 General Trends

By analysing behavioural clusters and the most common submission clusters the students' submissions were in, we noticed general trends as the final exam pass rate increased. For example, submissions in Submission Cluster 2, characterised by no submission attempt, were most common in students in behavioural clusters with the lowest pass rate. On the other hand, Submission Cluster 1 (early start, strong first attempt) was most common in behavioural clusters with higher pass rates. We used these trends to order the submission clusters: Submission Clusters 2 and 5, being the most and second most common submission clusters in poorly performing behavioural clusters, were placed on the bottom of the scale. Of the remaining four submission clusters, Submission Cluster 0 was least common in the top three behavioural clusters,

and so came next on the scale. This was followed by Submission Clusters 4, 3 and then 1, which became more prevalent in higher performing behavioural clusters. Figure 4 shows the percentage of submissions in each behavioural cluster that fell into each submission cluster. The behavioural clusters are ordered based on pass rate, and the submission clusters are ordered as described above. The prevalence of each submission cluster in different behavioural clusters is summarised in Table 7.

Table 7. Submission Clusters Typical of each Behavioural Cluster

Submission Cluster	Common in Behavioural Clusters with	Submission Cluster Description
0	Many different pass rates	Average students, medium/high effort.
1	High pass rates	Excellent students who started early from a very high mark.
2	Low pass rates	Did not submit.
3	High pass rates	Hard working students – from CR to HD.
4	Medium pass rates	Good students who started neither early nor late from a mid HD.
5	Low pass rates	Good students who started late from a low HD and improved slightly.

4.5.2 The median

We can also visualise the evolution of student behaviour over the semester in a meaningful way. We looked at the weekly behaviour of students in each behavioural cluster each week and found the "median" behaviour. This was achieved by taking the median of each original feature for these students, such as the first mark, last mark, time taken and percentage of early submissions. We then used this to create a median vector, and found which submission cluster the vector belonged to. We repeated this for all behavioural clusters and plotted the results. This can be seen in Figure 5. Note that submission clusters were previously ordered so the higher the submission cluster the more typical it is overall of the behavioural clusters with the highest pass rate.

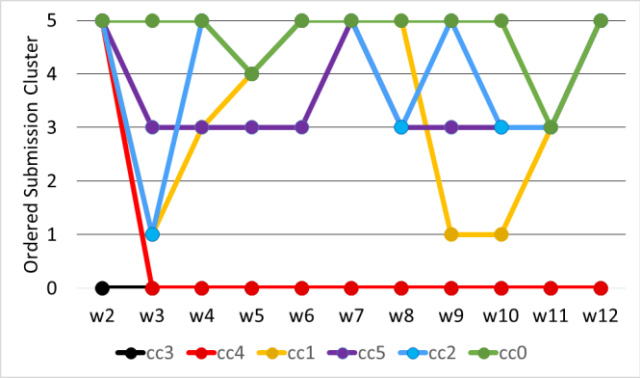

Figure 5. Changing student behaviour over the semester. Each colour represents a behavioural cluster. The median behaviour of students each week (i.e. the median submission cluster) is shown. The submission clusters are ordered so that higher corresponds to better performance.

Rather than the secondary clusters slowly diverging over time, we notice a clear separation from as early as week 3. The secondary clusters with the lowest (secondary clusters 3 and 4) and highest (secondary cluster 0) pass rates are already distinguishable from the other clusters at this time. This early separation of behaviours could facilitate early identification of students at risk of failing or performing poorly, allowing for intervention.

5. DISCUSSION

The scheme in our analysis can be separated into two parts:

(i) A submission clustering, where the approach and performance of each student in each weekly submission is treated as independent and then clustered to give typical task-level behaviours.

(ii) A behavioural clustering, where students are clustered based on the submission clusters they were in over the entire semester.

Through the example of a junior computer science course, we demonstrated the usefulness of this double-clustering method in allowing us to identify some important approaches students in this course took to their weekly tasks. We found that many students started sufficiently early and invested time to improve their attempts based upon instant feedback they received from the autograding system, benefiting from a significant improvement in the quality of their final attempts (Clusters 0 and 3, 26%). We also found that students often found the task sufficiently easy and that further improvements were of little value (Clusters 1, 4 and 5, totaling 43%), and that it was also common for students to not attempt the tasks at all (Cluster 2, 31%). A broader application of this analysis over multiple units of study and across multiple offerings of the same course would be useful in understanding how common such behaviours are in general as opposed to this specific offering.

Through the behavioural clustering, we were able to identify common behavioural patterns over the entire semester, and to draw links between these patterns and final exam outcomes. In particular, we identified behavioural patterns associated with high and low final exam grades. For example, students in behavioural clusters with high pass rates tended to consistently start early with a high mark, or start early and work hard until a high mark was achieved. Conversely, students in behavioural clusters with low pass rates often did not submit their tasks at all. Knowledge of the relationship between behavioural patterns and exam performance is essential in the identification of students at risk of performing poorly and important in the structuring of a course to maximise student learning and performance.

We compared submission clusters that used all data up to week 12 with submission clusters that used all data up to week 7, and found that they were quite similar. This suggests that the typical task-level behaviours of students did not vary much at the end of the semester and that, as a consequence, these behaviours could be identified early on in the semester. Moreover, we saw that the term-long behavioural clusters we found did not slowly diverge over time, but rather there was an immediate difference from as early as week 3. This suggests that both the submission and behavioural clustering could be performed early in the semester, with potentially similar results to the end of semester, allowing for early identification of students at risk of performing poorly and early intervention. We suggest an avenue of future research could be to apply this technique midway through the semester and evaluate its

effectiveness in facilitating interventions that could improve student outcomes.

We also suggest investigating how effective this method can be in general, by applying it to courses with different assessment structures and content, and also to compare the results obtained through these clustering methods to traditional measures of behaviour and engagement, such as tutorial attendance and feedback surveys, to evaluate how well they corroborate.

Although the reported analysis is for data from a system for assessing computer code submissions, it could just as readily applied to other systems in which students can make multiple submissions in response to feedback. For instance, many Learning Management Systems provide multiple-choice style questions for which students can receive feedback about their choices, and this style of question could be used in any discipline. Our analysis depends only upon records of the time and quality of each submission. While we include details such as number of compile errors as one measure of quality, this could readily be substituted with other measures.

6. CONCLUSION

In this paper we have presented a method for analysing student behaviour and the evolution of this behaviour over the semester, using data from autograding system logs. We have shown that this method can be useful in identifying common weekly behaviours of students, and following the changes of such behaviours over the semester. We have discussed the relationship between these behaviours and final exam results, and demonstrated how these behaviours might be detectable early enough in the semester for instructors to intervene. As such, we believe that the techniques discussed here may be implemented and improved upon to realise the full potential of increasingly common autograding systems in facilitating real improvement in student outcomes.

7. ACKNOWLEDGMENTS

This work was funded by the Human-Centred Technology Cluster of the University of Sydney.

8. REFERENCES

[1] Sherman, M., Bassil, S., Lipman, D., Tuck, N. and Martin, F. 2013. Impact of autograding on an introductory computing course. *Journal of Computing Sciences in Colleges*, 28 (6), 69-75.

[2] Enstrom, E., Kreitz, G., Niemela, F., Soderman, P., and Kann, V. 2011. Five years with Kattis - using an automated assessment system in teaching. *In Proceedings of the Frontiers in Education Conference (FIE)*, IEEE.

[3] Milojicic, D. 2011. Autograding in the cloud: interview with David O'Hallaron. I*EEE Internet Computing* 15 (1), 9-12.

[4] Gramoli, V., Charleston, M., Jeffries, B., Koprinska, I., McCrane, M., Radu, A., Viglas, A., and Yacef, K. 2016. Mining autograding data in computer science education. In *Proceedings of the Australasian Computing Education Conference (ACE)*.

[5] Koprinska, I., Stretton, J., and Yacef, K. 2015. Predicting student performance from multiple data sources. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, LNCS 9112, 678-681.

[6] Koprinska, I, Stretton, J., and Yacef, K. 2015. Students at risk: detection and remediation. In *Proceedings of the International Conference on Educational Data Mining (EDM)*, 512-515.

[7] Perera, D., Kay, J., Koprinska, I., Yacef, K., and Zaiane, O. 2009. IEEE Transactions on Knowledge and Data Engineering, 21(6), 759-772.

[8] Tomkins, S., Ramesh, A., and Getoor, L. 2016. Predicting post-test performance from online student behaviour: a high school MOOC case study, In Proceedings of the International Conference on Educational Data Mining (EDM).

[9] Radu, A. and Stretton, J. PASTA, School of Information Technologies, University of Sydney, http://www.it.usyd.edu.au/~bjef8061/pasta/. Accessed: 2016-05-02

[10] INFO1105: Data Structures (2015 - Semester 2). https://cusp.sydney.edu.au/students/view-unit-page/uos_id/79883/vid/309891. Accessed: 2016- 03- 05.