Using Genetic Algorithms and Decision Trees for *a posteriori* Analysis and Evaluation of Tutoring Practices based on Student Failure Models

Dimitris Kalles and Christos Pierrakeas
Hellenic Open University, Laboratory of Educational Material and
Educational Methodology, Sachtouri 23, 26222, Patras, Greece
{kalles, pierrakeas}@eap.gr

Abstract. Many students who enrol in the undergraduate program on informatics at the Hellenic Open University (HOU) fail the introductory course exams and drop out. We analyze their academic performance, derive short rules that explain success or failure in the exams and use the accuracy of these rules to reflect on specific tutoring practices that could enhance success.

1 Introduction

The Hellenic Open University's (HOU) primary goal is to offer university-level education using distance learning methods and to develop the appropriate material and teaching methods to achieve this goal. The HOU offers both undergraduate and postgraduate studies and its courses were initially designed and first offered in 1998 following the distance learning methodology of the British Open University. The HOU was founded in 1992 and currently (2005) nearly 25,000 students are enrolled.

The undergraduate programme in informatics is heavily populated, with more than 2,000 enrolled students. About half of them currently attend junior courses on mathematics, software engineering, programming, databases, operating systems and data structures. A key observation is that substantial failure rates are consistently reported at the introductory courses.

Such failures skew the academic resources of the HOU system towards filtering the input rather than polishing the output, from a quantitative point of view. Even though this may be perfectly acceptable from an educational, political and administrative point of view, we must analyse and strive to understand the mechanism and the reasons of failure. This could significantly enhance the ability of HOU to fine-tune its tutoring and admission policies without compromising academic rigour.

Please use the following format when citing this chapter:
Kalles, Dimitris, Pierrakeas, Christos, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 9–18

There are two key educational problems that have been identified as being core aspects of these failures. The first is that these courses are heavy on mathematics and adult students have not had many opportunities to sharpen their mathematical skills since high-school graduation (which has typically occurred at about 10 years prior to enrolling at HOU). The second is that the lack of a structured academic experience may have rendered dormant one's general learning skills and attitudes.

Our approach to investigating this problem uses increasingly rudimentary technology for data analysis. We use genetic algorithms to derive short decision trees that explain student failure [1, 2].

In this paper we expand that work by investigating differences in the accuracy of the induced models. We focus on short models that are easier to communicate among peers and question whether these differences might be attributed to the versatility of the tutoring practices. The results support our intuition about which practices better smooth out the disadvantages that arise due to some students' special circumstances. These results are now used as supporting data when we attempt to convince fellow tutors of the potential of some specific tutoring practices.

This paper is structured in three subsequent sections. In the next section, we briefly review the problem of predicting student performance at large, and the related techniques we have been using at HOU. We then single out three modules which have clearly different policies in dealing with students who have failed an exam and devise a set of experiments to observe whether these policies can be evaluated by a machine learning model. Finally, we argue about the ability to carry out these experiments at a larger scale and discuss the potential implications of our findings from an educational point of view.

2 Background

The work reported in this paper is part of an effort to analyze data at an institutional level, so we first briefly cover some essential background. We first present the application domain, then we present some key aspects of the technology used and, finally, we summarize the results obtained to date.

2.1 Operational issues

The educational philosophy of Open Universities around the world is to promote "life long education" and to provide adults with "a second educational chance" [3]. The method used is known as "distance learning" education, hence the widely used acronym ODL standing for Open-and-Distance-Learning.

In open and distance learning, dropout rates are definitely higher than those in conventional universities. Relatively recently, the Open Learning journal published a volume on issues on student retention in open and distance learning, where similarities and differences across systems is discussed, highlighting issues of institutions, subjects and geographic areas [4].

The vast majority (up to 98%) of registered students in the "Informatics" program, upon being admitted at HOU, selects the module "Introduction to

Informatics" (INF10). Following that, and according to university recommendations, they will typically select the modules "Fundamental Software Engineering" (INF11) and "Mathematics" (INF12). These modules are the most heavily populated and serve as test-beds for experimentation.

A module is the basic educational unit at HOU. It runs for about ten months and is the equivalent of about 3-4 conventional university semester courses. A student may register with up to three modules per year. For each module, a student is expected to attend five plenary class meetings throughout the academic year (a class contains about thirty students). Each meeting is about four hours long and may be structured along tutor presentations, group-work and review of assigned homework. Furthermore, each student must turn in some written assignments (typically four or six), which contribute towards the final grade, before sitting a written exam.

We have embarked on an effort to analyze the performance of high-risk students [1, 2, 5]. Key demographic characteristics of students (such as age, sex, residence etc), their marks in written assignments and their presence or absence in plenary meetings may constitute the training set for the task of explaining (and predicting) whether a student would eventually pass or fail a specific module. It is important to mention that the great majority of students dropped out after failing to deliver the first one or two written assignments. It is, thus, reasonable to assert that predicting a student's performance can enable a tutor to take early remedial measures by providing more focused coaching, especially in issues such as priority setting and time management.

2.2 Summarizing the technology: decision trees and genetic algorithms

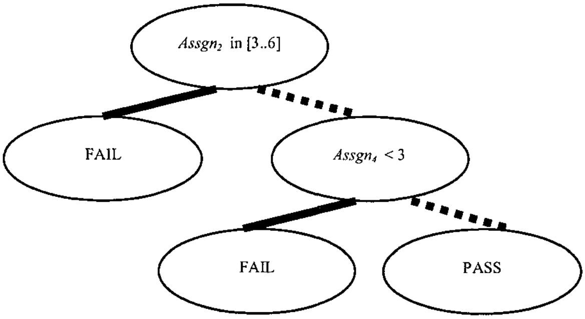

Fig. 1. A sample decision tree

A decision tree [6] for the failure analysis problem could look like the one in Figure 1. In essence, it conveys the information that a mediocre grade at an assignment, turned in at about the middle (in the time-line) of the module (containing 4 assignments altogether), is an indicator of possible failure at the exams, whereas a non-mediocre grade refers the alert to the last assignment. An excerpt of a training set that could have produced the above tree could be the one shown in Table 1.

Table 1. A sample decision tree training set

Assgn$_1$	Assgn$_2$	Assgn$_3$	Assgn$_4$	Exam
...	...	...	...	...
4.6	7.1	3.8	9.1	PASS
9.1	5.1	4.6	3.8	FAIL
7.6	7.1	5.8	6.1	PASS
...	...	...	...	...

Genetic algorithms can directly evolve binary decision trees [7] that explain and/or predict the success/failure patterns of junior undergraduate students. To do so, we evolve populations of trees according to a fitness function that allows for fine-tuning decision tree size vs. accuracy on the training set. At each time-point (in genetic algorithms dialect: *generation*) a certain number of decision trees (*population*) is generated and sorted according to some criterion (*fitness*). Based on that ordering, certain transformations (*genetic operators*) are performed on some members of the population to produce a new population. This is repeated until a predefined number of generations is reached (or no further improvement is detected).

These concepts form the basis of the GATREE system [8], which was built using the GAlib toolkit [9]. A mutation may modify the test attribute at a node or the class label at a leaf. A cross-over may exchange parts between decision trees.

The GATREE fitness function is:

$$fitness(Tree_i) = CorrectClassified_i^2 * \frac{x}{size_i^2 + x}.$$

The first part of the product is the actual number of training instances that a decision tree (a member of a population) classifies correctly. The second part of the product (the size factor) includes a factor x which has to be set to an arbitrary big number. Thus, when the size of the tree is small, the size factor is near one, while it decreases when the tree grows big. This way, the payoff is greater for smaller trees. Of course, this must be exercised with care since we never know whether a target concept can be represented with a decision tree of a specific size.

2.3 Summarizing past findings and setting the context

Initial experimentation [1] consisted of several Machine Learning techniques to predict student performance with reference to the final examination. The WEKA toolkit [10] was used and the key finding, also corroborated by our tutoring experience, is that success in the initial written assignments is a strong indicator of success in the examination. A surprising finding was that demographics were not important.

Follow-up experimentation [2] using the GATREE system [8] initially produced significantly more accurate and shorter decision trees. That stage confirmed the qualitative validity of the original findings (also serving as result replication) and set the context for experimenting with accuracy-size trade offs. That experimentation spanned three academic years, covered the three introductory modules INF10, INF11 and INF2, and validated that genetic induction of decision trees could indeed produce very short and accurate trees that could be used for explaining failures.

We have already documented that drop-out is a significant issue in ODL universities. What is most important, however, is that drop-out usually occurs early in the studies. Failure on a senior year course should simply postpone graduation as the fundamental commitment to studying has been already made. However, failure in a junior course, and for the HOU case, this refers to the INF10, INF11 and INF12 modules, can contribute to a decision to drop out both because the learning investment is not yet large enough to warrant a certain attitude of persistence and because the student may not have had the time to familiarize oneself with the distance learning mode of education (which, given time, allows one to dovetail studying more effectively with other activities).

By regulations, a student who fails a module examination can sit the exam on the following academic year. Such students are only assigned to student groups for examination purposes and the group tutor is responsible for marking their papers only; we thus refer to them as "virtual" students (should they fail their exam for a second year, they must take the module afresh, in which case they are conventionally assigned to a group and cease to be virtual).

Virtual students are not entitled to attending plenary sessions, and to having their assignments graded by the group tutor (as a matter of fact they are not even requested to submit assignments). In practice this regulation may be relaxed by a tutor, who may opt to extend an invitation to attend some plenary sessions to these virtual students usually. Usually, all tutors of a module will either accept or decline to relax the regulation. Of course, there is no focused follow-up of the progress of virtual students, as opposed to the case with typical students.

Any attempt to address these realities involves a political decision that must necessarily take into account the university's administrative regulations.

One step taken by tutors of the INF10 and INF11 modules is to hold a plenary marking session of tutors for each module after an examination, and to discuss variations in individual marking styles based on a predefined assignment of points to exam questions. This is especially important for problems that involve design or prose argumentation. We note that this practice is not widespread within HOU.

A further ad hoc step taken (during the 2003-4 academic year) by the INF11 tutors was to group all virtual students in one group and assign one experienced tutor to that group, as opposed to the usual practice of distributing virtual students across tutors. These students were fully supported by an asynchronous discussion forum and by synchronous virtual classrooms. The tutor did neither hold a physical meeting nor correct any assignments. This was in line with the HOU regulations and, coincidentally, served as a convenient constraint on the "degrees of freedom" of the educational experiment.

We now establish interesting indicators on the effectiveness of these approaches.

3 The experimental environment

We use GATREE for all experiments (even the basic version allows for unlimited experimentation with the *x* parameter in the fitness function, essentially treating *x* as an accuracy-vs.-size bias "knob").

For all experiments we used the default settings for the genetic algorithm operations (cross-over probability at 0.99, mutation probability at 0.01, error rate at 0.95 and replacement rate at 0.25). All experiments were carried out using 10-fold cross-validation, on which all averages are based. Because the data sets are reasonably large, ranging from 500 to 1000 student records, and because 10-fold cross-validation is a widely acceptable testing methodology, we opt to not report standard deviations. The experiments were made with a generations/population:150/150 configuration.

All data refer to the 2003-4 academic year. They do not differentiate between typical and virtual students.

Our methodology is the following: we attempt to use the student data sets to develop success/failure models represented as decision trees. We then use the differences between the models derived when we omit some attributes to reflect on the importance of these attributes. The results are then used to comment on alternative educational policies for dealing with virtual students.

We first try to deal with the issue whether we might be able to obtain an overall (typical and virtual students included) model that deals with explaining (and, ultimately, predicting) exam success, across the three modules that have three distinct policies.

The first experimental session attempted to produce short decision trees that could be used to explain the failure model of students in each module. For this, the *x* knob was set to 1000 (the minimum possible value). For each module, four (4) experimental batches were conducted and the results are shown in Table 3.

Table 2. Results for x=1000, gen/pop:150/150 GATREE decision trees

Data Set	Accuracy (in %)	Size (in nodes)
INF10: Basic	78.20	3
INF10: Basic_T	78.20	3
INF10: Basic_Y	82.58	6
INF10: Basic_TY	82.02	6
INF11: Basic	82.82	5
INF11: Basic_T	82.05	5
INF11: Basic_Y	81.28	6
INF11: Basic_TY	81.54	6
INF12: Basic_T	62.37	6
INF12: Basic_T	63.39	6
INF12: Basic_Y	67.97	6
INF12: Basic_TY	68.81	6

A few words on notation are in order (which apply for all experimental sessions reported in this paper). The *Basic* version of the training set consists of all student records, where the only available attributes are the assignment grades and the class attribute is the *pass/fail* flag. The *Basic_T* version of the training set includes the tutor as an attribute, whereas the *Basic_Y* version includes as an attribute the year of first sitting the exam for that module. The *Basic_TY* version includes both additions. The gen/pop configuration refers to the number of generations and the population size.

The first observation is that the basic model for INF10 simply has a root and two leaves! A slightly larger model, which also tests on the year, is enough to increase sizeably the explanation accuracy.

A casual first observation of the above findings seems to suggest that the tutor attribute is relatively not important (note that we acknowledge that we do not report our results with statistical significance, but we have opted to focus on educated selections of experiments that can demonstrate easily observable trends).

A further observation is that the INF11 module demonstrates a clear "smoothing" of model accuracies across the various versions of its training set. We take this to be a first indication of the success of the INF11 approach to virtual students as it essentially conveys the information that the failure explanation must be traced solely to academic performance (i.e. assignments).

Very short trees may be very concise to communicate but might lack the representational power to detect delicate regularities in the data. We have thus followed-up the experimental results above with increasing *x* to 10000 to allow for larger trees to be generated. However, for space reasons, we will directly jump to the case where this "tweaking" of the *x* knob, was accompanied by larger-scale experimentation in terms of generations and populations as well.

The results are shown in Table 4. (Note that we have dropped the reporting of model sizes as they were very close to the ones reported for the shorter experiments.)

Table 4. Results for gen/pop:300/300 GATREE decision trees

Data Set	Accuracy, x = 1000 (in %)	Accuracy, x = 10000 (in %)
INF10: Basic	78.20	77.42
INF10: Basic_T	78.20	77.30
INF10: Basic_Y	83.60	84.61
INF10: Basic_TY	83.37	83.60
INF11: Basic	82.05	79.74
INF11: Basic_T	81.28	80.26
INF11: Basic_Y	82.31	84.36
INF11: Basic_TY	81.03	83.33
INF12: Basic	62.54	65.08
INF12: Basic_T	63.73	64.07
INF12: Basic_Y	70.51	72.03
INF12: Basic_TY	70.68	73.05

The results are very interesting, to say the least.

Starting from the INF11 module, we see that the short trees are indeed excellent as far as consistency goes. When we go to larger trees, the year attribute creates a performance gap that was not evident before.

This has a two-fold interpretation. On one hand, the larger trees now produced seem to be less well-fitted than the smaller ones (note the accuracy reduction for non-year-inclusive data-sets). This could well be an indication of over-fitting. On the other hand, it suggests that the year attribute has importance; this would concur well with the explanation that students who have failed to pass through the examination filter may be unlikely to have confidence to pursue their studies actively.

Is this finding contradicting the shorter experiments? One needs to examine the results for the other modules to glimpse at the (negative) answer.

First, we observer that for INF12, the year attribute remains a top contributor to the model. For INF10 and INF11 short trees again suggest that the year attribute is less important than for INF12, quite markedly so for INF11, where the year attribute is essentially suppressed. For larger trees, both for INF10 and INF11, the importance of the year attribute seems to rise but at the expense of an overall reduction trend for the *Basic* models. This lends weight to the over-fitting argument but still is plausible, as we said above, since one cannot easily wipe out the *a priori* disadvantage of virtual students.

However, we also note that the increase in accuracy for the INF10 models that use the year attribute is easily seen to be less that the corresponding accuracy for the INF12 models. This observation combined with the observation that the average accuracies for INF10 are also larger than the average accuracies for INF12 may be also interpreted as an indicator that the plenary "marking" session of INF10 helps trim out potential grading inconsistencies. Of course, this may be also a contributor to the underlying quality of the INF11 models, but at the resolution level we are working, we cannot easily confirm or refute the level of this contribution.

Summarising, the importance of the year attribute is only evident for larger trees for the modules that employ the post-exam plenary marking session. Still, that rising importance is clearly less evident than in the INF12 module. Moreover, that evidence is still less proclaimed for the INF11 module that employs a further approach to dealing with virtual students.

4 Conclusions – Focusing on the application domain

We believe that, as of yet, we do not need to experiment with still larger trees, larger populations and more generations, just like we have so argued before [2]. We have observed that large trees give easily rise to the over-fitting phenomenon and that relatively few generations and reasonably small populations could deliver directly usable results. Furthermore, a small accurate model is a very important tool at the hands of a tutor, to assist in the task of continuously monitoring a student's performance with reference to the possibility of passing the final exam. Our setting of parameter x in the accuracy-size trade-off in this paper again confirms this view.

We intend to continue favouring GATREE compared to other software for the particular data analysis tasks, because it incurs a less steep learning curve on the part of a user. However, we have used other software as (simply) another way of replicating the results in the data sets that we have used [2].

We cannot yet answer whether the approach of the INF11 tutors is an approach that would have had replicable educational results in the other modules. The most obvious reason is that exact replication of the above experiments is impossible. Had we wanted to experiment with INF11 approach in INF10, we cannot hope to ever again observe the given set of students and their assignment to groups within modules, as well as the given set of tutors and their assignment to groups. This is one of the reasons that we progressively narrowed down our experiments: we started at only one undergraduate programme, then focused on the most junior and well-subscribed modules, then singled out the two ones that demonstrated one difference only at the policy level.

Having taken these careful steps, we believe that, when one focuses on limiting drop-out, the presented analysis suggest that the effective smoothing-out of the year-and-tutor factors in the success-failure model should benefit from a purely educational decision: by assigning an experienced tutor to directly deal with virtual students. The other alternative, which is to train all tutors to be more active in discussion fora and more proficient in virtual classroom techniques, may be a grand goal with far-reaching benefits, but could demand a substantial mentality shift of the tutors and substantial vocational training resources, entailing significant political decisions.

Are the conclusions and the advice too strong? We think not, taking into account that differences are in the order of several percentage points, with consistent standard deviations, whereas individual performances are in the order of 70% (and not, for example, 95%, where a few percentage points might be less important). Moreover, the validity of the results is strengthened by the fact that we have conducted the experiment in the most controlled of environments. An obvious extension of this work is to try to see whether differences are more or less pronounced in less controlled environments (for example, in senior year modules, where the student population is drawn from more than one academic admission stage).

This observation then sets the context for the wider goal of this research. We investigate the building an "early warning and reaction system" for students with "weak" performance. This research has also operational and political aspects, besides the obvious technical ones.

From both an operational and technical viewpoint, one must set a scheme to validate the performance of a model based on subsequent years' statistics and not simply on cross-validation testing. It is important to note that the approach is self-contained in the sense that it can be readily applied to data available at the university registry.

Deploying this scheme as an organization-wide process would also lend support to our preference for short models. We believe that a small accurate model is a very important tool at the hands of a tutor, to assist in the task of continuously monitoring a student's performance with reference to the possibility of passing the final exam. A small model is easier to communicate among peers, easier to compare with competing ones and can have wider applicability.

Political issues are much subtler, of course, and we have already pin-pointed one.

A sensitive point is that it would be unwise to simply consider the higher or lower overall *absolute* accuracy rate of (any) model in one module as an indicator of success of an approach, at least at this early stage of the research. It is for this reason that in the experiments described above we never pit one module's accuracy against another module's accuracy; besides referring to different student populations (including differences in population sizes), a module also refers to different tutors and to another scientific field.

We believe that such an approach would distract us from our goal. What is more important, we claim, is to detect and observe the trends within the module itself and try to understand what actions need to be taken at the module level.

In [2] we argued that using a system like GATREE and an approach like the one documented above to produce and operationally use success/failure models raises the fundamental question of whether we measure the performance of actors (students or tutors) or the performance of the system at large (the ODL system implemented in HOU). We also conjectured that it is the latter alternative that has the most potential from an educational point of view.

Given that we have successfully used raw data (student records) to *a posteriori* justify an educational policy, as opposed to compute an individual student model *per se*, we believe that this conjecture is now better founded.

References

1. Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using Machine Learning techniques. Applied Artificial Intelligence, 18:5, 411-426.
2. Kalles, D., & C. Pierrakeas (2005). Analyzing student performance in distance learning with genetic algorithms and decision trees (accepted for publication in the journal: Applied Artificial Intelligence).
3. Keegan, D. (1993). Theoretical Principles of Distance Education. Routledge, London.
4. Open Leaning (2004). Special issues on "Student retention in open and distance learning". 19:1, http://www.tandf.co.uk/journals/titles/02680513.asp.
5. Xenos, M., Pierrakeas, Ch. & 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.
6. Mitchell, T. (1997). Machine Learning. McGraw Hill.
7. Koza, J.R. (1991). Concept formation and decision tree induction using the genetic programming paradigm. Parallel problem solving from nature. Berlin: Springer Verlag.
8. Papagelis, A., & Kalles, D. (2001). Breeding decision trees using evolutionary techniques. In Proceedings of the International Conference on Machine Learning, Williamstown, Massachusetts.
9. Wall, M. (1996). GAlib: A C++ Library of Genetic Algorithm Components. M.I.T. http://lancet.mit.edu/ga/.
10. Witten, I., & Frank, E. (2000). Data mining: practical machine learning tools and techniques with Java implementations. San Mateo, CA: Morgan Kaufmann