A promising classification method for predicting distance students' performance.

Diego García-Saiz
Universidad de Cantabria
Avda. Los Castros s/n
Santander, Spain
diego.garcia@unican.es

Marta Zorrilla
Universidad de Cantabria
Avda. Los Castros s/n
Santander, Spain
marta.zorrilla@unican.es

ABSTRACT

Predicting the students' performance is still a challenging task despite being one of the oldest and most popular applications of data mining in education. One of the problems encountered when analyzing data from e-learning platforms is that it presents statistical outliers as a consequence of how students work in online courses. It causes that classifiers are built with less accuracy than desired. To solve this problem we propose a new method to eliminate outliers as previous step to build the classifier. In this work we describe our meta-algorithm and compare its performance with respect to several well-known classification techniques. The comparison is evaluated in terms of accuracy, true positive and true negative rate. The results obtained shows that our approach produces more accurate models.

1. INTRODUCTION

Since the advent of learning platforms, their use in educational centers has been constantly growing. Unlike traditional teaching, one of the advantages which these systems have is they store a huge quantity of data which, adequately analysed, can help both instructors and students: instructors can discover information to evaluate the teaching-learning process [6]; and, students can receive suitable feedback about their dedication in the course [4] and recommendations in order to achieve the learning objectives [5].

In the literature, there are several works which compare classification techniques using educational datasets but none of them analyse the data distribution in order to detect outliers [3; 1]. In our experimentation, we have detected that despite the data is clean (free of human errors), there are instances which can be considered as outliers in the statistical sense (e.g. students with one learning session can pass the course and students with a high time spent in the course fail) and these must be eliminated in order to improve the classifier accuracy. Therefore, we have designed and implemented a meta-algorithm which carries out both tasks: pre-processing and modelling. Our goal is to offer this meta-algorithm to educational community so they can build more accurate classification models.

Next, we describe our meta-algorithm and compare its accuracy, TPR and FPR with respect to that obtained with 5 of the most frequently used classification algorithms [7].

2. META-ALGORITHM DESCRIPTION AND EXPERIMENTATION

Before explaining our meta-algorithm, we show why a removal outlier phase is necessary. In this experimentation we work with a course hosted in Blackboard and taught in 2009, 2010 and 2011 at University of Cantabria, entitled "Introduction to Multimedia". We generated two different datasets. Dataset1 includes the following attributes: total time spent, number of sessions, average time per session, average time per week and average number of sessions per week. Dataset2 includes the same attributes but aggregated by month and by tool (content-page, forum and mail). For instance, total time spent in January reading content-pages. Both data sets have 194 instances (one per each student) and they don't contain missing data. We show in Figure 1, the distribution of students who passed and failed with respect to the total time spent and the total number of sessions carried out. As can be observed, there are a few students who failed despite spending a lot of time in the course and often connect, and students who passed (see red squares between lines draw in Figure 1) with an average time and a number of sessions similar to those who failed. The reasons of this bad behavior is intrinsic to the way of working in the web. Students connect to the e-learning platform and, after some clicks, they open another URL out of the course, and work in parallel in both. So that the total time does not correspond with the total time of work. And viceversa, students who work hardly but in a disconnected mode since they download the materials. Thus, if we want to improve our classification models we have to minimize the effect of this problem.

Our meta-algorithm works as follows. First, it carries out a correlation study and removes those attributes which are dispensable. Next, it builds a two-class classifier with all instances and determines which are the incorrectly classified instances of both classes. Next, it calculates the prototype of each class and the euclidean distance of each instance to its prototype. Then it obtains the average distance in each class, ED, and chooses that class which has a bigger value of ED. Once chosen the class with higher ED, named K, the incorrectly classified instances of the K class are selected and the meta-algorithm carries out a clustering using Kmeans in order to separate these instances in two groups. Built the two clusters, it calculates the centroid in both clusters and names N and M respectively. Then it removes from the training sets the instances belonging to the cluster whose centroid has a larger euclidean distance to the

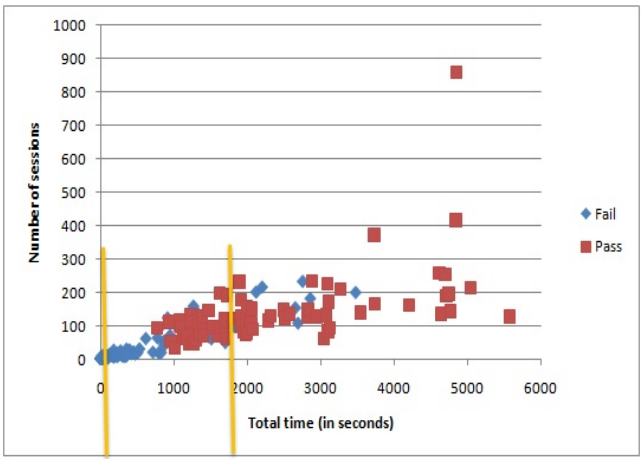

Figure 1: Class distribution according to total time and number of sessions

Table 1: Comparison of models obtained with Dataset1

Alg.	Orig.			Malg		
	Acc.	TPr	TNr	Acc.	TPr	TNr
J48	81.44	95.62	71.68	82.47+	96.30+	72.57+
NB	77.84	64.20	87.61	83.51+	83.95+	83.18-
OneR	78.87	87.65	72.57	78.35-	86.42-	72.57-
RTree	78.35	75.31	80.53	80.93+	91.36+	73.45-
JRip	81.95	92.59	74.34	85.05+	97.53+	76.11+
Avg.	79.69	83.07	77.34	82.06+	91.11+	75.57-

instances of the selected K class. That means, it eliminates the bad-classified instances, those that have a more irregular distribution with respect to the distribution of the instances set of K class. After removing the instances considered as outliers from the training sets using 10-CV, it builds the final classifier. Of course, outliers are only removed from the training sets, leaving the test sets with all initial instances. Table 1 shows the results of applying our meta-algorithm (Malg) using 5 different classification algorithms on Dataset1. As can be observed, our meta-algorithm improves in accuracy (Acc.) and TPrate (TPr) the results obtained with respect the original (Orig.) classification algorithm without removing instances, except with OneR. TNrate (TNr) is sometimes lower due to the fact that the negative class was chosen by the meta-algorithm in the preprocessing phase in this dataset. In particular, the instances with negative class eliminated correspond to the blue diamonds behind the second vertical line in Figure 1. We obtain similar results when we use Dataset2 (see Table 2). Using this dataset, our meta-algorithm improves the results of accuracy in all cases and sometimes worsens the TNrate with respect to the original classification algorithm.

3. CONCLUSIONS AND FUTURE WORK

The experimentation carried out in this work allow us to conclude that the preprocessing tasks generally improve the classification models when data sets suffer from statistical outliers. In our case study this is traduced to find students with both, a high total time spent and a high number of sessions in a virtual course who, at the end, failed.

Table 2: Comparison of models obtained with Dataset2

Alg.	Orig.			Malg		
	Acc.	TPr	TNr	Acc.	TPr	TNr
J48	87.11	96.29	80.53	87.63+	95.06-	82.30+
NB	77.84	65.43	86.73	80.93+	87.65+	76.11-
OneR	86.08	97.53	77.87	88.66+	100.00+	80.53+
RTree	82.47	76.54	86.73	87.63+	87.65+	87.61+
JRip	86.08	93.83	80.53	88.14+	98.77+	80.53=
Avg.	83.92	85.92	82.48	86.59+	93.87+	80.53-

The meta-algorithm implemented allows us to build more accurate classifiers, so instructors can predict better the student's performance of their courses and improve the teaching process. Nevertheless we are working in some improvements such as extending the meta-algorithm in order to work with a multi-valued class (more than two) or adapting it for working with multi-instance predictors, which will allow us to combine instances of courses whose organization is different and to obtain a common model for all of them. Finally, we implement a new template for our ElWM tool [2] which uses this meta-algorithm.

Out of educational data mining context, this meta-algorithm offers an opportunity to improve any classification model which presents statistical outliers in its training datasets.

4. REFERENCES

[1] M. Cocea and S. Weibelzahl. Cross-system validation of engagement prediction from log files. In *EC-TEL*, pages 14–25, 2007.

[2] D. García-Saiz and M. E. Zorrilla. E-learning web miner: A data mining application to help instructors involved in virtual courses. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. C. Stamper, editors, *EDM*, pages 323–324. www.educationaldatamining.org, 2011.

[3] W. Hämäläinen and M. Vinni. Comparison of machine learning methods for intelligent tutoring systems. In M. Ikeda, K. Ashley, and T.-W. Chan, editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 525–534. Springer, 2006.

[4] A. A. Juan, T. Daradoumis, J. Faulin, and F. Xhafa. A data analysis model based on control charts to monitor online learning processes. *Int. J. Bus. Intell. Data Min.*, 4:159–174, July 2009.

[5] X. Li, Q. Luo, and J. Yuan. Personalized recommendation service system in e-learning using web intelligence. In *Proc. of the 7th international conference on Computational Science, Part III*, pages 531–538, 2007.

[6] C. Romero and S. Ventura. 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, 2010.

[7] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14:1–37, December 2007.