

Behavior Pattern Mining during the Evaluation Phase in an e-Learning Course

Bernardete Ribeiro and Alberto Cardoso

CISUC, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
bribeiro@dei.uc.pt and alberto@dei.uc.pt

Abstract – There is a broad range of products available for e-learning which can be used in course curriculum at University level. While e-learning in education is well established, there are a few attempts to extract information in its evaluation phase. We look at a specific part of an e-learning designed course favoring students' evaluation phase for extraction of behavior pattern mining. Our approach uses neural networks (NN) and Support Vector Machines (SVM) to build prediction models able to track student's behavior. The data sets were obtained from student's logs in a Moodle designed Course of Discrete Structures of the Informatics Engineering Bachelor at the University of Coimbra. The results show the model is able to successfully predict students' final outcome while bringing useful feedback during course learning.

Index Terms – e-learning, data mining techniques, course evaluation.

INTRODUCTION

The overwhelming growth of Internet turns possible an enormous potential for the use of online education and e-learning. This rapid proliferation of information and their easy accessibility through World Wide Web is attracting more researchers towards web-based data source.

In recent times, numerous tools are being used by the scientific community to retrieve the required data from the available databases needed for web-designed courses. Increasingly more institutions provide their students with web-based learning management systems (LMS). These tools are important for facilitating information extraction, fact finding, relationship search, and concept discovery.

A broad range of commercial LMS around such as WebCT, Virtual-U and TopClass among others [1] which have been proved efficient. However, freely distributed learning management systems such as Moodle [2], Atutor, ILIAS [3] and educational adaptive hypermedia courses as ELM-ART and AHA are also becoming prominent [4].

These systems generate a vast amount of information daily that is very useful for analyzing student's pattern behavior if proper data mining tools are used. There are several views on how to sort out education mining tools. Following [4] we refer some most used types (i) personalization of learning systems [5] to help educators analyzing some aspect of the learning process, (ii) recommendation systems [6], to classify students and contents to recommend optimum resources, (iii) detection of

irregularities [7], to discover irregular browsing patterns. Moreover, these systems can be further categorized according to their direct goal function i.e. towards students or instructors.

We look at a specific part of an e-learning designed course favoring students' evaluation phase for extraction of behavior pattern mining. Data mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. It is a blend of statistics, artificial intelligence, and database research. With appropriate data preparation and strong algorithms, data mining can produce relevant analysis results and provide novel insights of high scientific value.

The paper presents a prediction model of the students learning performance based on the results obtained during the evaluation phase by using Moodle in a Course of Discrete Structures of the Informatics Engineering Bachelor at the University of Coimbra.

This paper presents five sections. A brief overview of data mining techniques is given in the next section. In the third section the experimental setup is described. In section four the prediction learning model for the evaluation phase is presented. The performance measures and analysis of results in terms of prediction accuracy are also given. The main findings are drawn in the last section.

DATA MINING TECHNIQUES

In the following the main techniques used for pattern behavior modeling are underlined. Supervised (neural Networks and Support Vector Machines) and unsupervised techniques (Clustering) were used to build the prediction models of students behavior. The difference between them is on, using or not, a target vector aiming at model tuning.

1. Neural Networks

Neural networks (NN) are inspired in biological models of brain functioning. They are capable of learning by examples and generalizing the acquired knowledge. Due to these abilities the neural networks are widely used to find out non-linear relations which otherwise could not be unveiled due to analytical constraints. The learned knowledge is hidden in their structure thus it is not possibly to be easily extracted and interpreted.

The structure of the multilayered perceptron (MLP), i.e. the number of hidden layers and the number of neurons, determines its capacity, while the knowledge about the

relations between input and output data is stored in the weights of connections between neurons. The values of weights are updated in the supervised training process with a set of known and representative values of input – output data samples.

The neural network architecture for the problem under consideration in this paper is illustrated in Figure 1.

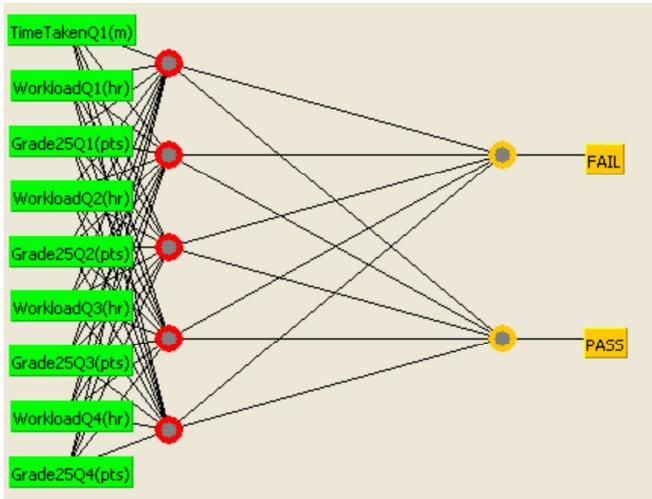


FIGURE 1

MULTILAYER PERCEPRON PREDICION MODEL OF STUDENTS OUTCOME BASED ON GRADING (PTS), WORKLOAD (HR) AND TIMETAKEN (MIN).

The training data is used to set up the values of the weights and thus build the model, whereas the testing data is used for testing of dependencies. At the beginning of the training process, the weights are set randomly. With respect to the squared difference between the target value and the calculated output value on the output neuron, weights of all connections are updated by a minimization algorithm.

The empirical risk minimization (ERM) can be performed by several algorithms. One of the most succeeded is Levenberg-Marquardt back-propagation minimization algorithm [8], which involves performing computations backwards through the network. The presentation of input – output samples to the model is repeated until the weights of the network stabilize and the average squared difference between the calculated output and target values converges to a minimum value. Input and output values are usually normalized for non-linear transfer functions to operate in active region.

II. Support Vector Machines

Support vector machines (SVM) are a new learning-by-example paradigm for classification and regression problems [9]. SVM have demonstrated significant efficiency when compared with neural networks. Their main advantage lies in the structure of the learning algorithm which consists of a constrained quadratic optimization problem (QP), thus avoiding the local minima drawback of NN.

The approach has its roots in statistical learning theory (SLT) and provides a way to build “optimum classifiers” according to some optimality criterion that is referred to as the maximal margin criterion [10]. An interesting development in SLT is the introduction of the Vapnik-

Chervonenkis (VC) dimension, which is a measure of the complexity of the model.

Equipped with a sound mathematical background [11], support vector machines treat both the problem of how to minimize complexity in the course of learning and how high generalization might be attained. This trade-off between complexity and accuracy led to a range of principles to find the optimal compromise. Vapnik and co-authors' [9] work have shown the generalization to be bounded by the sum of the training error and a term depending on the Vapnik-Chervonenkis (VC) dimension of the learning machine leading to the formulation of the structural risk minimization (SRM) principle. By minimizing this upper bound, which typically depends on the margin of the classifier, the resulting algorithms lead to high generalization in the learning process. Figure 2 illustrates the basic principle behind support vector machines with example instances lying on the margin on each side of a two class, two dimensional problems.

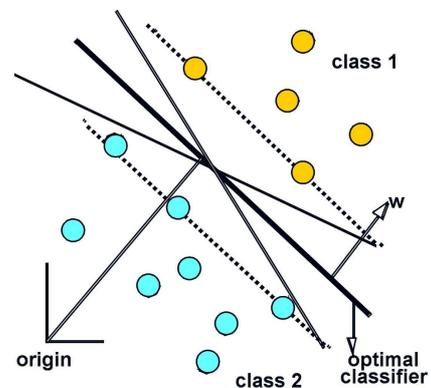


FIGURE 2

SVM MAXIMUM MARGIN CLASSIFIER.

A useful property of SVM is that loss functions lead to sparse solutions [12]. This means that, unlike regularization networks [13, 14], only a small fraction of the coefficients in the decision function are nonzero.

III. Support Vector Classification

We give a very brief review of SVM basic model principles; the reader is referred for more details to survey texts in [15, 16].

In the following we assume binary patterns where $y_i \in \mathbf{Y} \equiv \{\pm 1\}$. The learning method uses input-output training examples from the data set $D = \{(\mathbf{x}_i, y_i) \in \mathbf{X} \subseteq \mathbf{R}^N \times \mathbf{Y} : 1 \leq i \leq l\}$ such that f classifies correctly test data (\mathbf{x}, y) generated from the same underlying probability distribution $P(\mathbf{x}, y)$. In this framework, the SVM with Kernel $k(\cdot, \cdot)$ finds the minimizer of:

$$\frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathbb{F}_k}^2 \quad (1)$$

where V is a loss function that measures the discrepancy of the interpolating function output $f(\mathbf{x}_i)$ with respect to the

given output y_i , \mathbb{F}_k is the Reproducing Kernel Hilbert Space (RKHS) with Reproducing Kernel k and λ a positive parameter. The minimizer of (1) has the form:

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b \quad (2)$$

with $\alpha_i, b \in \mathbb{R}$.

The learning problem can be formulated minimizing the function (1) using the loss function defined by $V(y_i, f(\mathbf{x}_i)) = |1 - y_i f(\mathbf{x}_i)|_+$. The equivalent quadratic programming problem originally proposed in [8] is:

$$\min_{f \in \mathbb{F}_k, \xi} \Phi(f, \xi) = \frac{C}{l} \sum_{i=1}^l \xi_i + \frac{1}{2} \|f\|_k^2 \quad (3)$$

subject to constraints:

$$\begin{aligned} y_i f(\mathbf{x}_i) &\geq 1 - \xi_i \quad i = 1, \dots, l \\ \xi_i &\geq 0 \quad i = 1, \dots, l \end{aligned} \quad (4)$$

where C is the penalty constant (regularization parameter) and ξ the slack variable.

Introducing Lagrange multipliers:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (5)$$

with respect to α_i , under the constraints $0 \leq \alpha_i \leq \frac{C}{l}$, $i = 1, \dots, l$ and $\sum_{i=1}^l \alpha_i y_i = 0$, the solution has again the form (2). The empirical error measured by $\sum_{i=1}^l \xi_i$ is minimized while controlling capacity measured in terms of norm f .

IV. Clustering

Clustering techniques apply when the instances of data are to be divided into natural groups. The classical clustering technique is k-means where clusters are specified in advance prior to application of the algorithm. This corresponds to parameter k . Then k points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the Euclidian distance metric. Next the centroid, or mean, of each cluster center is calculated. These centroids are taken to be the new cluster centers for their respective clusters. The whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive runs. At this point the cluster centers have stabilized and will remain the same [17].

There are many variants of clustering even for the k-means algorithm depending upon the method of choosing the initial centers.

EXPERIMENTAL SETUP

We gathered data from the evaluation phase of a Discrete Structures course of Informatics Engineering Bachelor at the University of Coimbra during the first semester of 2006-2007.

The course was taught in Portuguese to 240 students. The evaluation phase we deal with corresponds to the evaluation of the practical component only. The theoretical part was evaluated in a final exam and is not considered here.

Moodle based Web-learning platform [18] has been partially used for course e-learning together with lecture notes in PowerPoint from previous years. However, our focus is on the evaluation phase which was specifically designed for the year under study.

Bachelor students were subjected to Quizzes graded with 25 points each. There were five Quizzes although only four were used for calculating the final grade corresponding to the practical component of the course.

After concluding the Quiz, students were encouraged to confer with the system feedback by checking the correct answers and instructors' comments.

The environment to execute the Quizzes is web based with strong restrictions to access other web services than the strictly necessary. The access to the Quizzes was restricted requiring a password and a network address associated to the selected rooms. Each Quiz had a fixed time to initiate and to finish and only one attempt was allowed to submit the final answers to the questions. The Quizzes were designed including several types of questions (Matching, Embedded Answers, Multiple Choice and Numerical).

An example of one of the five questions of Quiz 1 is shown in Figure 3. In this exercise, students are required to answer questions of Predicate Logic by understanding the given Tarski's World [19] through a convenient multi choice procedure. Each of the sub-questions has four possibilities and only one is correct. As Figure 3 caption indicates the corrected answers are shown.

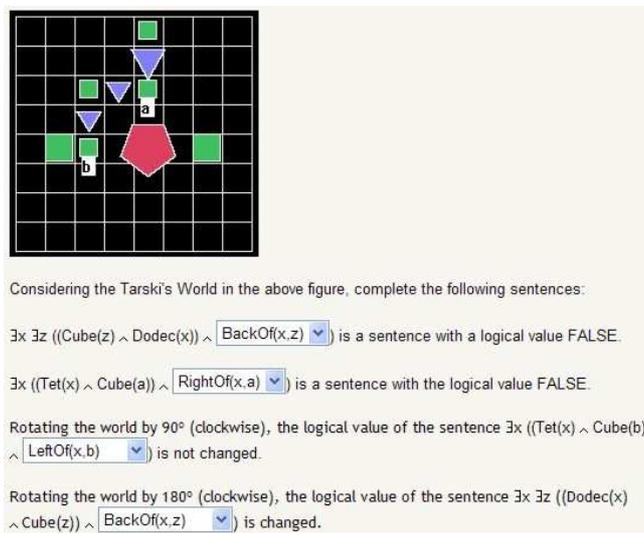


FIGURE 3
SAMPLE OF A QUESTION OF THE QUIZ Q1. IN THE FIGURE SCROLL BOXES
INDICATE CORRECT SELECTED ANSWERS

Another example is presented in Figure 4 where students are inquired about well-formed formulas. The main text book followed in the Discrete Structures course is written by Rosen [20].

Choose, given the following expressions, the true sentences:

1. $((P \rightarrow R) \vee Q)$ is a WFF (Well-Formed Formula).
2. $((P \rightarrow Q) \vee R) \vee P$ is not a WFF (Well-Formed Formula).
3. $(P \rightarrow (R \wedge (\neg P))) \rightarrow Q$ is not a WFF (Well-Formed Formula).

- Choose at least one answer.
- a. The expression 1 is a true sentence.
 - b. The expression 2 is a true sentence.
 - c. The expression 3 is a true sentence.

FIGURE 4
SAMPLE OF A QUESTION ON WFF OF THE QUIZ Q1.

The identification of relevant variables (also called features) is an essential component of construction of decision support models and computer-assisted discovery. Therefore from the student data logs we arranged the following features: i) grade obtained, ii) the time they used to solve the Quiz, iii) the number of hours of study for the respective Quiz, and iv) the final grade obtained by the student. The size of useful total amount of data is 186 since not all students were evaluated.

Table I illustrates the features used for the prediction model.

TABLE I
SAMPLE OF MOODLE LOGS DATA

Nc	TimeTakenQ1(m)	WorkloadQ1(hr)	Grade25Q1(pts)	WorkloadQ2(h)	Grade25Q2(pts)	WorkloadQ3(h)	Grade25Q3(p)	WorkloadQ4(hr)	Grade25Q4(pt)	class
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	43.0	5.0	13.02	12.0	8.0	12.0	5.9	12.0	9.45	FAIL
2	34.0	3.0	15.53	6.0	10.13	7.0	6.8	8.0	19.0	PASS
3	43.0	5.0	18.75	5.0	14.44	5.0	7.2	1.0	2.64	FAIL
4	44.0	0.0	13.28	2.0	15.08	0.0	0.0	0.0	0.0	FAIL
5	33.0	0.5	15.63	12.0	8.63	0.0	12.7	12.0	12.0	PASS
6	41.0	5.0	18.75	4.0	11.0	3.0	6.75	1.0	3.03	FAIL
7	39.0	2.0	7.03	2.0	8.75	1.0	0.43	0.0	0.0	FAIL
8	24.0	2.0	17.97	1.0	2.5	0.0	0.0	0.0	0.0	FAIL
9	28.0	1.0	17.71	0.0	14.5	1.0	5.95	3.0	17.8	PASS
10	42.0	3.0	15.63	6.0	7.65	8.0	14.3	8.0	13.95	PASS
11	35.0	2.0	5.08	3.0	10.75	3.0	3.83	2.0	0.0	FAIL
12	41.0	2.0	11.72	1.0	13.0	2.0	8.9	0.0	18.35	PASS
13	35.0	3.0	11.46	2.0	3.13	0.0	0.0	0.0	0.0	FAIL
14	41.0	10.0	10.94	5.0	13.38	6.0	0.83	12.0	9.6	FAIL
15	43.0	5.0	13.28	10.0	14.63	5.0	7.6	0.0	18.35	PASS
16	28.0	3.0	16.67	3.0	7.88	4.0	2.95	0.0	0.0	FAIL
17	43.0	0.0	16.63	4.0	19.13	12.0	6.0	4.0	4.8	FAIL
18	35.0	1.0	11.72	2.0	16.75	1.0	10.05	2.0	6.85	PASS
19	37.0	3.0	12.76	2.0	13.13	2.0	20.2	2.0	18.4	PASS
20	41.0	5.0	21.09	10.0	5.88	10.0	4.68	0.0	3.29	PASS
21	43.0	3.0	14.06	2.0	8.25	6.0	10.83	22.0	16.6	PASS
22	38.0	1.0	14.84	4.0	4.72	2.0	1.68	10.0	13.36	PASS
23	41.0	4.0	12.5	12.0	5.63	4.0	2.83	4.0	6.99	FAIL
24	42.0	5.0	11.72	7.0	13.38	10.0	0.0	0.0	11.05	PASS
25	26.0	1.0	11.72	1.0	1.5	1.0	8.33	1.0	13.0	FAIL
26	42.0	3.0	13.67	10.0	12.33	5.0	6.3	5.0	6.1	PASS
27	24.0	2.0	10.16	5.0	7.5	5.0	0.4	5.0	16.55	FAIL
28	39.0	4.0	20.31	5.0	19.33	10.0	13.88	5.0	6.1	PASS
29	30.0	1.0	7.81	2.0	11.53	2.0	3.63	0.0	0.0	PASS

The first Quiz Q0 was not compulsive and since we have not enough data therefore it is not included in Table I. However, we enter with it in the model since it indicates student's interest from the very beginning of the course.

Weka Data Mining Software Tools [16] were used in the study allowing multiple possibilities of choosing appropriate learning models.

RESULTS AND DISCUSSION

I. Performances Measures

The performance criteria used to evaluate the results were based on the metrics recall, precision, F1 and ROC curves. The first two measures are defined in terms of true positives (TP), false positives (FP) and false negatives (FN) as presented in (6) and (7).

$$recall = \frac{TP}{TP + FN} \quad (6)$$

$$precision = \frac{TP}{TP + FP} \quad (7)$$

A combined measure of above two is F1 and can be written by (8).

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (8)$$

Another popular measure is the ROC curve which is commonly used for binary problems. From the ROC curve it is very useful to calculate the Area Under Curve which gives a relative measure of performance and allows to compare efficiency among methods.

II. Analysis of Results

Figure 5 shows the knowledge flow system for the evaluation phase proposed. Several prediction models are illustrated although only the actual working model (SVM) is connected in the knowledge flow diagram.

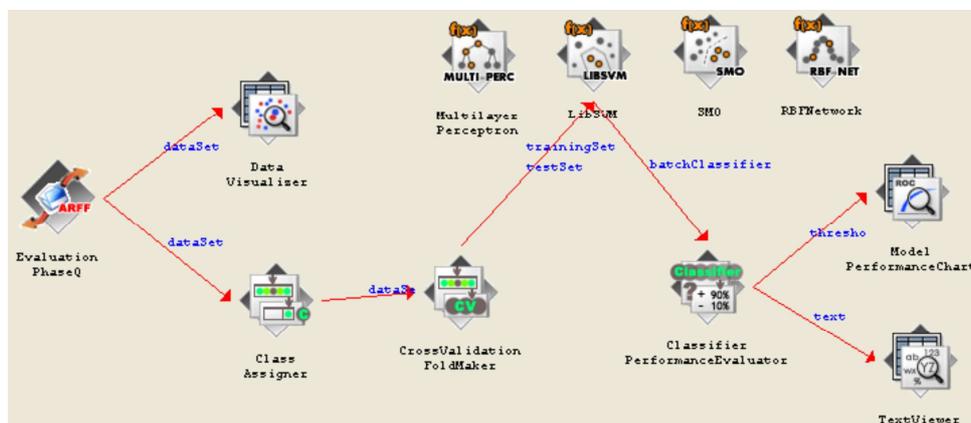


FIGURE 5
EVALUATION PHASE SYSTEM KNOWLEDGE FLOW

A first set of experiments have been set up by clustering variables (unsupervised models) to work out with the students' behavior. We found out that the two types of features describing student's behavior (initial interest in Quiz Q0 and workload (hr) along all the Quizzes) did not play a decisive role on the overall partition of the two clusters (FAIL/PASS) as compared to the scored grades. However, these types of variables provide useful feedback for instructors along the semester regardless their weak role on the overall models.

Two clustering algorithms (k-means and X-means) have been used. The best clustering result is with X-Means, a variant of k-means which takes into account the BIC (Bayesian Information Criteria) parameter. With X-means 85.61% of instances were correctly clustered whereas with k-means the result is 83.5% of correctly clustered instances. Clusters evaluation has been performed using prior information of classes.

A second set of experiments was performed using classification methods (supervised models). The results from the neural network prediction model are illustrated in Tables II and III. The performance attained is 82.26% for correctly classified instances. Regarding the SVM prediction model the performance is 86.56% as can be observed in Tables IV and V. This is due to superior ability in generalizing of SVM models as described earlier.

TABLE II
DETAILED ACCURACY BY CLASS IN NEURAL NETWORK

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.858	0.233	0.851	0.858	0.855	0.914	FAIL
0.767	0.142	0.778	0.767	0.772	0.914	PASS

TABLE III
CONFUSION MATRIX IN NEURAL NETWORK

	PREDICTED CLASS		Class
	FAIL	PASS	
ACTUAL CLASS	97	16	FAIL
	17	56	PASS

TABLE IV
DETAILED ACCURACY BY CLASS IN SVM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.903	0.192	0.879	0.903	0.891	0.951	FAIL
0.808	0.097	0.843	0.808	0.825	0.951	PASS

TABLE V
CONFUSION MATRIX IN SVM

	PREDICTED CLASS		Class
	FAIL	PASS	
ACTUAL CLASS	102	11	FAIL
	14	59	PASS

Figure 6 illustrates the ROC curves obtained with Neural Networks (two models have been analyzed (MLP) and (RBF) and with SVM.

The Area Under Curve (AUC) parameters are also indicated allowing to conclude superior performance of SVM. These indicators allow a fair comparison among used mining techniques.

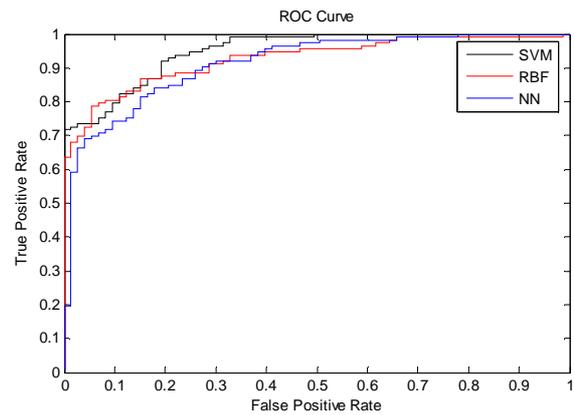


FIGURE 6
ROC CURVE FOR NN AND SVM CLASSIFIERS
PLOT AREA UNDER CURVE - SVM (0.954) RBF (0.926) NN(0.914)

CONCLUSIONS

There is a broad range of products available for e-learning which can be used in course curriculum at University level. Instructors gather a lot of information from web logs and, at a particular moment of time, they can understand and realize what the students' pattern behavior is. However, due to the amount of data stored along the course duration, relevant aspects are lost. Data mining techniques are crucial to build models of student's behavior based on their activity patterns.

While e-learning in education is well established, there are a few attempts to extract information in its evaluation phase. This work presents a prediction model for mining students' behavior pattern. The data obtained from the logs in a Moodle framework and data preparation for model construction is crucial for tracking students' behavior. The results show the model is able to successfully predict students' final outcome while bringing useful feedback during course making.

The experience was stimulating and worthwhile. Future work will focus on the improvement of the course designed features.

REFERENCES

- [1] J. A. S. Itmazi, *Sistema Flexible de gestión del elearning para soportar el aprendizaje en las universidades tradicionales*. PhD Thesis, University of Granada, Spain, 2005.
- [2] W. Rice, *Moodle E-Learning Course Development*. Packt Publishing, 2006.
- [3] P. Brusilovsky, "Adaptative Educacional Systems on the World-Wide-Web: A Review", in *Proc. Int. Conf. on Intelligent Tutoring Systems*, San Antonio, 1998.
- [4] H. García, C. Romero, S. Ventura and C. de Castro, "Using Rules Discovery for the Continuous Improvement of e-Learning Courses", in *Proc. IDEAL 2006*, LNCS, 2006, pp. 887-895.
- [5] J. Srivastava, B. Mobasher and R. Cooley, "Automatic Personalization Based on Web Usage Mining", *Communications of the Association of Computing Machinery*, pp. 142-151, 2000.
- [6] J. Li and O. R. Zaiane, "Combining Usage, Content and Structure Data to Improve Web Site Recommendation", in *Proc. Int. Conf. on Electronic Commerce and Web Technologies*, Spain, 2004.
- [7] V. Barnett and T. Lewis, *Outliers in Statistical Data*. John Wiley & Sons, 1994.

- [8] S. Haykin, *Neural Networks: A Comprehensive Foundation*. 2nd ed. (chapter 4), Prentice-Hall, New Jersey, 1999.
- [9] C. Cortes and V. Vapnik, "Support vector networks", *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [10] V. Vapnik and A. Chervonenkis, "Uniform convergence of frequencies of occurrence of events to their probabilities", *Dokl. Akad. Nauk SSSR*, vol. 181, pp. 915-918, 1968.
- [11] F. Cucker and S. Smale, "On the mathematical foundations of learning", *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1-49, 2001.
- [12] M. Pontil and A. Verri, "Properties of support vector machines", *Neural Computation*, vol. 10, pp. 955-974, 1997.
- [13] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks", *Science*, vol. 247, pp. 978-982, 1990.
- [14] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines", *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1-50, 2000.
- [15] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [16] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [17] I. H. Witten and E. Frank, *Data Mining – Practical Machine Learning Tools and Techniques*. 2nd ed., Elsevier, 2005.
- [18] J. Cole, *Using Moodle*. O'Reilly Media, Inc., 2005.
- [19] J. Barwise and J. Etchemendy, *Language, Proof and Logic*. CSLI Publications, 2002.
- [20] K. H. Rosen, *Discrete Mathematics and Its Applications*. McGraw-Hill Higher Education, 6th ed., 2006.