Th. Hadzilacos, D. Kalles, C. Pierrakeas and M. Xenos. “On Small Data Sets revealing Big Differences”, *Panhellenic Conference on Artificial Intelligence*, Heraklion, Greece, **2006**.

On Small Data Sets revealing Big Differences

Thanasis Hadzilacos[1,2], Dimitris Kalles[1], Christos Pierrakeas[1] and Michalis Xenos[1]

[1] Hellenic Open University, Patras, Greece
{thh, kalles, pierrakeas, xenos}@eap.gr
[2] Computer Technology Institute, Patras, Greece

Abstract. We use decision trees and genetic algorithms to analyze the academic performance of students throughout an academic year at a distance learning university. Based on the accuracy of the generated rules, and on cross-examinations of various groups of the same student population, we surprisingly observe that students’ performance is clustered around tutors.

1 Introduction

Small data sets are usually suspect when used in a machine learning context. We present an application in a complex educational environment where a collection of small data sets can reveal surprisingly sensitive information about the processes that generate the data.

In the Hellenic Open University (HOU) we attempt to analyze whether tutoring practices have an effect on student performance. Using decision trees and genetic algorithms, we report significant differences in tutoring practices and we reflect on the implications and on the potential of these findings.

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.

Initial experimentation at HOU [1] consisted of using several machine learning techniques to predict student performance with reference to the final examination. The WEKA toolkit [2] was used and the key finding, also corroborated by 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.

We then followed-up with experimentation [3] using the GATREE system [4], which 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.

A decision tree like the one in Figure 1 (similar to the ones actually produced by GATREE) tells us that a mediocre grade at an assignment, turned in at about the middle (in the time-line) of the module, 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 such a tree is shown in Table 1.

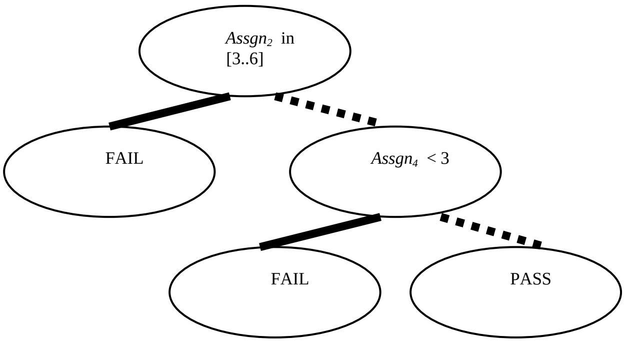

Figure 1. A sample decision tree.

Table 1. A sample decision tree training set

Assgn$_1$	**Assgn$_2$**	**Assgn$_3$**	**Assgn$_4$**	**Exa m**
…	…	…	…	…
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
…	…	…	…	…

2 The experimental environment

We use the student data sets to develop success/failure models represented as decision trees. We then calculate the differences between the models derived by data sets from different tutors to reflect on alternative educational policies.

The measurement is based on partitioned data sets. We have partitioned student populations into groups, according to how tutors are assigned to groups. This partitioning allows us to examine whether the grading practices of one tutor apply predictably well to a group supervised by another tutor at the same module. Table 2 shows how are these results are calculated.

Table 2. A template for tabulating cross-testing results

Data Set	**D$_0$**	**D$_1$**	**D$_2$**	**…**	**D$_n$**
D$_0$	…	…	…	…	…

D_1	…	CV_1	…	…	…
D_2	…	$V_{2,1}$	CV_2	…	…
…	…	…	…	…	…
D_n	…	…	…	…	…

A few words on notation are in order. D_i refers to the student group supervised by tutor *i*. D_0 refers to all students of a module. CV_i refers to the 10-fold cross-validation accuracy reported on data set D_i. $V_{i,j}$ refers to the validation accuracy reported on data set D_j, using the model of data set D_i.

We experimented with one senior (INF31) and two introductory (INF10, INF11) modules, chosen because one or more of the authors had a direct experience in tutoring a student group therein. INF31 data refer to the 2004-5 academic year. INF10 and INF11 data refer to the 2003-4 academic year (newer data has not been collected yet).

In INF31, six student groups were identified, with about 160 students in total. INF10 and INF11 have more student groups (over 30 and over 15 respectively) and stronger student populations (about 1000 and 500 respectively). The smallest groups had just over 10 students. The largest ones had just over 30 students.

We use GATREE for all experiments, with the default settings for the genetic algorithm operations (cross-over probability at 0.99 and mutation probability at 0.01) and set the bias to prefer short trees. Now, each table like Table 2 can be summarized by the average value of its cells; the initial results are shown in Table 3.

Table 3. INF31 accuracy results for decision trees

Data Set	INF31	INF10	INF11
Accuracy	92.05	83.60	82.31

Do we infer that senior course (INF31) tutors demonstrated a tighter homogeneity than their junior course (INF10, INF11) colleagues? Or, are the above findings the results of processes inherent (but, not yet identified) in the very different students populations (junior vs. senior)?

Herein lurks the danger of using statistics (even, sophisticated) without adequate domain knowledge. To further analyze the above data we went a step further. We noted that the overall exam success rate in INF31 is nearly 100%, whereas in the other two modules (before factoring in the students who drop out) the success rate is substantially below 50%. That sheds new light into the above findings. Indeed, it would be surprising if many differences could be spotted between groups who are overwhelmingly successful!

Now, note that the term “exam” actually aggregates two sessions; a student sits the second exam if the first one is unsuccessful. In this light, we observed that the near 100% rate of INF31 was due to the second exam, whereas the success rate for the second exam in INF10 and INF11 was very small (compared to the overall rate).

Now, the result for the first INF31 exam was $\boxed{62.31}$ (use Table 3 as a benchmark).

The findings are telling (and, incidentally, they suggest that any reference to standard deviations is superfluous). What initially appeared as homogeneity among tutors now turns out to be wide differences.

One can now reframe the question of homogeneity of tutoring practices, albeit in the opposite direction. However, we believe that the crux of the matter must be the identification of the gap. Bridging the gap is, as the cautious reader might suspect, more about human aspects than about technology.

3 Conclusions

The single most important contribution of this paper is the identification of a procedure for analyzing the level of homogeneity displayed by a group of tutors who are, theoretically, expected to smooth out differences that hinder their students' learning efforts. We have used an advanced combination of standard AI methods, in order to obtain information necessary for the reflection on educational policies.

There are some clear steps that can be taken to address differences. Some tutoring groups in HOU have invested in plenary virtual classes, whereas other groups are rotating the grading of exam papers. Either approach could be a legitimate candidate.

But, the purpose of this work is not to recommend a particular approach or to pit tutors' practices against each other; rather it is to develop a methodology whereby tutoring differences are raised and analyzed with respect to their importance. The role of the tutor in reflecting on one's own practices is as important as ever and, as this paper's experimental methodology has demonstrated, even small data sets can yield profound insight.

References

[1] Kotsiantis, S., Pierrakeas, C., & P. Pintelas (2004). Predicting students' performance in distance learning using Machine Learning techniques. *Applied Artificial Intelligence*, 18:5, 411-426.

[2] Witten, I., & E. Frank (2000). *Data mining: practical machine learning tools and techniques with Java implementations*. San Mateo, CA: Morgan Kaufmann.

[3] Kalles, D., & C. Pierrakeas (2006). Analyzing student performance in distance learning with genetic algorithms and decision trees (accepted for publication in: *Applied Artificial Intelligence*).

[4] Papagelis, A., & D. Kalles (2001). Breeding decision trees using evolutionary techniques. *Proceedings of the International Conference on Machine Learning,* Williamstown, Massachusetts.