Predicting Students Marks in Hellenic Open University

Conference Paper · January 2005

DOI: 10.1109/ICALT.2005.223 · Source: DBLP

2 authors:

Sotiris Kotsiantis
University of Patras
239 PUBLICATIONS **11,038** CITATIONS
SEE PROFILE

P. E. Pintelas
University of Patras
179 PUBLICATIONS **5,966** CITATIONS
SEE PROFILE

Predicting Students' Marks in Hellenic Open University

Sotiris B. Kotsiantis & Panayiotis E. Pintelas
Educational Software Development Laboratory
Department of Mathematics University of Patras
{sotos, pintelas}@math.upatras.gr

Abstract

The ability to provide assistance for a student at the appropriate level is invaluable in the learning process. Not only does it aids the student's learning process but also prevents problems, such as student frustration and floundering. 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. The scope of this work compares some of the state of the art regression algorithms in the application domain of predicting students' marks. A number of experiments have been conducted with six algorithms, which were trained using datasets provided by the Hellenic Open University. Finally, a prototype version of software support tool for tutors has been constructed implementing the M5rules algorithm, which proved to be the most appropriate among the tested algorithms.

1. Introduction

The application of Machine Learning Techniques in predicting students' performance proved to be helpful for identifying poor performers and it can enable tutors to take remedial measures at an earlier stage, even from the very beginning of an academic year using only students' demographic data, in order to provide additional help to the groups at risk [4]. The diagnosis of students' performance is increased as new curriculum data is entered during the academic year, offering the tutors more effective results. It was showed in [4] that the most accurate machine learning algorithm for identifying predicted poor performers is the Naïve Bayes Classifier. However, that work could only predict if a student passes a course module or not.

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.

Generally, the usage of regression analysis to classify data can be an extremely useful tool for researchers and Open University administrators. A plethora of data can be utilized simultaneously to classify cases and the resultant model can be evaluated for usefulness relatively easily. The ability to develop a predictive model based on the model produced through the regression analysis procedure increases its usefulness substantially. Open Universities can utilize this dynamic and powerful procedure to target services and interventions to students who need it most, thereby utilizing their resources more effectively.

The following section describes in brief the Hellenic Open University (HOU) distance learning methodology and the data of our study. Some very basic definitions about regression techniques are given in section 3. 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, section 6 discusses the conclusions and some future research directions.

2. Hellenic Open University and Data Description

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. For the purpose of our study the 'informatics' course provided the training set. A total of 354 instances (student's

Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)
0-7695-2338-2/05 $20.00 © 2005 **IEEE**

records) have been collected from the module 'Introduction to Informatics' (INF10) [12].

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.

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 hours 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.

'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.

The analysis of the demographic attributes showed that the ratio of men who passed the exams vs. men who failed is 48–52%, while for women this ratio drops to 39–61%. Moreover, it should be noted that the percentage of students below 32 years old that pass the exams is measured 46%, when the corresponding number for older students is 44%. Another interesting fact is related to student performance and their marital status. It is just as possible for a married student to pass the exams (51%) while a single student has only 41% probability to pass the module. A similar situation holds with the existence of children, a student with children has 52% probability to pass the module while a student without children has only 43%. This is probably due to the fact that the family obligations is known and has been taken under consideration prior to the commencement of the studies. It must be also mentioned that the workload separates the probabilities just in the middle.

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

On the contrary, as far as the demographic attributes are concerned, stronger correlation exists between student performance and the existence of previous education in the field of Informatics. The ratio of students who have previous education in the field of Informatics and pass the exams vs. them who fail is 51–49%, while for the remaining students this ratio drops to 28–72%. A similar correlation exists between the involvements in professional activities demanding the use of computer. The students who use the computer in their job have 52% probability to pass the module while the remaining students have only 32%.

Until now, we have described how each demographic attribute influences the prediction based on our dataset. In order to show in which direction

(pass or fail) each of the remaining attributes' values push the induction in Table 2 some practical probabilities are estimated. The interpretation of Table 2 is easy enough and it shows, for example, that a student with a mark more than 6 in WRI-4, has about 4 times more probabilities to pass than fail (0.65/0.17).

Table 2.Influence of each attribute

Attribute	*Value*	*Pass*	*Fail*
WRI-4	Mark<3	0.04	0.68
	3=<Mark=<6	0.31	0.15
	Mark>6	0.65	0.17
WRI-3	Mark<3	0.03	0.61
	3=<Mark=<6	0.21	0.2
	Mark>6	0.66	0.19
WRI-2	Mark<3	0.08	0.52
	3=<Mark=<6	0.15	0.26
	Mark>6	0.77	0.22
FTOF-4	Absent	0.23	0.76
	Present	0.77	0.24
FTOF-3	Absent	0.2	0.65
	Present	0.8	0.35
WRI-1	Mark<3	0.02	0.19
	3=<Mark=<6	0.14	0.35
	Mark>6	0.84	0.46
FTOF-2	Absent	0.22	0.54
	Present	0.78	0.46

Subsequently, in an attempt to show how much each attribute influences the induction, we ranked the influence of each one according to a statistical measure – RRELIEF [9]. The demographic attributes that mostly influence the induction are the 'sex' and the 'children'. In addition, it was found that 1st written assignment has not a large value of influence. The reason is that almost all students try harder with the first written assignment thus making the offered information of this attribute minimal and maybe confusing.

3. Regression Issues

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, ..., 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 [10], Neural Networks [5], Linear regression (LR) [2], Locally weighted linear regression (LWR) [1] and Support Vector Machines [7] are used.

The most well known model tree inducer is the M5´ [10]. M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5´Model trees [11]. BP is the most well known algorithms for training Neural Networks. The sequential minimal optimization algorithm (SMO) SMO differs from most SVM algorithms in that it does not require a quadratic programming solver. In [8] SMO is generalized so that it can handle regression problems (SMOreg).

4. Experiments Results

The learning algorithms are useful as a tool for identifying predicted poor performers [3]. 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 [4]. After the middle of the period, we can use existing regression techniques in order to predict the students' marks.

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 2000-1. 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 (2001-2) 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 [11] for our experiments. In Table 3, the most easily understandable measure - mean absolute error: $\left(\left|p_1 - a_1\right| + \ldots + \left|p_n - a_n\right|\right)/n$ where p_i: predicted values, a_i: actual values and $\bar{a} = \frac{1}{n}\sum_i a_i$ - of each algorithm for all the testing steps of the experiment is presented.

Table 3. Mean absolute error of each algorithm for all the testing steps

	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

According to the results, the M5rules is the most accurate regression algorithm to be used for the construction of a software support tool (even though in most of the testing steps there is not statistically significant difference between algorithms according to the corrected resampled t-test [6]). However, another advantage of M5rules except for its better performance is its better comprehensibility.

5. Software 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. 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 is available in the web page: http://www.math.upatras.gr/~esdlab/Regression-tool/). The tool (Figure 1) 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.

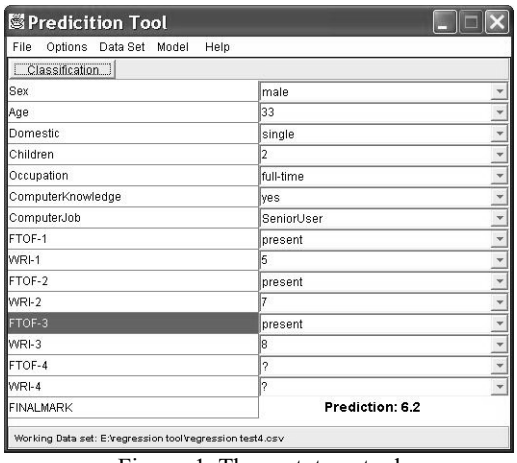

Figure. 1. The prototype tool

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' (Figure 2).

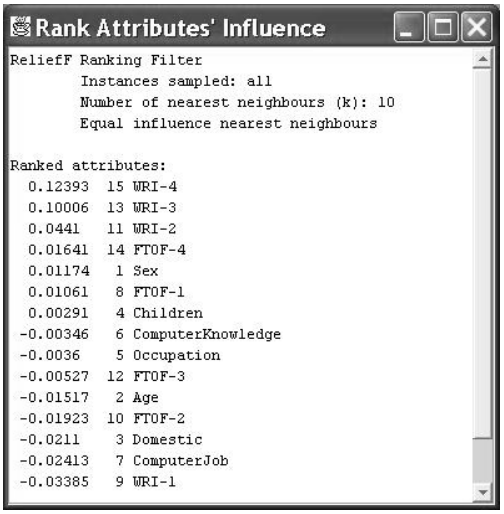

Figure 2. Ranking the attributes' influence to the final prediction in our use case

6. Conclusion

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. It would be interesting to compare our results with those from other open and distance learning programs offered by other open Universities. So far, however, we have not been able to find such results.

Generally, the education domain offers many interesting and challenging applications for data mining. 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.

In a next study we intend to apply data mining methods with the goals of answering the following two research questions:

1) Are there groups of students who use online resources in a similar way? Based on the usage of the resource by other students in the group, can we help a new student use the resources better?
2) Can we classify the learning difficulties of the students? Can we help instructors to develop the homework more effectively and efficiently?

7. References

[1] Atkeson, C. G., Moore, A.W., & Schaal, S. (1997). Locally weighted learning. Artificial Intelligence Review, 11, 11–73.

[2] Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, ISBN: 080394540X, Sage Pubns.

[3] Kotsiantis, S., Pierrakeas, C., Pintelas, P.(2003), Preventing student dropout in distance learning systems using machine learning techniques, Lecture notes in AI, Springer-Verlag Vol 2774, pp 267-274.

[4] Kotsiantis S., Pierrakeas C., Pintelas P. (2004), Predicting Students' Performance in Distance Learning Using Machine Learning Techniques, Applied Artificial Intelligence (AAI), Volume 18, Number 5 / May-June 2004, pp. 411 - 426.

[5] Mitchell, T. (1997), Machine Learning. McGraw Hill.

[6] Nadeau, C., Bengio, Y. (2003), Inference for the Generalization Error. Machine Learning, 52, 239-281.

[7] Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In: Kearns, M. S., Solla, S. A. & Cohn D. A. (Eds.), Advances in neural information processing systems 11. MA: MIT Press.

[8] Shevade, S., Keerthi, S., Bhattacharyya C., and Murthy, K. (2000). Improvements to the SMO algorithm for SVM regression. IEEE Transaction on Neural Networks, 11(5):1188-1183.

[9] Sikonja M. and Kononenko I. (1997), An adaptation of Relief for attribute estimation in regression, Proceedings of the Fourteenth International Conference (ICML'97), ed., Dough Fisher, pp. 296-304. Morgan Kaufmann Publishers.

[10] Wang, Y. & Witten, I. H. (1997). Induction of model trees for predicting continuous classes, In Proc. of the Poster Papers of the European Conference on ML, Prague (pp. 128–137).

[11] Witten, I.H., Frank, E. (2000), Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA.

[12] Xenos, M., Pierrakeas C. and 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, Computers & Education (39): 361–377.