Modelling student online behaviour in a virtual learning environment

Martin Hlosta[†]   Drahomira Herrmannova[†]   Lucie Vachova[††]
Jakub Kuzilek[†]   Zdenek Zdrahal[†]   Annika Wolff[†]

Knowledge Media Institute, The Open University[†]
Walton Hall
Milton Keynes, MK7 6AA
{martin.hlosta; d.herrmannova; jakub.kuzilek;
z.zdrahal; annika.wolff}@open.ac.uk

University of Economics, Prague[††][*]
Department of Exact Methods, Faculty of Management
Jarosovska 1117/II
Jindrichuv Hradec, 377 01
vachova@fm.vse.cz

ABSTRACT

In recent years, distance education has enjoyed a major boom. Much work at The Open University (OU) has focused on improving retention rates in these modules by providing timely support to students who are at risk of failing the module. In this paper we explore methods for analysing student activity in online virtual learning environment (VLE) – *General Unary Hypotheses Automaton (GUHA)* and *Markov chain-based analysis* – and we explain how this analysis can be relevant for module tutors and other student support staff. We show that both methods are a valid approach to modelling student activities. An advantage of the Markov chain-based approach is in its graphical output and in the possibility to model time dependencies of the student activities.

Categories and Subject Descriptors

D.4.8 [**Performance**]: Modelling and Prediction;
H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Student Data, Distance Learning, Predictive Models, Machine Learning, Information Visualisation

1. INTRODUCTION

The recent years have seen a massive growth of different possibilities of online education, such as the well known massive open online courses (MOOCs) [Cormier,2008]. The concept of distance education is however not new. The Open University is an institution with over forty years of experience with distance education, historically based on off-line materials and nowadays making an increasing use of the Internet. The great advantage of the online courses is in the fact they are accessible to virtually anyone with Internet access.

The other side of the coin is that the retention rates in these courses are often low. [Koller et al.,2013] mention, that an average retention rate of a Coursera[1] course is around 5%. The situation at traditional universities as well as at The Open University is significantly better, however, there is still a room for improvement.

There might be many reasons for the low retention rates, from the fact that the online courses are often offered to anybody interested to the fact that the performance of each student depends almost exclusively on how much are they willing to study on their own at home. Our work at The Open University aims at analysing students' activities in the online courses in order to gain insight into their behavioural patterns, which can be utilised for building prediction models.

1.1 Problem Description

The Open University[2] is the biggest university in the United Kingdom, offering several hundred distance learning modules, which can be studied both as standalone modules or as part of a university degree. Anybody can sign up for a module provided by The OU, without any previous education whatsoever. The students receive their study material and submit their assignments through an online virtual learning environment.

Students participating in a module are generally split into smaller study groups of no more than few tens of students, typically according to their geographic location. Each group has an assigned tutor. The tutors grade the students' assignments and exams, answer their questions in the online forums, provide general advice and guidance, etc.

In order to support the students who are at risk of failing the module The OU also implements various interven-

---

[*]This work was carried out at the Knowledge Media Institute.

[1]https://www.coursera.org/, a well known and one of the biggest platforms providing open online courses.
[2]http://www.open.ac.uk/

tions (such as phone calls from a specialised student support teams) during the course of the module. Because the number of students studying each module can reach several thousand (the modules used in our analysis have enrolment of around two thousand students) and the resources available for the interventions are limited, the interventions have to be carefully planned. Therefore, an important question one might ask is how to identify students at risk of failing the module so that the intervention is meaningful and efficient.

Improving student retention through these focused interventions and helping the tutors to focus on the students, who require help, provides many benefits, from improved student satisfaction to financial savings for the university. In order to support the identification of students who are currently at risk, we utilise several statistical and machine learning methods. The available data contain both the information about the students' activity in the VLE as well as their demographic information. However, for modelling student behaviour, only the data from the VLE was used. A more detailed description of our data set can be found in Section 3.

2. PREVIOUS WORK

The current work builds on previous research done at The Open University. The initial experiments with machine learning techniques were using the VLE and assessment data [Wolff and Zdrahal,2012]. One of the main findings of this research was that decision trees generally outperform the other methods [Wolff and Zdrahal,2012]. This research also included creation of a dashboard providing the university staff with real-time information about student performance.

Additional methods were tested in [Wolff et al.,2013b] and in [Wolff et al.,2013a]. In the latter work, demographic data were added to the predictions, however this research did not confirm that this data provide a significant increase in performance.

3. DATA SPECIFICATION

The analyses we performed were done using real data of several modules from The Open University. We examined a number of subsequent presentations of each module.

The available data contain two types of information:

- Information about the results of student assignments (TMAs – tutor marked assignments). There are several assignments in each module, typically between five and seven. Generally, the module is ended by a final exam.

- Data about student activity from the virtual learning environment (VLE).

The VLE data are aggregated by days and content type (e.g. forum, wiki, resource, ...). This means that for each day we know how many times did the student interact with given content type. For our analysis we summarise the data by weeks and content types. Summarising the data by weeks seemed to be reasonable, it simplifies our analyses without loosing too much detail. The features generated by the summarisation and used in the methods are:

- click counts aggregated by week,

- click counts aggregated by week and content type,

- binary flags indicating whether student was active in the VLE and in various content types.

4. METHODS

For analysis of student behaviour in the virtual learning environments, we have used two different approaches – GUHA [Hájek et al.,1966] and modelling based on Markov chains [Norris,1998].

4.1 Activity types analysis

As mentioned in Section 3, the VLE data contain information about the type of content the student accessed. The content type can be for example *forum, wiki, resource, quiz*, etc. In total there are 11 different content types. Using the binary flags, indicating whether student was active in given week and content type, we utilised Bayes Theorem [Bishop and Nasrabadi,2006] for determining the probability that the student will fail to complete the module. Moreover, we analysed each of the content types in terms of mean number of students succeeding in the module based on activity or inactivity in the given content type. Based on this investigation we have selected a set of content types which significantly influence students' performance, these were then used in the further analyses.

4.2 GUHA

General Unary Hypotheses Automaton (GUHA), originally published in [Hájek et al.,1966], is one of the oldest data mining methods for automatic discovery of new interesting hypotheses from the data. To achieve this goal, GUHA uses various different procedures (ASSOC, IMPL, CORREL). The choice of the procedure to use depends mainly on the user needs and his experience. We have selected the *ASSOC* procedure [Rauch and Simunek,2001], which allows to discover interesting associations between attributes in the data. The interestingness of the association is mostly based on their co-occurrence. The ASSOC procedure allows to limit the resulting rules by specifying constraints on the attributes. This property is important for our field of interest.

For our research, we used the ASSOC procedure that is implemented in the 4ft-Miner module within the Lisp-Miner software tool [3]. The specification of the constraints enabled us to restrict the rules only to those, which cover students that fail or succeed in the TMA. We used three basic types of features for the analysis introduced in Section 3.

The search space for both types of binary flags was reasonable to perform analysis. For weekly aggregated counts, it was necessary to reduce the search space via interval discretisation. For this purpose, we utilized LISp-miner and unsupervised equal frequency discretisation [Wong and Chiu,1987].

GUHA method produce large set of hypotheses. The example of such results with the categorized binary flags are depicted in the Figure 1. Unfortunately the information contained in the output is difficult to interpret. Moreover, the information of the time dimension is lost and this is even worse when using various content types. For us, this was the motivation to look for another modelling method.

4.3 Markov chain-based analysis of student activity
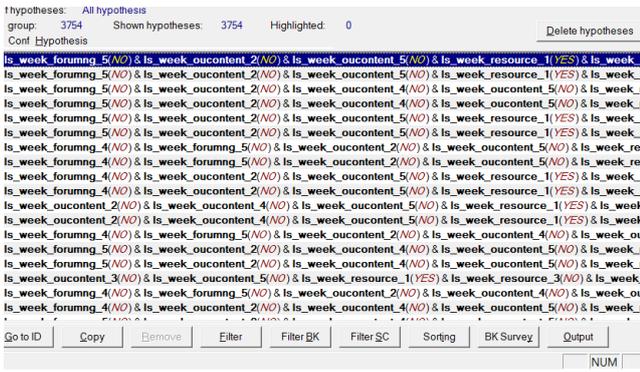
Figure 1: Screenshot with discovered rules from LISp-miner.

	Scenario	TMA 1 not submit	TMA 1 success
1.	zero in any of 0 - 4	51.6	42.6
2.	**zero only in 1 - 4**	**95.7**	**4.3**
3.	**zero only in 2 - 4**	**92.3**	**7.7**
4.	**zero only in 3 - 4**	**100.0**	**0.0**
5.	zero only in 4	54.3	45.7
6.	zero only in 0	15.4	71.8
7.	**zero only in 0 - 1**	**6.7**	**86.7**
8.	zero only in 0 - 2	15.4	69.2
9.	zero only in 0 - 3	57.1	42.9
10.	zero in at least one of 0 - 3, non-zero in 4	18.2	71.6
11.	zero in at least one of 0 - 2, non-zero in 3 - 4	14.6	75.1
12.	zero in at least one of 0 - 1, non-zero in 2 - 4	9.3	80.4

Table 1: Summary of examined situations for students behaviour evaluation (data in % of total number of course students)

In this part of analysis we examined the differences in intensity of student activity between students who were successful in the first TMA (TMA 1) and those who did not submit TMA 1. Students who failed in TMA 1 are not included in this analysis, due to the fact that they represent only a small portion of students who submit TMA 1. In this stage of research we analysed only students who at least once had zero VLE activity in one of the weeks under consideration. In VLE passive students represent the specific group important while looking for potentially at-risk students. On this group we have studied different scenarios - the list of them is displayed in Table 1. Moreover, the Table 1 shows percentage of students who behaved according to given Scenario and were successful in TMA 1, in comparison to those who did not submit TMA 1, as well. Numbers in column Scenario represents particular weeks of course.

Based on the data in Table 1, we can identify behaviour of at-risk students. This is evident especially in scenarios 2, 3, 4 and 7. This shows us that students who tend to reach zero VLE activity in later weeks are more probable not tu submit (scenarios 2, 3, 4). On the other hand those who have zero VLE activity in earlier weeks and later start to show interest
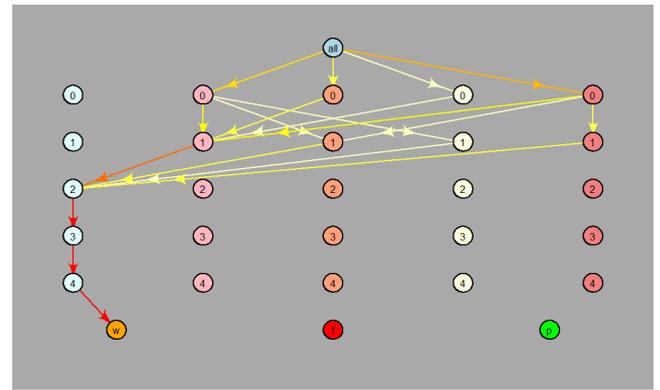

Figure 2: Representation of students behaviour in scenario 3 (TMA 1 was not submitted)
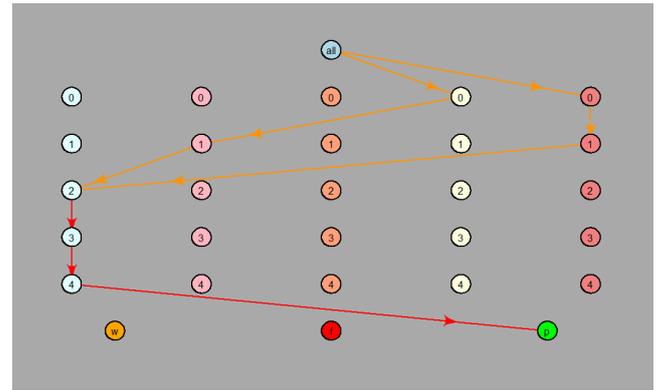

Figure 3: Representation of students behaviour in scenario 3 (TMA 1 was passed)

represented by VLE activity, raise their chance to success in TMA 1 (scenarios 6, 7, 8).

Figures 2 and 3 specify more closely the situations from the scenario 3 (with the TMA 1 not submitted, and TMA 1 passed, respectively). Colour tones of arrows differ (from white to red) depending on the percentage of students who moved in given direction. The more red the colour the bigger the percentage of students it represents. The rows represent the weeks in which VLE activity was measured. First row shows activity before the beginning of the course (Week 0), the other four rows capture the VLE activity in week 1, 2, 3 and 4 respectively. The columns represent the intervals of VLE activity - different colour of the node mean different interval. First column is zero VLE activity, while in other columns the activity is divided into intervals with the cut points in multiples of 30.

This type of analysis enables us to look for specific patterns in students' behaviour. In the similar way we can analyse the different types of VLE activities as shown in Figure 4. In this case the nodes represent not the intensity of student activity, but capture the interest of student in specific content type. And (unlike the previous two figures) this directed acyclic graph as a whole depicts the Markov chain [Norris,1998].

5. CONCLUSION
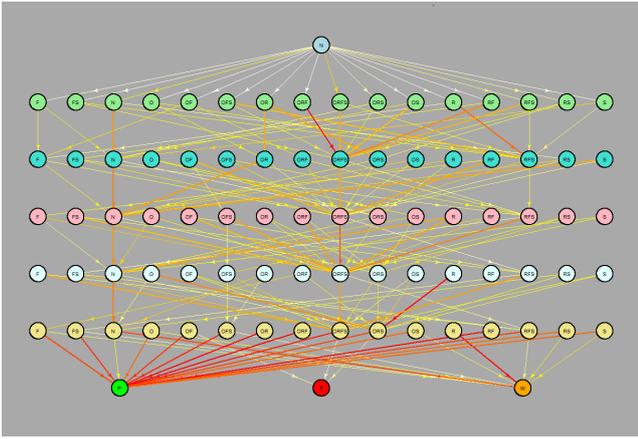
Figure 4: Markov chain for the various activity types combinations in weeks 0-5.

In this paper we have examined two methods for analysing activity of students in the online virtual learning environment before the first tutor marked assignment – GUHA and Markov chain-based graphical models. Both methods provide useful insights into the students' behaviour during their studies. The benefit of the latter lies mostly in its graphical output, which might be easier to interpret and could potentially provide support in planning interventions, and in the possibility to model time dependencies of the student activities. We believe that the understanding of the student behavioural patterns will also be useful for building better predictive models of student performance.

References

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 1. springer New York.

Dave Cormier. 2008. The CCK08 MOOC–connectivism course, 1/4 way. *Dave's Educational Blog*, 2.

Petr Hájek, I. Havel, and Michal Chytil. 1966. The guha method of automatic hypotheses determination. *Computing*, 1(4):293–308.

Daphne Koller, Andrew Ng, Chuong Do, and Zhenghao Chen. 2013. Retention and intention in massive open online courses: In depth. *EDUCAUSE*, June.

James R Norris. 1998. *Markov chains*. Number 2008 in Cambridge series in statistical and probabilistic mathematics. Cambridge university press.

Jan Rauch and Milan Simunek. 2001. Mining for association rules by 4ft-miner. In *INAP*, pages 285–295.

Annika Wolff and Zdenek Zdrahal. 2012. Improving retention by identifying and supporting "at-risk" students. *EDUCAUSE Review Online*, July/Summer.

Annika Wolff, Zdenek Zdrahal, Drahomira Herrmannova, and Petr Knoth. 2013a. Predicting student performance from combined data sources. In Alejandro Peña-Ayala, editor, *Educational Data Mining: Applications and Trends*, number 524 in Studies in Computational Intelligence, pages 175–202. Springer International Publishing, Cham.

Annika Wolff, Zdenek Zdrahal, Andriy Nikolov, and Michal Pantucek. 2013b. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Third Conference on Learning Analytics and Knowledge (LAK 2013)*. ISBN 978-1-4503-1785-6.

Andrew K. C. Wong and David K. Y. Chiu. 1987. Synthesizing statistical knowledge from incomplete mixed-mode data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(6):796–805, June.