Learning Analytics for the Prediction of the Educational Objectives Achievement

Manuel Fernández-Delgado, Manuel Mucientes, Borja Vázquez-Barreiros and Manuel Lama
Center for Research in Information Technologies (CiTIUS)
University of Santiago de Compostela (Spain)
Emails: {manuel.fernadez.delgado, manuel.mucientes, manuel.lama}@usc.es

***Abstract*—Prediction of students' performance is one of the most explored issues in educational data mining. To predict if students will achieve the outcomes of the subject based on the previous results enables teachers to adapt the learning design of the subject to the teaching-learning process. However, this adaptation is even more relevant if we could predict the fulfillment of the educational objectives of a subject, since teachers should focus the adaptation on the learning resources and activities related to those educational objectives. In this paper, we present an experiment where a support vector machine is applied as a classifier that predicts if the different educational objectives of a subject are achieved or not. The inputs of the problem are the marks obtained by the students in the questionnaires related to the learning activities that students must undertake during the course. The results are very good, since the classifiers predict the achievement of the educational objectives with precision over 80%.**

I. Introduction

One of the most challenging issues in the design of a subject consist in developing the learning activities and contents that would be appropriate to achieve the set of educational objectives related to that subject [1]. Adaptive learning strategies [2] try to minimize the impact of non appropriate learning designs by assigning complementary learning activities [3], [4] and/or adaptive contents [5] to each student based on different criteria such as the student's performance. This adaptation is typically implemented through adaptive rules that select the learning resources depending on the values of certain student properties such as the time needed to complete previous learning activities, the marks obtained in exercises or questionnaires, and so forth. However, this adaption would be even more effective if we could *predict* if students will achieve the educational objectives of the subject [6]. This kind of prediction would allow teachers to introduce learning activities *before* the students undertake the planned learning activities, improving the learning design of the course.

In this context, we consider that learning activities undertaken by learners are related to one or several educational objectives and that an educational objective is achieved through one or several learning activities. Also, if a particular learning activity is successfully completed by a student, the educational objectives attached to that activity might be achieved or not, depending on whether these educational objectives are associated with other learning activities. Thus, the information about the level in which the educational objectives are achieved through the enforcement of the learning activities would allow teachers to introduce new learning activities and/or contents with the aim of improving the learning-teaching process.

Prediction of the students' performance is one of the first challenges faced in educational data mining [7]. This performance can be understood in very different ways such as the number of errors made by students in exams, final grades, students' marks in a subject, and so forth. Examples of approaches to predict students' marks —which is the purpose of this paper— includes neural networks [8], [9], Bayesian networks [10], [11], rule-based techniques [12], [13], linear regression [14], and so forth. For the purpose of this paper, the main drawback of these approaches it that they do not focus on predicting the fulfillment of the educational objectives of a course, as they assume that each educational objective is related to *an only* learning activity, and vice versa.

In this paper, we present a set of classifiers that are able to predict the degree of fulfillment of each educational objective in the course —one classifier for each objective—. The classifiers were learned with a Support Vector Machine (SVM) [15], using as inputs the marks of the assignments for each learning activity, and providing as output information about the achievement of the educational objectives by each learner. In these classifiers the degree of fulfillment of the educational objectives is binary, i.e., the objective was accomplished or not. To validate the performance of this technique we have conducted an experiment with 56 students. The results of the classification are good, with a high precision for all the educational objectives. The worst results are for the most unbalanced problems, where most of the students achieve (or not) the objective. Also, the educational objectives that are more difficult to predict are those associated with questionnaires which require more creativity.

This paper is structured as follows: in Section II the educational scenario in which the prediction technique has been applied is described; in Section III the results of applying the prediction technique in that educational scenario are discussed; and, finally, in Section IV we summarize the contributions of the paper.

II. Description of the experiment

The experiment described in this section was implemented for Automata Theory and Formal Languages (ATFL), a core subject of the Computer Science Engineering (CSE) Degree of the University of Santiago de Compostela (USC). We obtained data from 56 students that attended the subject during the first semester of the 2011-2012 academic course. The students have to attend 25 hours of theoretical lectures structured in seven parts (Fig. 1) and 25 hours of practical lectures with 12 assignments and their corresponding educational objectives.

978-1-4799-3922-0/14/$31.00 ©2014 IEEE

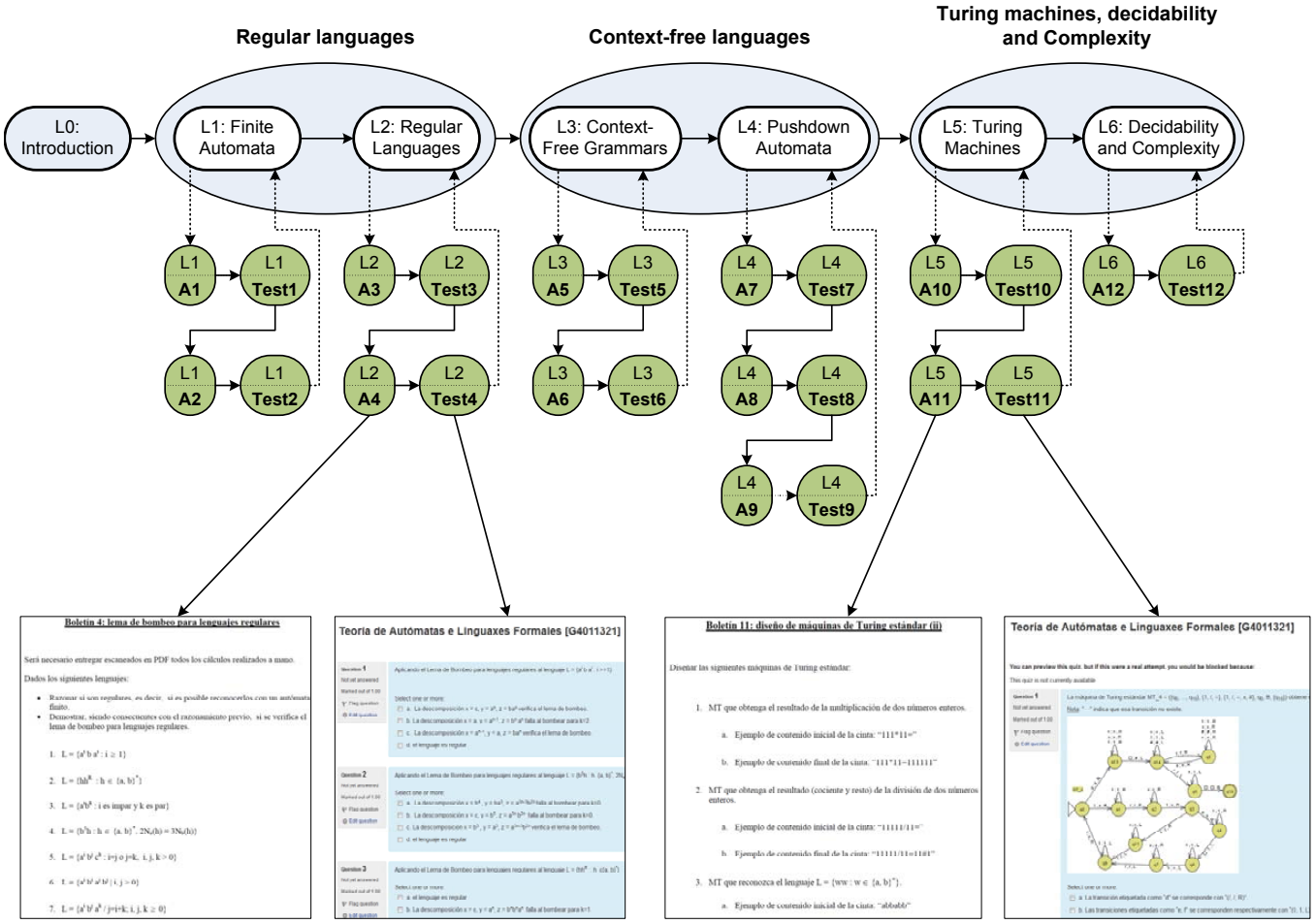

Figure 1. Blocks diagram of the subject ATFL.

The relationship between the lectures (L0-L6), assignments (A1-A12) and educational objectives (O1-O8) is the following:

- Introduction (L0). This part has no assignments or educational objectives.
- Finite Automata (L1).
  - Design of finite state automata (A1); design of finite automata (O1).
  - Real applications of finite state automata (A2); design of finite automata (O1).
- Regular Languages (L2).
  - Minimization of finite state automata, regular expressions (A3); connection between finite automata and regular expressions (O2).
  - Pumping lemma for regular languages (A4, O3).
- Context-Free Grammars (L3).
  - Design of context-free grammars (A5, O5).
  - Design of context-free grammars and Chomsky normal form (A6); design of context-free grammars (O5).
- Pushdown Automata (L4).
  - Design of pushdown automata that accept by final state (A7); design of pushdown automata (O4).
  - Design of pushdown automata that accept by empty stack (A8); design of pushdown automata (O4).
  - Pumping lemma for context-free languages (A9, O6).
- Turing Machines (L5).
  - Design of standard Turing machines (A10, O7).
  - Design of complex standard Turing machines (A11); design of standard Turing machines (O7).
- Decidability and Complexity (L6).
  - Design of context-sensitive and unrestricted grammars (A12, O8).

The assessment of the subject is the sum of the theoretical and practical parts. The practical part is evaluated with the completion of 12 questionnaires (T1-T12), each of them associated with the respective assignment (A1-A12). These questionnaires are multiple choice tests. On the other hand, the theoretical part is evaluated with a final questionnaire

with 8 topics (Q1-Q8) associated with each of the educational objectives (O1-O8). Each topic has several questions that must be answered like a true-false test.

Our starting hypothesis was that the marks of the assignments (T1-T12) are able to predict the fulfillment of the educational objectives. Therefore, the objective of the experiment was to build a model of the educational objectives using machine learning techniques. Given a set of marks for the assignments, the model automatically provides the degree of fulfillment of each educational objective. With this model the teacher can identify the lacks of the students on the educational objectives before the final exam. Therefore, it gives the chance to the teacher to fill the gaps of the students through reinforcement activities.

As the subject has eight educational objectives, we have eight machine learning problems, and each one will obtain a model for the corresponding educational objective. Any supervised machine learning algorithm needs two datasets: i) the training dataset (D_{tra}), which contains the examples used to learn the model; ii) the test dataset (D_{tst}), which has the examples used to validate the learned model. For both datasets an example is defined as: $\left(T_1^i \ \ldots \ T_{12}^i \ Q_j^i\right)$, where $T_k^i \in [0,\ 10]$ is the mark of the i-th student in the k-th questionnaire, and Q_j^i is the mark of the i-th student in the j-th topic of the final questionnaire. As was stated before, the degree of fulfillment of each educational objective is represented by the mark obtained in the corresponding topic of the final questionnaire.

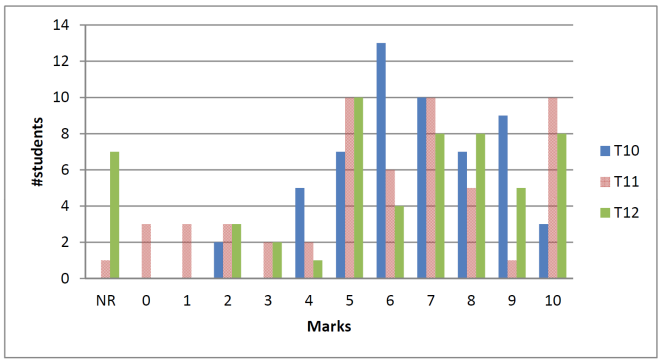

Figure 2. Histograms of the marks for the assigments T10-T12. The value NR indicates that the student did not complete the assigment.

III. Results

Fig. 2 shows the histogram of the marks for 3 of the 12 assignments. On the other hand, Fig. 3 presents the histogram of the degree of fulfillment of each educational objective, which is represented by the marks obtained in the final questionnaire. The degree of fulfillment of the educational objectives is binary, i.e., the objective was accomplished ($Q_j^i = 1$) or not ($Q_j^i = 0$).

Given that the classes are highly unbalanced, we followed a resampling strategy called SMOTE (Synthetic Minority Oversampling TEchnique) [16], generating an artificial data set from the original patterns in which all the classes have the same population. Specifically, if n_i is the number of patterns of class i and $N = \max_i\{n_i\}$, we generated N artificial patterns $\mathbf{x}_{i1}^a, \ldots, \mathbf{x}_{iN}^a$ for each class i, being $\mathbf{x}_{ij}^a$ defined by:

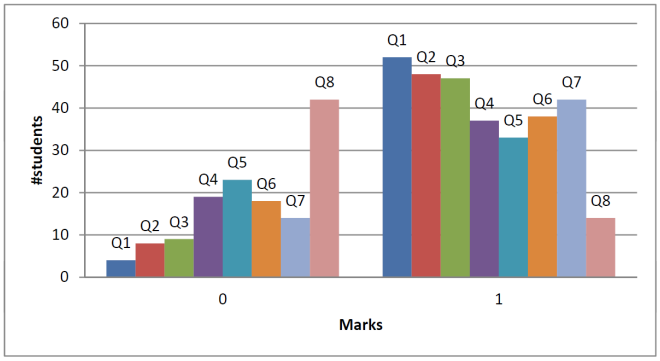

Figure 3. Histogram of the marks for the educational objectives (Q1-Q8).

$$\mathbf{x}_{ij}^a = \mathbf{x}_i + \varepsilon_j(\mathbf{x}_i^* - \mathbf{x}_i) \tag{1}$$

where $\mathbf{x}_i$ is an original pattern of class i (selected randomly), $\mathbf{x}_i^*$ is the nearest neighbor of $\mathbf{x}_i$ (selected among the original patterns of class i), and ε_j is a random number in $(0, 1)$. For each of the educational objectives, the resulting artificial dataset is the training dataset (D_{tra}) of the corresponding machine learning problem.

The classifiers were learned with a Support Vector Machine (SVM). In particular, we used the R[1] (v. 2.15.3) implementation of the SVM provided by the function `ksvm` in the package `kernlab`[2] (v. 0.9-18) [15]. The SVM uses a Gaussian kernel, tuning the kernel spread σ with values $\{2^i\}_{i=-16}^{8}$ and the regularization parameter C with values $\{2^i\}_{i=-5}^{14}$. The selected values of C and σ were those which achieved the best average accuracy over the four trials of a 4-fold cross validation on the artificial dataset. Finally, the SVM was trained on the whole artificial dataset (D_{tra}) with the best values of C and σ, and tested on the original dataset (D_{tst}).

Table I. Precision (P_i) and recall (R_i) for the eight educational objectives.

	Class 0		Class 1	
Educational objectives	P_0	R_0	P_1	R_1
Q1	100.0	66.7	96.2	100.0
Q2	100.0	72.7	93.8	100.0
Q3	100.0	100.0	100.0	100.0
Q4	94.7	75.0	83.8	96.9
Q5	87.0	76.9	81.8	90.0
Q6	100.0	81.8	89.5	100.0
Q7	92.9	86.7	95.2	97.6
Q8	92.9	100.0	100.0	82.4

Table I shows the results of the experiments for each of the educational objectives. The performance of the classifiers was measured with two well known metrics: precision (P) and recall (R). Precision is defined as:

$$P = \frac{TP}{TP + FP} \tag{2}$$

[1]http://www.r-project.org

[2]http://cran.r-project.org/web/packages/kernlab

where TP is the number of true positives —the number examples of the class that were correctly classified—, and FP is the number of false positives —the number of examples of other classes that were incorrectly classified in that class. The higher the precision, the higher the probability that an example classified in the class was correctly identified. On the other hand, recall is defined as:

$$R = \frac{TP}{TP + FN} \tag{3}$$

where FN is the number of false negatives —the number of examples of the class that were incorrectly classified. Thus, recall measures the ability of the classifier to detect examples from that class.

The results of the classification are good, with a high precision for all the educational objectives. The worst result is for objective O1 (questionnaire Q1), as it is the most unbalanced problem and, therefore, although the precision for class 1 is really high, the few misclassified examples have a great impact in the recall of class 0 examples ($R_0 = 66.7$). Also, the educational objectives related with the design of automata (specially O1 and O4) are more difficult to predict, as the associated questionnaires cannot be solved following a number of predefined steps. For Q2 the problem is again very unbalanced but, also, the corresponding educational objective —connection between finite automata and regular expressions (O2)— is particularly difficult to predict because the only test directly related with it is T3 —minimization of finite state automata, regular expressions.

IV. Conclusions

We have presented an approach for predicting the degree of fulfillment of the educational objectives of a subject through Support Vector Machines. Our proposal uses as inputs the marks of the assignments of each learning activity and provides as outputs whether the educational objective was fulfilled or not. We have conducted an experiment with 56 students in the subject of Automata Theory and Formal Languages. Results show a great predictive ability of the classifiers, with very high precisions (always over 80%) and good recalls, although some of the problems are highly unbalanced and/or very difficult to predict.

As future work, we plan to extend our proposal to predict not only the fulfillment of the educational objectives, but also the degree in which the educational objective was met. Thus, the classifiers will indicate if the objective was fully accomplished, if it was not met, or a partial fulfillment of the educational objective. This is a multi-class problem, and will be tackled with a one vs. one approach: each classifier is able to distinguish between a pair of classes.

Acknowledgment

This work was supported by the Spanish Ministry of Economy and Competitiveness under the project TIN2011-22935 and by the European Regional Development Fund (ERDF/FEDER) under the project CN2012/151 of the Galician Ministry of Education.

References

[1] S. Toohey, *Designing Courses for Higher Education.* Philadelphia, USA: Open University Press, 1999.

[2] N. Henze and W. Nejdl, "A logical characterization of adaptive educational hypermedia," *The New Review of Hypermedia and Multimedia*, vol. 10, no. 1, pp. 77–113, jun 2004.

[3] H. Nijhavan and P. Brusilovsky, "A framework for adaptive e-learning based on distributed re-usable learning activities," in *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2002)*, M. Driscoll and T. Reeves, Eds. Montreal, Canada: AACE, 2002, pp. 154–161.

[4] D. Burgos, C. Tattersall, and R. Koper, "How to represent adaptation in e-learning with ims learning design," *Interactive Learning Environments*, vol. 15, no. 2, pp. 161–170, 2007.

[5] P. Brusilovsky, "KnowledgeTree: A distributed architecture for adaptive e-learning," in *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, Eds. ACM Press, 2004, pp. 104–113.

[6] M. Jovanovica, M. Vukicevica, M. Milovanovica, and M. Minovica, "Using data mining on student behavior and cognitive style data for improving e-learning systems: A case study," *International Journal of Computational Intelligence Systems*, vol. 5, no. 3, pp. 597–610, 2012.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, 2010.

[8] M. R. Beikzadeh, S. Phon-Amnuaisuk, and N. Delavari, "Data mining application in higher learning institutions," *Informatics in Education*, vol. 7, no. 1, pp. 31–54, 2008.

[9] L. Lykourentzou, L. Giannoukos, and G. Mpardis, "Early and dynamic student achievement prediction in e-learning courses using neural networks," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 372–380, 2009.

[10] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 529–535, 2010.

[11] R. Stevens, A. Soller, A. Giordani, L. Gerosa, M. Cooper, and C. Cox, "Developing a framework for integrating prior problem solving and knowledge sharing histories of a group to predict future group performance," in *Proceedings of the 1st International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2005)*, San Jose, CA, USA, 2005, pp. 1–9.

[12] Á. Nebot, F. Mugica, and F. Castro, "Fuzzy predictive models to help teachers in e-learning courses," in *International Joint Conference on Neural Networks (IJCNN 2010)*, Barcelona, Spain, 2010, pp. 1814–1820.

[13] T. A. Etchells, Á. Nebot, A. Vellido, P. J. G. Lisboa, and F. Mugica, "Learning what is important: feature selection and rule extraction in a virtual course," in *Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, M. Verleysen, Ed., Bruges, Belgium, 2006, pp. 401–406.

[14] N. Myller, J. Suhonen, and E. Sutinen, "Using data mining for improving web-based course design," in *Proceedings of the International Conference on Computers in Education (ICCE 2002)*, Kinshuk, R. Lewis, K. Akamori, R. Kemp, T. Okamoto, L. Henderson, and C. H. Lee, Eds., Auckland, New Zealand, 2002, pp. 959–963.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.