JMLR: Workshop and Conference Proceedings 17 (2011) 26–32 2nd Workshop on Applications of Pattern Analysis

Comparing classification methods for predicting distance students' performance

Diego García-Saiz DIEGO.GARCIAS@ALUMNOS.UNICAN.ES and **Marta Zorrilla**
MARTA.ZORRILLA@UNICAN.ES
Mathematics, Statistics and Computation Department, University of Cantabria. Avda. de los Castros s/n, 39005 Santander, Spain

Editor: Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Abstract

Virtual teaching is constantly growing and, with it, the necessity of instructors to predict the performance of their students. In response to this necessity, different machine learning techniques can be used. Although there are so many benchmarks comparing their performance and accuracy, there are still very few experiments carried out on educational datasets which have very special features which make them different from other datasets. Therefore, in this work we compare the performance and interpretation level of the output of the different classification techniques applied on educational datasets and propose a meta-algorithm to preprocess the datasets and improve the accuracy of the model.

Keywords: Educational data mining, Classification techniques, Data mining tools for naive users

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 managed, can help both instructors and students. Instructors can discover information to validate/evaluate the teaching-learning process (Romero and Ventura, 2010); and, students can receive suitable feedback about their dedication in the course (Juan et al., 2009) and recommendations in order to achieve the learning objectives (Li et al., 2007).

The problems arise when instructors with little or no knowledge about data analysis techniques want to predict the students' performance or to prevent student dropout among others. As is known, the process of knowledge discovery from data (KDD) involves the repeated application of the several steps (creating a target dataset, choosing the data mining technique, tuning and executing the algorithm on the dataset and interpreting mined patterns) and regretfully, in each step of the process, there is such a number of decisions to be taken with little or no formal guidance, that, often, the outcome of the process just results in the necessity to start over: either with more (or less, but cleaner) data, or using different algorithms, or setting different values for some parameters.

As Romero and Ventura (2010) cite "Data Mining tools are normally designed more for power and flexibility than for simplicity" and as a consequence of this fact instructors have difficulties discovering how their students behave and progress in the course. Even more, they also mention the necessity of these tools being integrated in e-learning systems. A challenging task in this direction is being carried out by the authors of this paper who are developing a tool, called ElWM (Elearning Web Miner) (Zorrilla and García, 2011) which tries to meet this need.

ElWM currently offers three templates which allow instructors to discover the student profile, the pattern of resources which are frequently used together in the course and the session profile, by simply sending a data file according to the templates provided by the system or by choosing the course if the system is connected to the e-learning environment and requesting the results. The tool itself is able to perform all the mining process (towards data mining without parameters). These templates include the definition of the dataset, the preprocessing tasks and the algorithm to use and the configuration of its parameters, these being selected from an previous analysis of the dataset.

In this work, we carry out a study of different classification techniques applied to educational datasets with the aim of finding the most suitable ones to answer prediction questions. Finally, we propose a meta-algorithm to get more accurate models.

The paper is organized as follows. In Section 2 we describe the main characteristics of educational datasets. Section 3 explains our case study, describes the datasets used and discusses the results obtained. Section 4 relates other research works carried out in this field and finally, Section 5 draws the most important conclusions of our study.

2. Educational datasets

One of the main challenges in data mining is scaling the algorithms to work with large datasets. In the case of educational datasets, the situation is very distinct mainly due to the dataset size is very small (ranges between 50 and 100 transactions, one for each student enrolled in the course) and the difficulty for carrying out the experimentation phase as a consequence of the fact that the data is very dynamic and can vary a lot among samples (different course design, different methods of assessment,different resources used, etc.) (Merceron and Yacef, 2008). At best, the dataset size can be of the order of 200 or 300 students, if the course has remained unchanged during two or more editions.

Another characteristic is that this data is generally numeric (time spent in session, mark in tests, and so on), although it can also have some categorical variables (boolean-answer test, answers to surveys), therefore our study must tackle both cases. And lastly, it is usually very clean since it comes from databases, so that few or no pre-processing tasks are required (Hämäläinen and Vinni, 2006). Regarding missing data, it is generally presented in datasets as a consequence of the fact that not all students answer surveys or complete required tasks and there are a big number of dropouts in the first weeks of virtual courses (Dekker et al., 2009).

3. Case study

In this section, we describe the datasets used in our experimentation, indicate the selected classification algorithms, describe our meta-algorithm and discuss the results obtained.

3.1. Datasets

For the case studies, we used the data from a course offered in the last three academic years (2007-2010) at the University of Cantabria. This is entitled ”Introduction to multimedia methods” and is open to all degrees. The average number of students enrolled per year is about 70, but only 25% of these deliver all the tasks and a lower number pass the course.

We worked with three datasets, two of them (Mu0910 and Mu0710 with 65 and 164 instances respectively) gather the global activity carried out by students of each course during the term and the final mark (pass/fail). And the third (Mu0910S with 65 instances) has 5 attributes more related to the student's learning style (Active/Reflectivee; Sensitive/Intuitive; Visual/Verbal; Sequential/Global) and his degree obtained by means of a survey. The global activity is measured by means of the following attributes: Total time spent, Number of sessions carried out, Average number of sessions per week, Average time spent per week and Average time per session. All datasets have the mark attribute as the variable to predict.

3.2. Classification techniques

As our intention is to choose the best algorithms for educational datasets which can be integrated in our ElWM tool, we have to search among those that can support categorical and numeric data, handle incomplete data, offer a natural interpretation to instructors and be accurate working with small samples. Therefore, we analyse four of the most common machine learning techniques, namely Rule-based algorithms, Decision Trees, Bayesian classifiers and Instance-based learner classifiers. We rule out the use of Artificial Neural Networks and Support Vector Machine techniques because of their lack of a comprehensive visual representation. The chosen algorithms were OneR, J48, Naïve Bayes, BayesNet TAN and NNge. Weka (Hall et al., 2009) was used to perform the analysis.

3.3. Classification Algorithms Comparison

We tested the five aforementioned algorithms using different parameter settings and different numbers of folds for cross validation, in order to discover whether they have a great effect on the result. Finally, we set the algorithms with default parameters and used 10-fold cross validation (Kohavi, 1995) for estimating generalization performance and the students' t-test for comparing the models generated. The statistical significance of differences in performance among OneR and the rest of classifiers was tested with the two-sided paired t-tester in Weka, using a significance level of 10%. Next, we show the classification accuracy and rates obtained with the five algorithms for the Mu0910 dataset in Table 1.

As can be observed, Bayes algorithms perform better in accuracy and TP rate and FP rate, and is comparable to J48 algorithm although it is worse at predicting fails (89% pass rate and 69% fail rate) than Naive Bayes (68% pass rate and 80,5% fail rate) which is the best in this aspect, whereas BayesNet TAN gets the most balanced result (84% pass rate and 70% fail rate). It must be highlighted that OneR suffered from overfitting in this dataset, so that it should be discarded as a suitable classifier for very small datasets.

Next, we discuss the results obtained on Mu0710 dataset. On one hand, the Table 2 shows that BayesNet TAN is the most efficient algorithm for this dataset with 196 instances, followed (very closed) by J48. On the other hand, Naïve Bayes is the one with the worst

Table 1: Accuracy and rates of Mu0910 dataset

Algorithm	TPRate	FPRate	TNRate	FNRate	Accuracy
OneR B 3	0.66	0.47	0.53	0.34	65.79
J48 C 0.25	0.74 v	0.17 *	0.83 v	0.26 *	74.21 v
Naïve Bayes	0.77 v	0.25	0.75	0.23 *	77.29 v
BayesNet TAN	0.76 v	0.18 *	0.82 v	0.24 *	76.36 v
NNge	0.70	0.41	0.59	0.30	70.10

v,* statistically significant improvement or degradation with respect to OneR

prediction according to t-test. It can also be observed that NNge improves its performance in this dataset although the great number of rules which it offers as output (more than 40) makes it less interpretative for instructors than the rest of the models.

Table 2: Accuracy and rates of Mu0710 dataset

Algorithm	TPRate	FPRate	TNRate	FNRate	Accuracy
OneR B 6	0.78	0.20	0.80	0.22	77.86
J48 C 0.25	0.79	0.17	0.83	0.21	79.36
Naïve Bayes	0.76	0.27 v	0.73 *	0.24	76.40
BayesNet TAN	0.81 v	0.15 *	0.85 v	0.19 *	81.26 v
NNge	0.78	0.22	0.78	0.22	78.04

v,* statistically significant improvement or degradation with respect to OneR

Finally, the classification prediction and rates for the Mu0910S dataset are shown in Table 3. Here we can see that Naïve Bayes and J48 are the two most accurate algorithms. In this case, the increment of the number of attributes has resulted in Naïve Bayes model gaining in accuracy despite this dataset presenting missing data. Surprisingly, BayesNet TAN has reduced its accuracy remarkably and NNge performs better than BayesNet TAN.

Table 3: Accuracy and rates of Mu0910S dataset

Algorithm	TPRate	FPRate	TNRate	FNRate	Accuracy
OneR B 3	0.65	0.48	0.52	0.35	65.29
J48 C 0.25	0.76 v	0.23 *	0.77 v	0.24 *	75.83 v
Naïve Bayes	0.81 v	0.20 *	0.80 v	0.19 *	80.90 v
BayesNet TAN	0.73	0.29	0.71	0.27	73.36
NNge	0.75	0.36	0.64	0.25	74.98

v,* statistically significant improvement or degradation with respect to OneR

Thus, we can conclude that Bayes Networks are suitable for small datasets (less than 100 instances), performing better Naïve Bayes when the sample is smaller. As consequence of the fact that BayesNet TAN model is more difficult to interpret for a non-expert user and J48 is similar in accuracy to it, we decide to use Naïve Bayes and J48 according to the size of the dataset and the kind of attributes.

3.4. Meta-algorithm

With the aim of improving the accuracy of these models, we proceeded to make a classified data analysis and 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). So that, we built a meta-algorithm to preprocess the dataset and eliminate these ”outliers”. This consists of building an initial classifier using 10-fold cross-validation and eliminating, randomly, a certain percentage of the instances which belong to the worse classified class, and next, building a second classifier with the filtered dataset using also 10-fold cross-validation. This percentage was initially set at 20%, chosen by experimentation.

All models built up showed that the classifiers' accuracy is higher. For example, applying Naïve Bayes on Mu0710 dataset in both stages, the pass and fail rates increased from 63.75% to 75.00% and from 85.09% to 86.84% respectively; and applying J48 in both stages, the pass and fail rates increased from 91.25% to 95.00% and from 71.93% to 86.81% respectively.

Better results were got when the instances ruled out were chosen taking into account the value of the most significant attribute for the algorithm. This means, to eliminate the 20% of instances with the highest values for this attribute when the worse classified class is ”pass” and the lowest ones when the worse classified class is ”fail”. In this case we used the classifierSubSetEval algorithm from Weka to select the attribute. It established that the total time spent was the most important attribute to classify the Mu0710 dataset in both techniques, J48 and Naïve Bayes. The models obtained with this preprocessing applying Naïve Bayes in both stages, got an improvement in the fail rate from 85.09% to 87.72% and got worse in the pass rate, from 63.75% to 71.88% in comparison with the previous results. But, applying J48 in both stages, both rates increased, the pass rate from 91.25% to 100% and the fail rate from 71.93% to 86.81%. It must be highlighted that, with J48,we eliminated the bad classified ”fail” instances, whereas, with Naïve Bayes, we ruled out the bad classified ”pass” instances. When we eliminated the bad classified ”fail” instances with Naïve Bayes (although the ”fail” instances are better classified than the ”pass” ones), we achieved an improvement in pass and fail rates from 63.75% to 90.72% and from 85.09% to 91.25% respectively. That is, we obtained better results than in the previous tests carried out with Naïve Bayes. Another alternative tested consisted on applying this preprocessing in each cross-validation fold of the first classifier. But, in this case, the instances ruled out in each fold were of different class and the results obtained were worse than the previous ones. The pass and fail rates with Naïve Bayes were 66.13% and 86.27% respectively, and 89.47% and 78.57% with J48.

So, we can conclude that eliminating ”outliers” of the class which present more irregularities following the second method mentioned is the best idea. Although more experimentation must be done, this meta-algorithm seems to be good to define our template for predicting students' performance. This improvement is remarkable when the dataset is bigger, but it is still notable with smaller datasets, like Mu0910, where our meta-algorithm achieves an improvement up to 6% in pass and fail rate.

4. Related Work

The use of data mining techniques in educational data is relatively new. As far as we know, there are only a few studies that have tackled this problem. Among these, we can mention the following: Kotsiantis et al. (2003) compared six classification algorithms to predict drop outs. The number of instances in the dataset was 350 and this contained numeric and categorical data. Naïve Bayes and neural networks were the two methods which performed best. Another similar study was done by Hämäläinen and Vinni (2006). In it, the authors compared five classification methods for predicting the course outcomes and used datasets which were very small (125 rows and 88 rows respectively). For numerical data, they used multiple linear regression and support vector machine classifiers, and for categorial data, three variations of Naïve Bayes classifier. They recommend Naïve Bayes classifiers, which are robust, can handle mixed variables and produce informative results (class probabilities).

Cocea and Weibelzahl (2007), tested eight methods to predict the engagement of students in virtual courses using data from log files. As in the aforementioned cases, all techniques offer good levels of prediction, with IBk algorithm being the most accurate. The size of the datasets were 341 and 450 respectively. All the attributes except the class variable were numeric. Finally, in Dekker et al. (2009), the authors present a case study to predict student drop-out demonstrating the effectiveness of several classification techniques and the cost-sensitive learning approach on several datasets over 500 instances with numerical and nominal attributes. They show that rather simple classifiers (J48, CART) give a useful result compared to other algorithms such as Bayes Net or JRip.

5. Conclusions

Although there are so many benchmarks comparing the performance and accuracy of different classification algorithms, there are still very few experiments carried out on Educational datasets. In this work, we compare the performance and the interpretation level of the output of different classification techniques applied on educational datasets in order to determine which one is more suitable for wrapping in our ElWM tool.

Our experimentation shows that there is not one algorithm that obtains a significantly better classification accuracy. In fact, the accuracy depends on the sample size and the type of attributes. When the sample size is very small (less than 100 instances) and contains numeric attributes, Naïve Bayes performs adequately, on the other hand, when the dataset is bigger, BayesNet TAN is a better alternative. J48 is suitable for datasets with more instances and/or with the presence of nominal attributes with missing data, although in this last context Naïve Bayes is the best but less interpretable. Due to the special characteristics of the datasets used, the best results are obtained with the meta-algorithm proposed using both Naïve Bayes and J48 to preprocess and to predict, being better if the preprocessed task is carried out according the most significant attribute for the algorithm used to preprocess.

Our near future work is to extend this experimentation with other datasets to validate these conclusions and next, to study how the meta-algorithm can set itself according to the dataset and, lastly, to add a template for predicting the students' success in our tool.

Acknowledgments

This work has been financed by Spanish Ministry of Science and Technology under project 'TIN2008 - 05924'

References

Mihaela Cocea and Stephan Weibelzahl. Cross-system validation of engagement prediction from log files. In *EC-TEL*, pages 14–25, 2007.

G.W. Dekker, M. Pechenizkiy, and J.M. Vleeshouwers. Predicting students drop out: a case study. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 41–50, 2009.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11: 10–18, November 2009. ISSN 1931-0145.

Wilhelmiina Hämäläinen and Mikko Vinni. Comparison of machine learning methods for intelligent tutoring systems. In Mitsuru Ikeda, Kevin Ashley, and Tak-Wai Chan, editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 525–534. Springer Berlin / Heidelberg, 2006.

Angel A. Juan, Thanasis Daradoumis, Javier Faulin, and Fatos 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. ISSN 1743-8195.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145, 1995.

Sotiris B. Kotsiantis, Christos Pierrakeas, and Panayiotis E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *KES*, pages 267–274, 2003.

Xinye Li, Qi Luo, and Jinsha Yuan. Personalized recommendation service system in e-learning using web intelligence. In *Proceedings of the 7th international conference on Computational Science, Part III: ICCS 2007*, pages 531–538, 2007. ISBN 978-3-540-72587-9.

Agathe Merceron and Kalina Yacef. Interestingness measures for associations rules in educational data. In *Proceedings of the 1st International Conference on Educational Data Mining*, pages 57–66, 2008.

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. ISSN 1094-6977. doi: 10.1109/TSMCC.2010.2053532.

Marta Zorrilla and Diego García. *Data Mining Service to Assist Instructors involved in Virtual Education.* IGI Global Publisher, 1 edition, 2011.