Artif Intell Rev (2012) 37:331–344
DOI 10.1007/s10462-011-9234-x

Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades

S. B. Kotsiantis

Published online: 12 May 2011

Abstract Use of machine learning techniques for educational proposes (or educational data mining) is an emerging field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns. The stored data (virtual courses, e-learning log file, demographic and academic data of students, admissions/registration info, and so on) can be useful for machine learning algorithms. In this article, we cite the most current articles that use machine learning techniques for educational proposes and we present a case study for predicting students' marks. Students' key demographic characteristics and their marks in a small number of written assignments can constitute the training set for a regression method in order to predict the student's performance. Finally, a prototype version of software support tool for tutors has been constructed.

Keywords Machine learning · Educational data mining · Decision support tools

1 Introduction

Use of machine learning techniques for educational proposes is an promising field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns (Baker and Yacef 2009). The stored data (virtual courses, e-learning log file, demographic and academic data of students, admissions/registration info, and so on) can be useful for machine learning algorithms. Recently, Romero et al. (2008) described the process of mining e-learning data step-by-step, as well as how to apply the main data mining techniques, such as statistics, visualization, classification, clustering and association rule mining of Moodle data.

Systems may treat user activities as individual items (either in an aggregated or event-based way) or consider activity sequences (Soller 2007). A further distinction can be made

S. B. Kotsiantis (✉)
Educational Software Development Laboratory, Department of Mathematics, University of Patras,
Patras, Greece
e-mail: sotos@math.upatras.gr

by the way data is analyzed later; during the last years a trend towards the combined use of data mining and machine learning techniques for the analysis of activity data can be observed (Romero and Ventura 2010).

Hershkovitz and Nachmias (2009) offer another classification of applications for educational data mining using two dimensions; one is whether a group or individual are the subject of research and the second is the time reference regarding learning—whether the end point or process are of interest.

This paper uses existing regression techniques in order to predict the students' marks in a distance learning system. It compares some of the state of the art regression algorithms to find out which algorithm is more appropriate not only to predict student's performance accurately but also to be used as an educational supporting tool for tutors. For the purpose of our study the 'informatics' course of the Hellenic Open University (HOU) provided the data set. Finally, a prototype version of software support tool for tutors has been constructed implementing the M5rules algorithm (Witten et al. 2011), which proved to be the most appropriate among the tested regression algorithms.

As more and more students enter online learning environments, databases concerning student access and study patterns will grow. Having information such as test scores and other student information available electronically can provide tutors with powerful tools for decision-making. We hope that the information produced by such decision support tools can be usefully applied by tutors to minimize the number of students who are probable to fail by providing them with extra teaching material or any other support.

The following section provides a recent survey of the usage of machine learning techniques for educational proposes. Section 3 describes in brief the Hellenic Open University (HOU) distance learning methodology and the data of our case study. Section 4 presents the experiment results for all the tested algorithms and at the same time compares these results. Section 5 presents the produced educational decision support tool. Finally, Sect. 6 discusses the conclusions and some future research directions.

2 Review of usage of machine learning techniques for educational proposes

Methods in usage of machine learning techniques for educational proposes include classification and regression algorithms, association rules, sequential pattern analysis, as well as clustering and web mining (Baker and Yacef 2009).

In e-learning, classification has been used for grouping students as hint-driven or failure-driven and finding students' common misconceptions (Yudelson et al. 2006); identifying learners with little motivation and finding remedial actions in order to lower drop-out rates (Cocea and Weibelzahl 2006) and for predicting course success (Hamalainen and Vinni 2006).

2.1 Use of classification algorithms for educational proposes

Hsia et al. (2008) used machine learning techniques to analyze the course preferences and course completion rates of enrollees in extension education courses at a university in Taiwan. First, extension courses were classified into five broad groups. Records of enrollees in extension courses from 2000–2005 were then analyzed by three data mining algorithms: Decision Tree, Link Analysis, and Decision Forest. Decision tree was used to find enrollee course preferences, Link Analysis found the correlation between course category and enrollee profession, and Decision Forest found the probability of enrollees completing preferred courses.

El-Alfy and Abdel-Aal (2008) propose an approach that uses abductive network modeling to automatically identify the most-informative subset of test items that can be used to effectively assess the examinees without seriously degrading accuracy. Using a training dataset of 1,500 cases (examinees) and 45 test items, the presented approach automatically selected only 12 items which classified an evaluation population of 500 cases with 91% accuracy.

Studies of attrition from courses have tended to concentrate on causation, trying, largely unsuccessfully, to elicit what causes drop out. However, the problem may more fruitfully be cast in terms of predicting who is likely to drop out. One powerful method for attempting to make predictions is rule induction. Moseley and Mead (2008) report the use of the Answer Tree package from SPSS for that purpose.

A dropout prediction method for e-learning courses, based on three popular machine learning techniques and detailed student data, is also proposed by Lykourentzou et al. (2009). The machine learning techniques used are feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP.

Using 5 years of institutional data along with several data mining techniques (both individuals as well as ensembles), Delen (2010) developed analytical models to predict and to explain the reasons behind freshmen student attrition. The comparative analyses results showed that the ensembles performed better than individual models, while the balanced dataset produced better prediction results than the unbalanced dataset.

The purpose of Wang et al. (2009) is to propose an adaptive system analysis for optimizing learning sequences. The analysis employs a decision tree algorithm, based on students' profiles, to discover the most adaptive learning sequences for a particular teaching content. The profiles were created on the basis of pretesting and posttesting, and from a set of five student characteristics: gender, personality type, cognitive style, learning style, and the students' grades from the previous semester.

The need for providing learners with web-based learning content that match their accessibility needs and preferences, as well as providing ways to match learning content to user's devices has been identified as an important issue in accessible educational environment. Guo and Zhang (2009) proposed a framework that utilizes machine learning algorithm for representing and extracting a dynamic learning process and learning pattern to support students' deep learning, efficient tutoring and collaboration in web-based learning environment.

Lee et al. (2009) propose to analyze learners' preferences with a machine learning algorithm. Findings in their study show that Field Independent learners frequently use backward/forward buttons and spent less time for navigation. On the other hand, Field Dependent learners often use main menu and have more repeated visiting.

Sacín et al. (2009) describe another machine learning approach in the context of educational systems that aims at predicting how suitable a specific course is for a specific student (based on the system's prediction of success for the respective course) via classification, in order to provide personalized recommendations.

Jantan et al. (2010) attempt to determine the potential classification techniques for academic talent forecasting in higher education institutions. Academic talents are considered as valuable human capital which is the required talents can be classified by using past experience knowledge discovered from related databases.

2.2 Use of regression algorithms for educational proposes

Campbell (2007) conducted a regression analysis of student academic performance and selected online activity data. These authors demonstrated that while student SAT scores are mildly predictive of future student success, the inclusion of a second variable—LMS logins—tripled the predictive power of this model. They also presented data indicating that students entering a course with low to moderate SAT scores could achieve success (measured by final grade) through above-average levels of effort (as indicated by number of LMS logins). Regression has been also applied for end-of-year accountability assessment scores (Anozie and Junker 2006).

Macfadyen and Dawson (2010) make an analysis of LMS tracking data from a Blackboard Vista-supported course identified 15 variables demonstrating a significant simple correlation with student final grade. Regression modeling generated a best-fit predictive model for this course which incorporates key variables such as total number of discussion messages posted, total number of mail messages sent, and total number of assessments completed and which explains more than 30% of the variation in student final grade.

There are a considerable number of machine learning approaches applied to personalization (Frias-Martinez et al. 2006). The personalization task can be viewed as a prediction problem: the system must attempt to predict the user's level of interest in, or the utility of, specific content categories, pages, or items, and then rank these according to their predicted values (Brusilovsky and Millán 2007).

2.3 Use of association rules for educational proposes

In order to discover common learning misconceptions of learners, Chen et al. (2007a) employ the association rule to mine the learner profile for diagnosing learners' common learning misconceptions during learning processes.

In the work of Buldu and Üçgün (2010), by being carried out Apriori algorithm upon the data of students of Istanbul Eyup I.M.K.B. Vocational Commerce High School, the rules have been produced and from the results obtained the relation between the courses that the students failed have been revealed.

From infrequent data, one can find a set of rare itemsets that will be useful for teachers to find out which students need extra help in learning. Weng (2011) developed a new algorithm based on the Apriori approach to mine fuzzy specific rare itemsets from quantitative data. The patterns are useful to discover learning problems.

Tseng et al. (2007) combined Fuzzy Set Theory, Education Theory and a machine learning approach to find grade fuzzy association rules. They proposed a Two-Phase Concept Map Construction (TP-CMC) algorithm with which to automatically construct a concept map of a course based upon historical testing records. Chen and Bai (2010) also present a new method to automatically construct concept maps based on data mining techniques for adaptive learning systems.

García et al. (2010) describe a collaborative educational data mining tool based on association rule mining for the ongoing improvement of e-learning courses and allowing teachers with similar course profiles to share and score the discovered information.

2.4 Use of clustering algorithms for educational proposes

Anaya and Boticario (2011) propose a method to analyze collaboration using machine learning techniques, so that it may be transferred to other collaborative learning environments. One approach grouped students according to their collaboration using clustering techniques, and the other approach provided metrics, built upon decision tree algorithms, which assigned a collaboration value to each student to support comparison on students' collaborative behavior.

Chen et al. (2007b) present a data-mining-based learning performance assessment scheme by combining four computational intelligence theories, i.e., gray relational analysis (GRA), K-means clustering scheme, fuzzy association rule mining, and fuzzy inference, in order to identify the learning performance assessment rules using the gathered Web-based learning portfolios of an individual learner. In other words, this scheme can help teachers assess the learning performance of the individual learner precisely utilizing only the learning portfolios in a Web-based learning environment. More significantly, teachers could understand the factors influencing learning performance in a Web-based learning environment according to the obtained interpretable learning performance assessment rules.

Hsu (2008) developed an online personalized English learning recommendation system capable of providing ESL students with reading lessons suited to their own interests that would therefore increase their motivation to learn. Amershi and Conati (2009) have found that undergraduate computer-science students' interactions with exploratory applets distinguish among successful and less successful learners: the former pause for longer times after presentation of a new problem and tend to follow suggested sequences and guidelines.

Lin et al. (2011) use two-stage clustering (SOM and K-means) under data mining theory to collect personnel training data of Automobile Corporation in Taiwan and China with data mining and analysis. The results under the two algorithms can serve as reference for future education training courses. In the end, in combination of back-propagation neural network to develop education training prediction model, the research offers reference for writing knowledge management system to enhance effects of personnel participation in training at corporations.

2.5 Use of sequential pattern mining for educational proposes

Sequential pattern mining has been previously used in e-learning although for different goals e.g to recommend sequences of resources for users to view in order to learn about a given topic (Cummins et al. 2006).

A different sequential pattern mining approach is described by Perera et al. (2009) where activity data monitored by the system is exploited to support mirroring, i.e., to extract and present patterns that characterize the behavior of successful groups.

Monitoring and interpreting sequential learner activities has the potential to improve adaptivity and personalization within educational environments. Köck and Paramythis (2011) present an approach based on the modeling of learners' problem solving activity sequences, and on the use of the models in targeted, and ultimately automated clustering, resulting in the discovery of new, semantically meaningful information about the learners.

2.6 Use of web mining techniques for educational proposes

The application of text mining techniques in e-learning can be used for grouping documents according to their topics and similarities and providing summaries (Hammouda and Kamel 2006).

Romero et al. (2009) propose an advanced architecture for a personalization system to facilitate Web mining. A specific Web mining tool is developed and a recommender engine is integrated into the AHA! system in order to help the instructor to carry out the whole Web mining process. The authors' objective is to be able to recommend to a student the most appropriate links/Web pages within the AHA! system to visit next. Web mining applications in virtual education environments were formed in accordance with descriptive research method in Sevindik et al. (2010).

Cohen and Nachmias (2010) provide a tool based on Web-log analysis for conducting comprehensive evaluations of Web based learning processes. By means of log-based measurement the instructors may learn about content usage, interpersonal interactions, learning assessments, student attitude, and reuse of course Websites and learning objects.

3 Case study

Zorrilla (2009) describes an application of data warehousing and OLAP technology applied to the e-learning field. The application tries to solve the lack a face-to-face student-teacher relationship providing instructors with information with which they can track and assess their students' progress and evaluate the design and planning of their virtual courses.

The mission of the Hellenic Open University (HOU) is to offer university level education using the distance learning methodology. The basic educational unit of the HOU is the course module (referred simply as module from now on) that covers a specific subject in graduate and postgraduate level. A module is equivalent to three semester academic lessons of Hellenic Universities while a student may register with up to three modules per year. The 'informatics' course of HOU is composed of 12 modules and leads to a Bachelor Degree. For the purpose of our study the 'informatics' course provided the training set. A total of 354 instances (student's records) have been collected from the module 'Introduction to Informatics' (INF10) (Xenos et al. 2002).

Regarding the INF10 module of HOU during an academic year students have to hand in 4 written assignments, optional participate in 4 face to face meetings with their tutor and sit for final examinations after an 11-month-period. A student with a mark ≥ 5 'passes' a lesson or a module while a student with a mark <5 'fails' to complete a lesson or a module.

Generally, a student must submit at least three assignments (out of 4). Subsequently, the tutors evaluate these assignments and a mark greater or equal to 20 should be obtained in total in order that each student successfully completes the INF10 module. Students who meet the above criteria may sit the final examination test.

3.1 Dataset of our study

The attributes (features) of our dataset are presented in Table 1 along with the values of every attribute. The set of the attributes was divided in 3 groups. The 'Registry Class', the 'Tutor Class' and the 'Classroom Class'. The 'Registry Class' represents attributes which were collected from the Student's Registry of the HOU concerning students' sex, age, marital status, number of children and occupation. In addition to the above attributes, the previous—post high school—education in the field of informatics and the association between students' jobs and computer knowledge were also taken into account. If a student has attended at least a seminar (of 100 h or more) on Informatics after high school then he/she would qualify as 'yes' in computer literacy. Moreover, students who use software packages (such as word processor) at their job without having any deep knowledge in informatics were considered as 'junior-users', while students who work as programmers or in data processing departments were considered a 'senior users'. The remaining students' jobs were listed as 'no' concerning association with computers.

Table 1 The attributes used and their values

Student's Registry (demographic) attributes	
Sex	Male, female
Age	24–46
Marital status	Single, married, divorced, widowed
Number of children	None, one, two or more
Occupation	No, part-time, fulltime
Computer literacy	No, yes
Job associated with computers	No, junior-user, senior-user
Attributes from tutors' records	
1st Face to face meeting	Absent, present
1st Written assignment	No, 0–10
2nd Face to face meeting	Absent, present
2nd Written assignment	No, 0–10
3rd Face to face meeting	Absent, present
3rd Written assignment	No, 0–10
4th Face to face meeting	Absent, present
4th Written assignment	No, 0–10
Class	
Final examination test	0–10

'Tutor Class' represents attributes, which were collected from tutors' records concerning students' marks on the written assignments and their presence or absence in face-to-face meetings. Finally, the 'class attribute' represents the result on the final examination test.

4 Experiments' results of regression algorithms

The problem of regression consists in obtaining a functional model that relates the value of a target continuous variable y with the values of variables $x_1, x_2, \ldots, x_n$ (the predictors). This model is obtained using samples of the unknown regression function. These samples describe different mappings between the predictor and the target variables.

For the propose of our comparison the six most common regression techniques namely Model Trees (Malerba et al. 2004), Neural Networks (Paliwala and Kumar 2009), Linear regression (Weisberg 2005), Locally weighted linear regression (Atkeson et al. 1997) and Support Vector Machines Shevade et al. 2000 are used. In the following we will briefly describe these regression techniques.

Linear regression is the simplest statistical technique used to find the best-fitting linear relationship between the class and its predictors (other features).

$$y = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Find values of beta that minimize Q:

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})^2$$

Note that nominal features with n values are converted into $n - 1$ binary features and a Wald test is used to test the statistical significance of each coefficient (β_ι) in the model

(Weisberg 2005). A standard linear regression method may employ an attribute deletion strategy, which simplifies the prediction task.

Model trees are the counterpart of decision trees for regression tasks. Model trees are trees that classify instances by sorting them based on attribute values. Instances are classified starting at the root node and sorting them based on their attribute values. The most well known model tree inducer is the M5 (Wang and Witten 1997). A model tree is generated in two stages. The first builds an ordinary decision tree, using as splitting criterion the maximization of the intra-subset variation of the target value (Malerba et al. 2004). The second prunes this tree back by replacing subtrees with linear regression functions wherever this seems appropriate. If this step is omitted and the target is taken to be the average target value of training examples that reach this leaf, then the tree is called a "regression tree" instead. Although the models trees are smaller and more accurate than the regression trees, the regression trees are more comprehensible (Witten et al. 2011).

M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5Model trees (Witten et al. 2011). The algorithm is able to deal with both continuous and nominal variables, and obtains a piecewise linear model of the data.

Artificial Neural Networks (ANNs) are another method of inductive learning based on computational models of biological neurons and networks of neurons as found in the central nervous system of humans (Paliwala and Kumar 2009). A multi layer neural network consists of large number of units (neurons) joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed, output units where the results of the processing are found, and units in between called hidden units. Regression with a neural network takes place in two distinct phases. First, the network is trained on a set of paired data to determine the input-output mapping. The weights of the connections between neurons are then fixed and the network is used to predict the numerical class values of a new set of data.

Locally weighted linear regression (LWR) is a combination of instance-based learning and linear regression (Atkeson et al. 1997). Instead of performing a linear regression on the full, unweighted dataset, it performs a weighted linear regression, weighting the training instances according to their distance to the test instance at hand. This means that a linear regression has to be done for each new test instance, which makes the method computationally quite expensive. However, it also makes it highly flexible, and enables it to approximate non-linear target functions.

The sequential minimal optimization algorithm (SMO) has been shown to be an effective method for training support vector machines (SVMs) on classification tasks defined on sparse data sets (Platt 1999). SMO differs from most SVM algorithms in that it does not require a quadratic programming solver. Shevade et al. (2000) generalize SMO so that it can handle regression problems. This implementation globally replaces all missing values and transforms nominal attributes into binary ones.

For the regression methods, there isn't only one regressor's criterion. Table 2 represents the most well known. Fortunately, it turns out for in most practical situations the best regression method is still the best no matter which error measure is used.

With the help of machine learning the tutors will be in position to know from the beginning of the module, based only on curriculum-based data of the students whose of them will complete the module with enough accurate precision, which reaches 64% in the initial forecasts and exceeds 80% before the middle of the period (Kotsiantis et al. 2004). After the middle of the period, we can use existing regression techniques in order to predict the students' marks.

Table 2 Regressor criteria (p: predicted values, a: actual values, $\bar{a} = \frac{1}{n}\sum_i a_i$)

Mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
Root mean squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_1-a_1)^2}{n}}$
Relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
Root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_1-a_1)^2}{(a_1-\bar{a})^2+\ldots+(a_1-\bar{a})^2}}$

Table 3 Mean absolute error

	M5′	*BP*	*LR*	*LWR*	*SMOreg*	*M5rules*
WRI-2	1.83	2.15	1.89	1.84	1.84	1.83
FTOF-3	1.74	2.08	1.83	1.79	1.78	1.74
WRI-3	1.55	1.79	1.6	1.53	1.56	1.55
FTOF-4	1.54	1.8	1.56	1.5	1.55	1.54
WRI-4	1.23	1.65	1.5	1.4	1.44	1.21

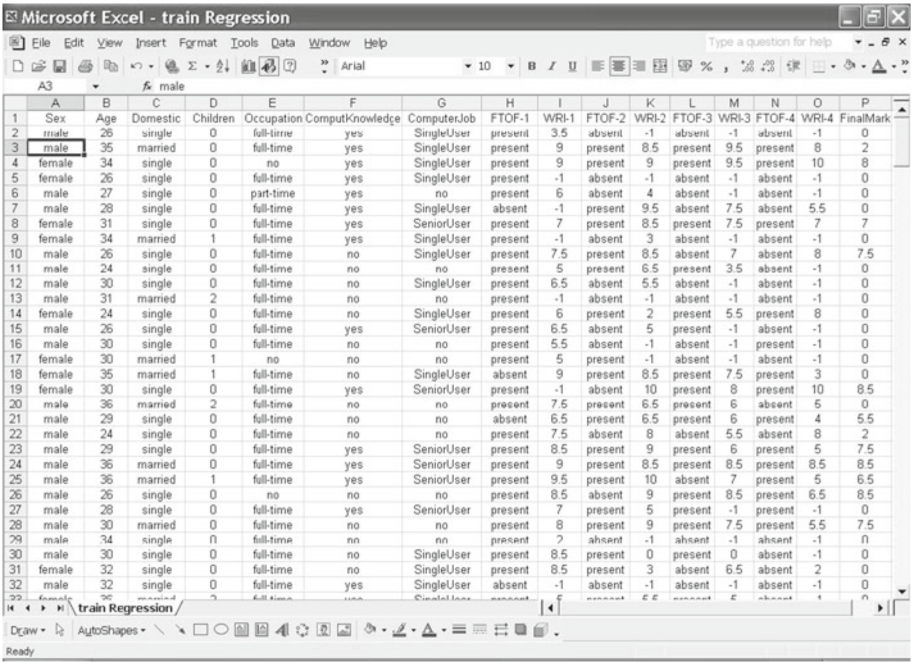
The experiments took place in two distinct phases. During the first phase (training phase) the algorithms were trained using the data collected from the academic year 2000–2001. The training phase was divided in 5 consecutive steps. The 1st step included the demographic data, the two first face-to-face meetings and written assignments as well as the resulting class (final mark). The 2nd step additionally included the third face-to-face meeting. The 3rd step additionally included the third written assignment. The 4th step additionally included the fourth face-to-face meeting and finally the 5th step that included all attributes described in Table 1.

Subsequently, ten groups of data for the new academic year (2001-2002) were collected from 10 tutors and the corresponding data from the HOU registry. Each one of these 10 groups was used to measure the accuracy within these groups (testing phase). The testing phase also took place in 5 steps. During the 1st step, the demographic data as well as the two first face-to-face meetings and written assignments of the new academic year were used to predict the class (final student mark) of each student. This step was repeated 10 times (for every tutor's data). During the 2nd step these demographic data along with the data from the third face-to-face meeting were used in order to predict the class of each student. This step was also repeated 10 times. During the 3rd step the data of the 2nd step along with the data from the third written assignment were used in order to predict the student class. The remaining steps use data of the new academic year in the same way as described above. These steps are also repeated 10 times.

It must be mentioned that we used the free available source code by (Witten et al. 2011) for our experiments. We have tried to minimize the effect of any expert bias by not attempting to tune any of the algorithms to the specific data set. Wherever possible, default values of learning parameters were used. This naïve approach results in lower estimates of the true mean absolute error, but it is a bias that affects all the learning algorithms equally.

In Table 3, the most easily understandable measure—mean absolute error—of each algorithm for all the testing steps of the experiment is presented.

According to the results, the M5rules is the most accurate regression algorithm to be used for the construction of a software support tool. An advantage of M5rules except for its better performance is its comprehensibility.


Microsoft Excel - train Regression
File Edit View Insert Format Tools Data Window Help
Type a question for help
Arial
10
A3
male
train Regression
Draw
AutoShapes
Ready

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Sex	Age	Domestic	Children	Occupation	ComputKnowledge	ComputerJob	FTOF-1	WRI-1	FTOF-2	WRI-2	FTOF-3	WRI-3	FTOF-4	WRI-4	FinalMark
2	male	26	single	0	full-time	yes	SingleUser	present	3.5	absent	-1	absent	-1	absent	-1	0
3	male	35	married	0	full-time	yes	SingleUser	present	9	present	8.5	present	9.5	present	8	2
4	female	34	single	0	no	yes	SingleUser	present	9	present	9	present	9.5	present	10	8
5	female	26	single	0	full-time	yes	SingleUser	present	-1	absent	-1	absent	-1	absent	-1	0
6	male	27	single	0	part-time	yes	no	present	6	absent	4	absent	-1	absent	-1	0
7	male	28	single	0	full-time	yes	SingleUser	absent	-1	present	9.5	absent	7.5	absent	5.5	0
8	female	31	single	0	full-time	yes	SeniorUser	present	7	present	8.5	present	7.5	present	7	7
9	female	34	married	1	full-time	yes	SingleUser	present	-1	absent	3	absent	-1	absent	-1	0
10	male	26	single	0	full-time	no	SingleUser	present	7.5	present	8.5	absent	7	absent	8	7.5
11	male	24	single	0	full-time	no	no	present	5	present	6.5	present	3.5	absent	-1	0
12	male	30	single	0	full-time	no	SingleUser	present	6.5	absent	5.5	absent	-1	absent	-1	0
13	male	31	married	2	full-time	no	no	present	-1	absent	-1	absent	-1	absent	-1	0
14	female	24	single	0	full-time	no	SingleUser	present	6	present	2	present	5.5	present	8	0
15	male	26	single	0	full-time	yes	SeniorUser	present	6.5	absent	5	present	-1	absent	-1	0
16	male	30	single	0	full-time	no	no	present	5.5	absent	-1	absent	-1	present	-1	0
17	female	30	married	1	no	no	no	present	5	present	-1	absent	-1	absent	-1	0
18	female	35	married	1	full-time	no	SingleUser	absent	9	present	8.5	present	7.5	present	3	0
19	female	30	single	0	full-time	yes	SeniorUser	present	-1	absent	10	present	8	present	10	8.5
20	male	36	married	2	full-time	no	no	present	7.5	present	6.5	present	6	absent	5	0
21	male	29	single	0	full-time	no	no	absent	6.5	present	6.5	present	6	present	4	5.5
22	male	24	single	0	full-time	no	no	present	7.5	absent	8	absent	5.5	absent	8	2
23	male	29	single	0	full-time	yes	SeniorUser	present	8.5	present	9	present	6	present	5	7.5
24	male	36	married	0	full-time	yes	SeniorUser	present	9	present	8.5	present	8.5	present	8.5	8.5
25	male	36	married	1	full-time	yes	SeniorUser	present	9.5	present	10	absent	7	present	5	6.5
26	male	26	single	0	no	no	no	present	8.5	absent	9	present	8.5	present	6.5	8.5
27	male	28	single	0	full-time	yes	SeniorUser	present	7	present	5	present	-1	present	-1	0
28	male	30	married	0	full-time	no	no	present	8	present	9	present	7.5	present	5.5	7.5
29	male	34	single	0	full-time	no	no	present	2	absent	-1	absent	-1	absent	-1	0
30	male	30	single	0	full-time	no	SingleUser	present	8.5	present	0	present	0	absent	-1	0
31	female	32	single	0	full-time	no	SingleUser	present	8.5	present	3	absent	6.5	absent	2	0
32	male	32	single	0	full-time	yes	SingleUser	absent	-1	absent	-1	absent	-1	absent	-1	0

Fig. 1 The CSV file of the use case

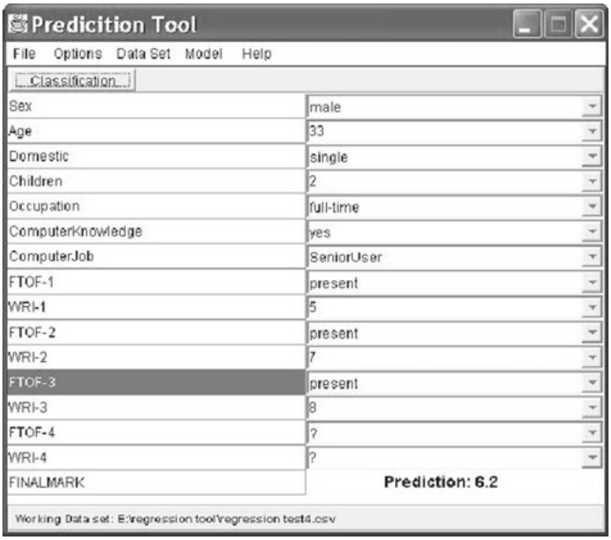
5 Decision support tool

A prototype version of the software support tool has already been constructed and is in use by the tutors. The tool expects the training set as a spreadsheet in CSV (Comma-Separated Value) file format (Fig. 1). The tool assumes that the first row of the CSV file is used for the names of the attributes. There is not any restriction in attributes' order. However, the class attribute must be in the last column. It must be mentioned that the used attributes are not a conclusive list. An extension can introduce new attributes that were not in the current database, but are collectable by tutors and may potentially contribute to the prediction of academic achievement. For example, measures of different intellectual abilities, interests, motivation, and personality traits of students.

Once the database is in a single relation, each attribute is automatically examined to determine its data type (for example, whether it contains numeric or symbolic information). A feature must have the value ? to indicate that no measurement was recorded. After opening the data set that characterizes the problem for which the user wants to take the prediction, the tool automatically uses the corresponding attributes for training.

After the training of the model, the user is able to see the produced regressor. The tool (Fig. 2) can also predict the output of either a single instance or an entire set of instances (batch of instances). It must be mentioned that for batch of instances the user must import an Excel cvs file with all the instances he/she wants to have predictions.

Subsequently, in an attempt to show how much each attribute influences the induction, the tool ranks the influence of each one according to a statistical measure—RRELIEF (Robnik-Šikonja and Kononenko 2003). The key idea of the RRELIEF algorithm is to estimate the quality of attributes according to how well their values distinguish between the instances



Sex	male
Age	33
Domestic	single
Children	2
Occupation	full-time
ComputerKnowledge	yes
ComputerJob	SeniorUser
FTOF-1	present
WRI-1	5
FTOF-2	present
WRI-2	7
FTOF-3	present
WRI-3	8
FTOF-4	?
WRI-4	?
FINALMARK	**Prediction: 6.2**

Working Data set: E:\regression tool\regression test4.csv

Fig. 2 The prototype tool

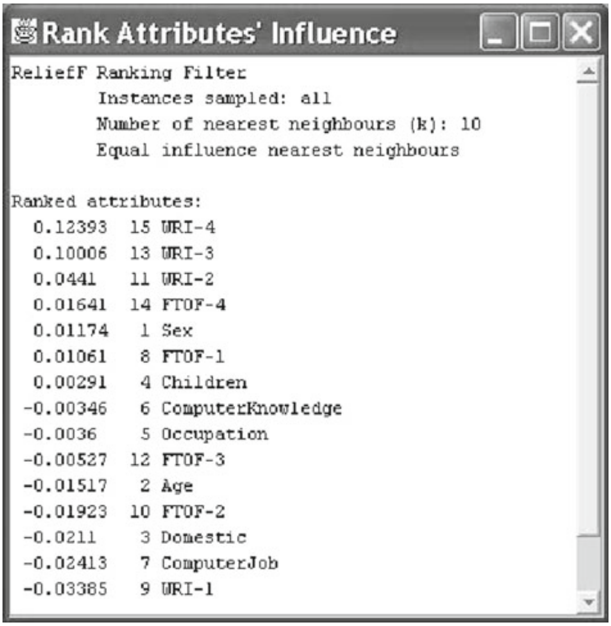
that are near to each other. In regression problems the predicted value (class) is continuous; therefore the (nearest) hits and misses cannot be used. Instead of requiring the exact knowledge of whether two instances belong to the same class or not, we can introduce a kind of probability that the predicted values of two instances are different. This probability can be modeled with the relative distance between the predicted (class) values of the two instances.

The ranking of the attributes' influence brought considerable benefits; by helping the tutors to better understand the characteristics of the population that mostly affect academic achievement. For example, the prototype tool for the used dataset shows that the attributes that mostly influence the induction are the 'WRI-4' and the 'WRI-3' (Fig. 3).

What is more, the implemented tool can present useful information about the imported data set such as the presence or not of missing attribute values, the frequency of each attribute value etc. Finally, the tool provides on-line help for novice users.

6 Conclusion

Generally, the education domain offers many interesting and challenging applications for machine learning. Firstly, an educational institution often has many diverse and varied sources of information. There are the traditional databases (e.g. students' information, teachers' information, class and schedule information, alumni information), online information (online web pages and course content pages) and more recently, multimedia databases. Secondly, there are many diverse interest groups in the educational domain that give rise to many interesting mining requirements. For example, the administrators may wish to find out information such as admission requirements and to predict the class enrollment size for timetabling. The students may wish to know how best to select courses based on prediction of how well they will perform in the courses selected. With so much information and so many diverse needs, it is foreseeable that an integrated machine learning system that is able to cater for the special



```
ReliefF Ranking Filter
        Instances sampled: all
        Number of nearest neighbours (k): 10
        Equal influence nearest neighbours

Ranked attributes:
  0.12393   15 WRI-4
  0.10006   13 WRI-3
  0.0441    11 WRI-2
  0.01641   14 FTOF-4
  0.01174    1 Sex
  0.01061    8 FTOF-1
  0.00291    4 Children
 -0.00346    6 ComputerKnowledge
 -0.0036     5 Occupation
 -0.00527   12 FTOF-3
 -0.01517    2 Age
 -0.01923   10 FTOF-2
 -0.0211     3 Domestic
 -0.02413    7 ComputerJob
 -0.03385    9 WRI-1
```

Fig. 3 Ranking the attributes' influence to the final prediction in our use case

needs of an education institution will be in great demand particularly in the twenty-first century.

This paper aims to fill the gap between empirical prediction of student performance and the existing regression techniques. Our data set is from the module INFO but most of the conclusions are wide-ranging and present interest for the majority of programs of study of Hellenic Open University and more generally for all the distance education programs.

In a next study we intend to apply machine learning methods in the Hellenic Open University data repository with the goal of answering the following research question: Can we classify the learning difficulties of the students? If so, can we show how different types of problems impact students' achievement? Can we help instructors to develop the homework more effectively and efficiently?

Appendix

The tool is available in the web page: http://www.math.upatras.gr/~esdlab/oldEsdlab/Regression-tool/ The Java Virtual Machine (JVM) 1.2 or newer is needed for the execution of the program.

References

Amershi S, Conati C (2009) Combining unsupervised and supervised classification to build user models for exploratory learning environments. J Educ Data Min 1(1):18–71

Anaya AR, Boticario JG (2011) Application of machine learning techniques to analyse student interactions and improve the collaboration process. Expert Syst Appl 38:1171–1181

Anozie N, Junker BW (2006) Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Educational data mining AAAI workshop, California, USA, pp 1–6

Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artificial Intell Rev 11(1–5):11–73

Baker RSJD, Yacef K (2009) The state of educational data mining in 2009: a review and future visions. J Educ Data Min 1(1):3–17

Brusilovsky P, Millán E (2007) User models for adaptive hypermedia and adaptive educational systems. In The adaptive web. LNCS 4321, Springer, Berlin, pp 3–53

Buldu A, Üçgün K (2010) Data mining application on students' data. Procedia Soc Behav Sci 2:5251–5259

Campbell J (2007) Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study. Doctoral thesis, Purdue University, Indiana, USA

Chen S-M, Bai S-M (2010) Using data mining techniques to automatically construct concept maps for adaptive learning systems. Expert Syst Appl 37:4496–4503

Chen CM, Hsieh YL, Hsu SH (2007a) Mining learner profile utilizing association rule for web-based learning diagnosis. Expert Syst Appl 33(1):6–22

Chen C-M, Chen Y-Y, Liu C-Y (2007b) Learning Performance Assessment Approach Using Web-Based Learning Portfolios for E-learning Systems, IEEE Trans Syst Man Cybern C Appl Rev 37(6) November

Cocea M, Weibelzahl S (2006) Can log files analysis estimate learners' level of motivation? In: Proceedings of the workshop week Lernen—Wissensentdeckung—Adaptivität, Hildesheim, pp 32–35

Cohen A, Nachmias R (2010) What can instructors and policy makers learn about Web-supported learning through Web-usage mining, Internet and Higher Education. doi:10.1016/j.iheduc.2010.07.008

Cummins D, Yacef K, Koprinska I (2006) A sequence based recommender system for learning resources. Aust J Intell Inform process Syst 9:49–56

Delen D (2010) A comparative analysis of machine learning techniques for student retention management. Decis Support Syst 49:498–506

El-Alfy EM, Abdel-Aal RE (2008) Construction and analysis of educational tests using abductive machine learning. Comput Educ 51:1–16

Frias-Martinez E, Chen SY, Liu X (2006) Survey of data mining approaches to user modeling for adaptive hypermedia. IEEE Trans Syst Man Cybern C Appl Rev 36(2):734–748

García E, Romero C, Ventura S, de Castro C (2010) A collaborative educational association rule mining tool, A collaborative educational association rule mining tool, Internet and Higher Education. doi:10.1016/j.iheduc.2010.07.006

Guo Q, Zhang M (2009) Implement web learning environment based on data mining. Knowl-Based Syst 22:439–442

Hamalainen W, Vinni M (2006) Comparison of machine learning methods for intelligent tutoring systems. In: Proceedings of the eighth international conference in intelligent tutoring systems, Taiwan, pp 525–534

Hammouda K, Kamel M (2006) Data mining in e-learning. In: Pierre S (ed), E-learning networked environments and architectures: a knowledge processing perspective, Springer Book Series: Advanced information and knowledge processing, pp 1–28

Hershkovitz A, Nachmias R (2009) Learning about online learning processes and students' motivation through web usage mining. Interdiscip J E-Learning and Learning Objects, 5,197–215, Special series of Chais Conference 2009 best papers

Hsia T-C, Shie A-J, Chen L-C (2008) Course planning of extension education to meet market demand by using data mining techniques—an example of Chinkuo technology university in Taiwan. Expert Syst Appl 34:596–602

Hsu MH (2008) A personalized English learning recommender system for ESL students. Expert Syst Appl 34(1):683–688

Jantan H , Hamdan AR, Othman ZA (2010) Classification and prediction of academic talent using data mining techniques, KES 2010, Part I, LNAI 6276, pp 491–500

Kock M, Paramythis A (2011) Activity sequence modeling and dynamic clustering for personalized e-learning, User Model User-Adap Inter. doi:10.1007/s11257-010-9087-z

Kotsiantis S, Pierrakeas C, Pintelas P (2004) Predicting Students' Performance in distance learning using machine learning techniques. Appl Artif Intell (AAI) 18(5):411–426

Lee MW, Chen SY, Chrysostomou K, Liu X (2009) Mining students' behavior in web-based learning programs. Expert Syst Appl 36:3459–3464

Lin WT, Wang SJ, Wub YC, Ye TC (2011) An empirical analysis on auto corporation training program planning by data mining techniques. Expert Syst Appl 38:5841–5850

Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V (2009) Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput and Educ 53:950–965

Macfadyen LP, Dawson S (2010) Mining LMS data to develop an “early warning system” for educators: a proof of concept. Comput Educ 54:588–599

Malerba D, Esposito F, Ceci M (2004) Top–down induction of model trees with regression and splitting nodes. IEEE Trans Pattern Anal Mach Intell 26(5):612–625

Moseley LG, Mead DM (2008) Predicting who will drop out of nursing courses: a machine learning exercise. Nurse Educ Today 28:469–475

Paliwala M, Kumar UA (2009) Neural networks and statistical techniques: a review of applications. Expert Syst Appl 36(1):2–17

Perera D, Kay J, Koprinska I, Yacef K, Zaïne OR (2009) Clustering and sequential pattern mining of online collaborative learning data. IEEE Trans Knowl Data Eng 21(6):759–772

Platt J (1999) Using sparseness and analytic QP to speed training of support vector machines. In: Kearns MS, Solla SA, Cohn DA (eds) Advances in neural information processing systems 11. MIT Press, MA

Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of reliefF and RReliefF. Mach Learn 53(1–2):23–69

Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. Expert Syst Appl 33(1):135–146

Romero C, Ventura S (2010) Educational data mining: a review of the state-of-the-art. IEEE Trans Syst Man Cybernet C Appl Rev 40(6):601–618

Romero C, Ventura S, García E (2008) Data mining in course management systems: moodle case study and tutorial. Comput Educ 51(1):368–384

Romero C, Ventura S, Zafra A, de Bra P (2009) Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. Comput Educ 53:828–840

Sevindik T, Demirkeser N, Cömert Z (2010) Virtual education environments and web mining. Procedia Soc Behav Sci 2:5120–5124

Shevade S, Keerthi S, Bhattacharyya C, Murthy K (2000) Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Netw 11(5):1183–1188

Soller A (2007) Adaptive support for distributed collaboration. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The AdaptiveWeb, vol 4321 of Lecture Notes in Computer Science. Springer, Berlin pp 573–595

Tseng SS, Sue PC, Su JM, Weng JF, Tsai WN (2007) A new approach for constructing the concept map. Comput and Educ 49(3):691–707

Vialardi Sacín C, Bravo Agapito J, Shafti L, Ortigosa A (2009) Recommendation in higher education using data mining techniques. 2nd international conference of educational data mining 2009, Spain, 1–3 July, 2009, 190–199

Wang Y-h, Tseng M-H, Liao H-C (2009) Data mining for adaptive learning sequence in english language instruction. Expert Syst Appl 36:7681–7686

Wang Y, Witten IH (1997) Induction of model trees for predicting continuous classes, In: Proceedings of the poster papers of the european conference on ML, Prague . Prague: university of economics, Faculty of informatics and statistics, pp 128–137

Weisberg S (2005) Appl Linear Regres, 3rd Edition, ISBN: 978-0-471-66379-9

Weng C-H (2011) Mining fuzzy specific rare itemsets for education data, Knowl-Based Syst. doi:10.1016/j.knosys.2011.02.010

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques (Third Edition), Morgan Kaufmann, January, ISBN 978-0-12-374856-0

Xenos M, Pierrakeas C, Pintelas P (2002) A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the Hellenic Open University. Comput Educ 39:361–377

Yudelson MV, Medvedeva O, Legowski E, Castine M, Jukic D, Rebecca C (2006) Mining student learning data to develop high level pedagogic strategy in a medical ITS. In: Proceedings of AAAI workshop on educational data mining, Boston, pp 1–8

Zorrilla ME (2009) Data Warehouse Technology for E-Learning. In: Zakrzewska D et al. (eds) Meth and Support Tech for Data Analys, SCI 225, pp 1–20