Predicting Students' Performance: Incremental Interaction Classifiers

Miguel Sanchez-Santillan
Department of Computer Science
University of Oviedo
UO167275@uniovi.es

Rebeca Cerezo
Department of Psychology
University of Oviedo
cerezorebeca@uniovi.es

MPuerto Paule-Ruiz
Department of Computer Science
University of Oviedo
paule@uniovi.es

JCarlos Nuñez
Department of Psychology
University of Oviedo
jcarlosn@uniovi.es

L@S 2016, April 25-26, 2016, Edinburgh, Scotland UK
ACM 978-1-4503-3726-7/16/04.
http://dx.doi.org/10.1145/2876034.2893418

Abstract
One of the Educational Data Mining (EDM) main aims is to predict the final student´s performance, analyzing their behavior in the Learning Management Systems (LMSs). Many studies make use of different classifiers to reach this goal, using the total interaction of the students. In this work we study if it is possible to build more accurate classification models in order to predict the output, analyzing the interaction in an incremental way. We study the data gathered for two years with three kinds of classifying algorithms and we compare the total interaction models with the incremental interaction models.

Author Keywords
Educational Data Mining; eLearning; Classifiers

Introduction
Currently the LMS store the students´ interaction in databases or log files. The analysis of this interaction through Data Mining techniques allows us a better understanding of the students´ behavior in the LMS. EDM is a discipline that studies and develops methods to reach such goals [1].

Moodle created variables

Time Lesson. Minutes spent in lesson.

Time Quiz. Minutes spent in quiz.

Time Forum. Minutes spent in forum.

Days Lesson. The days that go by since block is unlocked until the student consults the lesson for the first time.

Days Quiz. Days until student consults the quiz for the first time.

Days Finish Quiz. Days until student sends the quiz for the first time.

Days Forum. Days until student checks the forum for the first time.

Days Post. Days until student writes in the forum for the first time.

Relevant Actions. Number of log's entries generated by the student.

Time spent. The days that go by since the students make the first interaction until they carry out the last interaction.

As a matter of fact, one of the EDM particular goals is to predict the students´ final performance [2] using classifiers. Such classifiers analyze the whole data to create accurate models for every possible result.

During the past few years researchers have carried out diverse studies about how to create classification models which will enable us to predict or improve the students´ performance. [3-5]. Using them means a great support for teachers, as they can be used as indicators of the students´ behavior in the LMS [6].

Generally speaking, these studies classification models are built using the total interaction that students generate through the LMS. However, inasmuch as the course year moves forward, the students' participation in the LMSs tends towards steep drop-offs and highly unequal patterns of participation [7].

Henceforth, in this work we study if it is possible to obtain more accurate classification models through the analysis of the students' interaction in an incremental way.

Case Study

We have used Moodle to create a training program which is divided in 11 blocks. Each week a new block is unlocked for 15 days and students must carry out three basic tasks to complete each part: read a lesson, complete a quiz and add a post in the forum.

We gathered the students' interaction in two different years; 2012 (N= 111 students) and 2013 (N= 84 students). The structure and procedure have been the same in both cases.

Data

We created a combination for every year and every block with the collected students' interaction. Therefore, the datasheet drawn from the last block represents the total students' interaction in the training program. Moreover, we generated another 11 datasheets with the total of the sample. As a whole, we got 33 different datasheets.

We have made use of 11 variables in our study. From Moodle's log we generated 10 variables for each block, as shown in the sidebar. We discretized students' final grade as "FAIL" and "PASS". Grades below 5 were labeled as "FAIL" meanwhile grades with a value of 5 or higher were labeled as "PASS".

We use WEKA in order to analyze the datasheets using three different algorithms; JRIP (a RIPPER implementation [8], J48 (an open source Java implementation of the C4.5 [9]) and Bayesian Network [10]. We chose these algorithms because the classification models that they provide are built in a different way; classification rules, decision trees and neural networks, respectively. Results have been tested through cross-validation with 10 folds.

2012 course results

Total of Blocks	JRIP	J48	BayesNet
1	75.67%	73.87%	77.47%
2	74.77%	72.97%	77.47%
3	**79.27%**	77.47%	**79.27%**
4	**81.98%**	75.67%	**81.08%**
5	75.57%	77.47%	**79.27%**
6	**77.47%**	75.67%	80.18%
7	76.57%	76.57%	77.47%

Total of Blocks	JRIP	J48	BayesNet
8	74.77%	74.77%	**79.27%**
9	**77.47%**	78.37%	**79.27%**
10	**78.37%**	78.37%	**79.27%**
11	76.57%	78.37%	78.37%

Table 1. Classifiers accuracy results for each incremental model using 2012 course datasheet.

As shown in Table 1, the classification of the total interaction with JRIP is overcome by models 3, 4, 6, 9 and 10. As regards to the algorithm J48 any model classifies better than the total model. To finish with, the results obtained by BayesNet are even further off. The total model is surpassed by the models 3, 4, 5, 6, 8, 9 and 10.

2013 course results

Total of Blocks	JRIP	J48	BayesNet
1	61.90%	**70.23%**	64.28%
2	61.90%	61.90%	65.47%
3	63.09%	61.90%	65.47%
4	63.09%	60.71%	65.47%
5	60.71%	57.14%	65.47%
6	61.90%	57.14%	65.47%
7	63.09%	60.71%	65.47%
8	**70.23%**	60.71%	64.28%
9	**66.66%**	59.52%	64.28%
10	61.90%	**64.38%**	**67.85%**
11	65.47%	64.28%	65.47%

Table 2. Classifiers accuracy results for each incremental model using 2013 course datasheet.

As shown in Table 2, the classification of the total interaction with JRIP is overcome by models 8 and 9. As regards to the algorithm J48 only models 1 and 10 surpass the total model. To finish with, the results obtained by BayesNet show that the total model is surpassed by the model 10.

2012 and 2013 datasheets merged

Total of Blocks	JRIP	J48	BayesNet
1	**70.25%**	66.15%	**70.76%**
2	67.69%	**67.18%**	**71.79%**
3	68.71%	**71.28%**	**70.25%**
4	68.71%	**70.25%**	**68.71%**
5	68.71%	**70.76%**	**70.25%**
6	**71.79%**	66.15%	**69.23%**
7	68.71%	**70.76%**	67.17%
8	**70.25%**	**67.18%**	66.15%
9	**70.76%**	**72.82%**	67.17%
10	65.64%	66.66%	**69.23%**
11	68.71%	66.66%	67.69%

Table 3. Classifiers accuracy results for each incremental model using 2012 and 2013 courses datasheets merged.

As shown in Table 3, the classification of the total interaction with JRIP is overcome by models 1, 6, 8 and 9. As regards to the algorithm J48 the total model is surpassed by the models 2, 3, 4, 5, 7, 8 and 9. To finish with, the results obtained by BayesNet show that the total model is surpassed by the models 1, 2, 3, 4, 5, 6 and 10.

Discussion and future work

Just as the results show, it is possible to obtain better classification models using an incremental interaction. Employing this kind of incremental models will be useful to improve the feedback that teachers get in respect of the students' performance. Nonetheless, this case study is based on a course already structured in different blocks, and so it will be needed to test the accuracy of incremental models in different cases. For instance in those courses in which the interaction is not only limited by time intervals, being this way the students able to access, more or less freely, to the different resources.

We have employed three different algorithms to test the generation of the models, nevertheless we are going to explore more algorithms and we are going to design a system that will automate the process of models generation with the maximum number of possible algorithms.

Acknowledgements
This work has been funded by the Department of Science and Innovation (Spain) under the National Program for Research, Development and Innovation: EDU2014-57571-P. We have also received funds from the European Union, through the European Regional Development Funds (ERDF); and the Principality of Asturias, through its Science, Technology and Innovation Plan (grant GRUPIN14-100 and GRUPIN14-053).

References
[1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, 1(1), 3-17.

[2] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

[3] Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums.Computers & Education, 68, 458-472.

[4] Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), 43-47.

[5] Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. In WOCN Conference (pp. 1-5). IEEE.

[6] Ranjan, J., Khalil, S. (2008). Conceptual Framework of Data Mining Process in Management Education in India: An Institutional Perspective. In Information Technology Journal, 7, 1, 16-23.

[7] Clow, D. (2013). MOOCs and the funnel of participation. In LAK Conference (pp. 185-189). ACM.

[8] Cohen WW (1995). Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning. pp. 115-123.

[9] S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M]. Nanjing: Nanjing University Press, 1997.

[10] R.R. Bouckaert. Bayesian Belief Networks: from Construction to Inference. Ph.D. thesis, University of Utrecht, 1995.