On Experimenting with Data Mining in Education

T. Gnardellis*[1] , B. Boutsinas*[1,2]
1: IS & AI Lab, Dept. of Computer Eng. & Informatics, UOP
265 00, Rio, Patras.
2: Dept. of Business Administration, UOP
265 00, Rio, Patras.
E-mail: gnardell@ceid.upatras.gr, vutsinas@bma.upatras.gr

* This work is supported by Enterprise Programmes for Research and Technology II, General Secretariat for Research & Technology, Hellenic Ministry of Development.

Abstract
The main characteristic of formal education is teaching within an administrative framework. Therefore a lot of teaching and administrative activities must be carried out during different education processes, from library services evaluation to building cognitive student models. Efficient development of such education processes usually rest either on intuition or on performing data analysis techniques in order to reveal key concepts and their relationships. Following the latter approach in this paper, we investigate the application of data mining techniques within the education framework. Data mining techniques are used to automatically extract knowledge, in the form of relationships and patterns, from large databases. In this paper, we do not aim at an exhausted evaluation of all application areas within education. Instead, we rather aim at interesting the reader in such an idea by presenting some practical cases.
Keywords: data mining, education software, student modeling

Introduction

Data Mining is a process through which one can extract valuable knowledge from a large database. The necessity for the development of data mining evolved due to the immense and quick growth of the volume of stored corporate data. Ordinary querying methods could no longer produce results showing hidden patterns in such vast amounts of data. Using advanced methods derived from artificial intelligence, pattern recognition and statistics, data mining can construct a comprehensively descriptive model on input data. The data model can be produced in various forms and serves the purpose of describing and predicting behaviour of the data object.

On the other hand, teaching and administrative activities that must be carried out during different education processes, usually rest on performing data analysis techniques in order to reveal key concepts and their relationships. Data mining techniques are best suited for extracting from data such key concepts along with their relationships. In this paper, we investigate the application of data mining techniques within the education framework, aiming at interesting the reader in such an idea, presenting some practical cases. The rest of the paper is organized as follows. The data mining process is described briefly in next subsection. Application areas of data mining, concerning education processes is described in the following section. The paper ends with a case study and some concluding remarks.

The Data Mining process

The process of knowledge discovery involves several steps [12] one of which is applying the data mining technique that we have chosen, according to the nature of the data and the kind of knowledge we would like to extract. These steps are shown in the following diagram.

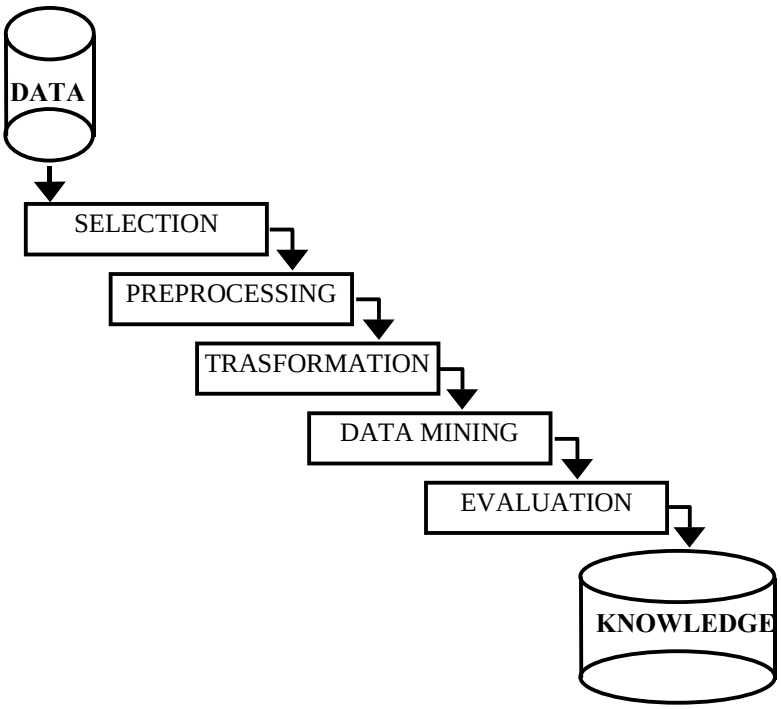

Fig. 1: The Knowledge Discovery Process

The first three steps in the above diagram involve data handling. The data are rarely stored in a form suitable for data mining. They have to be selected from various sources and then combined into one dataset, upon which various transformations may be applied. For example we may have numerical values which are generally better to be transformed into categorical values, thus producing more comprehensive results. These steps are highly important for the overall success of the process, perhaps equally important to the actual data mining step. The data mining step is the application of the actual algorithm on the pre-processed and probably transformed dataset. The analyst that performs the process has to carefully examine which algorithm to apply and how to specify its parameters, factors that can dramatically affect the quality of the results.

Pre-processing steps have a strict technical context, as well as the specification of the critical parameters of data mining algorithms. Due to them, there is an impression that data mining is not an automatic process, in most of cases. Some users may argue that data mining is mostly a model driven exercise, since the researcher needs to have a clear understanding of the domain, the semantics of the dataset (in order to perform data preprocessing and transformation) and finally to have a clear objective for the data mining process in order to drive the knowledge extraction. Such arguments do not hold true if we refer to an integrated data mining system (eg. Clementine™, Keppler™, etc), with an appropriate user interface.

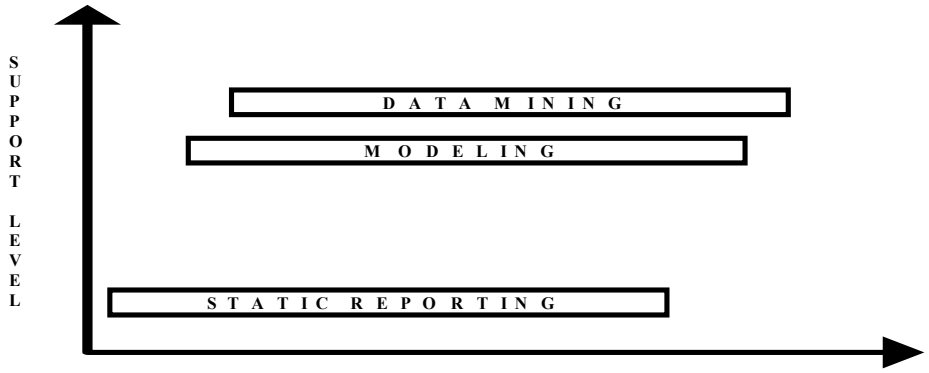

Fig. 2: User support in data analysis techniques

After the data mining step is completed, the output may be subjected to a filtering step during which it is evaluated. Since data mining is an automated, data-driven process, it may produce results that could appear evident or even naïve. Therefore, it is compulsory to filter out only the knowledge that is truly useful and understandable. In some cases the process may be repeated by transforming the data in a different way and altering the parameters specified.
Data mining is a revolutionary technique in terms of the level of support it provides to the user. As shown in the diagram of Fig. 2, it exceeds other popular data analysis techniques such as modeling. When we refer to the level of support in the diagram we mean the degree to which each process is automated. Static reporting is the process where the user foresees the patterns among the data and issues pre-designed queries to get the actual report. Modeling is performed using statistical software such as SPSS. It is more advanced than static reporting but is still user-driven since the user must make a work hypothesis upon which the system operates. Data mining on the other hand is completely data-driven. The user has to make no assumptions in regard to the results that he wants to receive. He simply feeds the data to the process and the results are produced depending solely on the form and values of the input data.
Data mining techniques are divided into various categories. The main categories are *classification*, *clustering* and *association*. These kinds of techniques are preferred because they generally produce the best results.
Classification is a process that aims at defining a model used for classifying data cases into one class of a set of predefined classes. Each unclassified data case can then be classified by the model, according to its values in certain fields. Classification algorithms are applied on "training" data sets that contain cases that have already been classified. The model they produce can be in several forms such as trees or sets of classification rules.
Clustering also divides input data into groups. The main difference is that there are no predefined classes in this case. Clustering algorithms examined the data with the purpose of finding similarities among the different cases justifying their grouping into *clusters*.

Finally, association aims at discovering dependencies among the values of different attributes. It produces rules that have the form $X \Rightarrow Y$, meaning that where there is X there is probably Y as well.

There have been found uses for data mining in a significant number of business and research activities. Marketing is an area that has been greatly assisted by data mining. Through the data

mining process, customer-related data can be "mined upon" in order to discover patterns in customer behavior. By taking advantage of that knowledge, marketing policies can be redesigned for increased efficiency.
In retail sales, data mining can be used for "market basket" analysis. By applying an association algorithm on retail data we can see which products are usually purchased together. We can use that information for advertising purposes or for better design of market shelves.
Other areas have benefited from data mining such as manufacturing, finance and medicine. In general, we can apply data mining wherever data can be shaped or interpreted as various instances of one concept, enabling us to discover knowledge for the behavior of that concept.

Application Areas

In this paper, we do not aim at an exhausted evaluation of all application areas of data mining, concerning education processes. Instead, we rather concentrate on some case studies, in order to interest the reader in the idea of applying data mining techniques within the field of education.
More specifically we investigate the use of data mining in three practical cases:

- Evaluation of Education Software
- Modeling in Intelligent Tutoring Systems (ITSs)
- Evaluation of Syllabus

Evaluation of Education Software

Evaluation of Education Software is a very important administrative activity within the education framework of the "information age" period. This is because such evaluations influence interest in research and application in the field of Education Software. But evaluations are most important because influence choices to adopt a particular software, therefore influence, indirectly, teaching activities and, hence, how students learn.

Following Tyler's suggestion [1], since 1942, there are six main purposes of educational evaluation studies:

1. to assess the effectiveness of a program periodically so that needs for improvement could be identified
2. to verify whether hypotheses upon which a program was based had proved tenable
3. to generate information for better guiding people in a program
4. to assure those involved with a program that it was worthwhile
5. to enhance public relations
6. to describe more clearly the directions in which a program was moving.

Although the above suggestions relate to an educational program, i.e. a curriculum, and, today, such view is somewhat different for Intelligent Tutoring Systems, (or Intelligent Agents, or Microworlds and so forth), evaluations of educational software retain this broad view in effect [3].
To achieve most of the above purposes, one needs to gather data to describe the software and then to examine those data. Gathered data should describe attributes of the software at each of its levels. Usually, one collects a file of every act that a student performs in the interface during a session with the education software. Of course, data should be easy to collect, be reliable and appropriate. Processes for gathering data is beyond the scope of this paper. In general, gathered data should be of multiple kinds from multiple sources that can be used collectively to sharpen portrayals of a context, an input, a process, an outcome, or links between these [2]. Obviously, such gathering processes leads to the construction of a large database with a great number of fields.
Of course, there are traditional statistical analysis methods, sometimes dedicated to educational evaluation studies (eg [4]), that can be applied to gathered data. However, data

mining techniques (classification and/or associations) are best suited for extracting knowledge for databases of the kind imposed by such gathering processes.

Modeling in Intelligent Tutoring Systems

The capability to create a model of the student [7] is an important attribute of most tutoring systems, since such systems need to have a very robust model of the user if they are to carry out actions that depend upon a knowledge of personal preferences or interests. Usually, such knowledge is empirical and it is derived from the study of typical student behaviours and the student's responses to tests and problems. The key idea is that this knowledge constitutes a knowledge base, that is embodied in the ITS. Then, based in this knowledge base, the ITS attempts to single out matters that the student may not have well understood and which therefore need further attention. The aim of proposed, in the literature, ITSs, is to identify a domain-independent knowledge representation schema that accommodates apparently different approaches. A lot of representation formalisms have been proposed, from propositional (eg. [8]) to non-monotonic ones (eg. [9]). Depending on the nature of the student's knowledge, different student models have been proposed, with most important the overlay and the buggy model [10].

The overlay model is based on the assumption that the student's knowledge partially covers the expert's knowledge, which can be fragmented into small units. The student's knowledge can be defined as a subset of these units. Historical data are kept about the skills the student has mastered. These data are then analysed in order to indicate certain discontinuities that usually constitute a complete rule based architecture. For this analysis, three sources of evidence are used: comparison of the student's and the expert's behaviour, dependencies among skills (mastered units), interacting with the student by test cases. It is obvious, how clustering and/or classification tasks can be used in the first case and association rules in the second.

For instance, for the first case, consider that mastering of units of the expert's knowledge is represented by certain attributes while mastering of units of the student's knowledge by others. Then, using the expert's knowledge attributes as class attributes, a classification task could extract rules that define which units of the student's knowledge must be mastered in order for certain expert's knowledge units to be covered. Moreover, for the second case, it is trivial how association rules can define dependencies among such attributes, namely among mastered units.

Buggy model considers the student's knowledge as a perturbation of the expert's knowledge. Therefore, a student model consists of an expert model and a list of predefined misconceptions and missing conceptions. Obviously, considering such misconceptions, missing conceptions and units of an expert's knowledge as attributes, data mining tasks can be applied, similarly to the overlay model. Moreover, a classification task can be used in building the library of bugs. The latter is the main limitation of the Buggy model, since in order for such a library to be built, one must either study the literature or use a learning theory to predict bugs or analyse the student's problem solving behaviour [11]. All of the previous are quite difficult tasks.

Another important concept that can be studied, in the context of an ITS, is the modelling of teaching strategies. It is showed [5] that a critical feature for an ITS to be acceptable to a school teacher, is that it should have flexible teaching behaviour and that teaching behaviour should be easily configurable. Without a teaching strategy, an ITS would become more of an intelligent help system, providing intervention only on student request. Note, also, that some exceptional ITSs (eg [6]) have multiple teaching strategies. Most of these systems include an authoring element so that teachers and courseware designers can control the teaching behaviour. Usually, such authoring element inspired by work in the area of expert systems. Domain knowledge is, usually, acquired from experts that are asked to present their attitudes

to formal or informal methods of teaching. Obviously, a clustering task could be effectively used in extracting patterns of major attitudes along with prerequisites for their applicability.
Also, the student's Plan recognition capabilities are valuable for intelligent tutoring, for making assumptions about the student's plans and inferring unobservable properties of the student. More specifically, Plan recognition interprets points (impasse points) where the student encounters some difficulty completing tasks. Usually, Plan recognition techniques are rigid [13], since they assume that the student is following a known plan step by step. Therefore, they have difficulty interpreting deviations from the plan. A classification procedure, based on data gathered during student monitoring, could identify the student's plan from the student's actions. Such a classification task, capable of classifying unseen examples with great classification accuracy, could cope with deviations from standard plans.

Evaluation of Syllabus

Evaluation of Syllabus is, of course, a very important administrative activity. It is, usually, related to the evaluation of whole educational projects. To our knowledge, there is not any formal methodology. Empirical studies must be based on defining certain measures and must define measuring processes. Certainly, data mining tasks can be used toward this direction.
Instead of formally describing the problem, we present, in the next section, an extensive use of data mining tasks in a certain case study, aiming at inspiring the reader the previous ideas.

Case Study

In the following case study, we will demonstrate how one can efficiently go through all the steps of the knowledge discovery process, in order to receive quality results from data mining procedures. The input data are student records from the Ovrya High School, Achaia, Greece recorded during the school periods 1998-1999 and 1999-2000. The data were given in two Microsoft Access™ databases, one for each school period.
Since we did not have various data sources to choose from, there was no need for a selection step in the process. There was, however, a great deal of tasks to perform in the preprocessing and transformation steps in order to bring the information included in the data in the best possible form for data mining.
The data were in identical structures in both databases so the procedure was repeated for both databases up to the point where we unified data from the two school periods into one dataset, as described later in this section. Initially, there were separate tables for student data, course data and trimester grades. These tables had to be combined into one dataset using multiple queries. That brought us to a point where we had a dataset with records including all three trimester grades per course, per student. That is to say, there was one record for student A's trimester grades in course A, another for the same student's grades in course B etc. Afterwards, we calculated an extra field containing the final grade.
However, that was hardly the dataset form required to conduct data mining procedures. Dependencies and associations in data that data mining is used to extract lie among different fields of the dataset, not among different rows. So what we needed to do was, in a way, to sum up the information in that dataset into another dataset where each record would contain all the final grades for each student. Since that required complicated data manipulation scripts, we transferred our datasets (one for every school period) to an Oracle™ database where data manipulation is quite easier. After creating the new summarized datasets, we had to categorize all the grade attributes, since these were obviously numerical ones.
Finally, we combined the two datasets into one and eliminated null records and fields. The null fields were courses that although recorded in the original courses table, it turned out there were no grades for them. In the end, we had one summarized and categorized dataset, containing 428 records free of null values, ready to be mined upon.

Applying the data mining algorithms on our dataset was a relatively simple process. The fields that interested us were obviously the ones related with the grades of each student. We also conducted a few runs to find dependencies among the sex of the students and their performance in various courses. We will now present the results of each algorithm on the data after having fulfilled the evaluation step. The results presented are a minor selection of the full amount of the data mining results, and their purpose is to demonstrate what kind of knowledge we are looking for out of this process.

We applied the classification algorithm C4.5 to mine classification rules from the data. C4.5 firstly mines a decision tree by dividing the record set in each node according to an attribute's values. The rules are extracted by considering each path of the tree as a classification rule.

Several runs were performed for classification since each time one field has to be the class field, the field, that is, whose values determine the class to which each record belongs. Thus, we can examine what kind of students do well in History, for example, by selecting History as a class field and other courses as fields on whose values the rules will be based. Some of the results where more or less expected, easy for one to assume. The following classification rules that were mined can serve as an example.

If RELIGIOUS MATTERS=EXCELLENT and
MODERN GREEK=EXCELLENT
Then HISTORY=EXCELLENT
(70/8)

If RELIGIOUS MATTERS=FAIL and
COMPOSITION=FAIL and
ANCIENTGREEK-TRANSLATION=FAIL and
MODERN GREEK=FAIL
Then HISTORY=FAIL
(16/1)

The first rule states that 62 out of the 70 students who received excellent grades in Religious Matters and Modern Greek did the same in History. The second one states an opposite situation, that 15 out of the 16 students who failed in Religious Matters, Composition, Ancient Greek – Translation and Modern Greek, failed in History as well.

Those rules represent knowledge that is easy to understand and one could rush to say that they be discarded as useless because they do not imply something new. However, these rules gives us the chance to verify with real data what we previously knew only as a vague assumption. The next two rules might seem more useful to the sceptic eye.

If RELIGIOUS MATTERS=EXCELLENT and
MODERN GREEK=PASS
Then HISTORY=VERY GOOD
(3/0)

If MODERN GREEK=VERY GOOD and
ANCIENTGREEK=FAIL and
ANCIENTGREEK-TRANSLATION=VERY GOOD
Then COMPOSITION=VERY GOOD
(26/4)

The first rule states that all three students who were excellent in Religious Matters and merely passed Modern Greek were very good in History. That is something that could be explained perhaps if we think about the nature of the courses and how these are being taught. The second rule states that students who were very good in Modern Greek, Ancient Greek – Translation

and failed in regular Ancient Greek were very good in Composition. That clearly demonstrates that students who are very good in Modern Greek courses can face severe difficulties in Ancient Greek.

The clustering algorithm k-modes was applied so as to discover the different clusters that students can be grouped to. Kmodes divides a record set into clusters by specifying the *center* record for each cluster. Its methodology is based on minimizing the average distance of a cluster's records to its center.

We instructed the algorithm to divide students into four clusters. The algorithm returned the «center» for each cluster, that is the average description of the students who belong to that cluster. Here are the four centers that k-modes returned.

	1st Cluster	2nd Cluster	3rd Cluster	4th Cluster
Religious Matters	PASS	EXCELLENT	PASS	VERY GOOD
History	PASS	EXCELLENT	FAIL	VERY GOOD
English	VERY GOOD	EXCELLENT	PASS	EXCELLENT
French	PASS	EXCELLENT	FAIL	VERY GOOD
Mathematics	PASS	EXCELLENT	FAIL	VERY GOOD
Gymnastics	PASS	PASS	PASS	EXCELLENT
Ancient Greek – Translation	PASS	EXCELLENT	FAIL	VERY GOOD
Ancient Greek	PASS	EXCELLENT	FAIL	VERY GOOD
Modern Greek	PASS	EXCELLENT	FAIL	VERY GOOD
Composition	PASS	EXCELLENT	FAIL	VERY GOOD
Music	PASS	PASS	FAIL	EXCELLENT
Artistic Matters	PASS	PASS	FAIL	EXCELLENT
Informatics	FAIL	EXCELLENT	FAIL	VERY GOOD

As was instructed, the algorithm divided the students into four clusters. Those who failed in almost every course, those with passing grades, very good grades and excellent grades. But it is interesting to see that excellent students merery passed gymnastics, artistic matters and music that require other talents except intellectuality and being good at studying. On the other hand, students who are characterised as very good in almost all other courses, were excellent in these three.

Apriori tries to find the association among different values of the attributes by evaluating the number of appearances of the various combinations. The algorithm resulted with tremendously numerous rules, associating grades in some courses with grades in others. The majority of these rules associated good grades from various courses or failing grades or passing grades and so on, verifying our conclusions from the previous two techniques. For example:

History = EXCELLENT, Informatics = EXCELLENT and Relig. M. =EXCELLENT

English = PASS, Artistic M. = PASS and History = FAIL

There were, however, in this case as well the «interesting» exceptions, such as:

Mathematics = FAIL, Gymnastics = EXCELLENT and French = PASS

All of the above results were taken by using our own implementations of the corresponding algorithms, which were developed under the project *"Diogenis"* sponsored by the General Secretariat for Research & Technology, Hellenic Ministry of Development. We need to stress once again that these results were not extracted based on any kind of user intervention. The

outcome of this process was purely data-driven, and that is what could be referred to as the beauty of data mining. It was demonstrated that it offers the highest level of support to the user.

Conclusion

We presented some practical cases of applying data mining techniques within the education framework. It seems that in whatever teaching or administrative activity, where a data analysis process is needed, data mining techniques can be used instead. Each application area that was previously mentioned constitutes a whole new research domain. In this paper, we aim at interesting the reader and not at describing, formally, the problems. We plan to further investigate these application areas in a later correspondence.

Acknowledgement

The authors wish to thank F. Stavridis for providing the data as well as the reviewers for the useful comments.

References

[1] Tyler R.W., "General statement on evaluation", Journal of Educational Research, 35, 1942, pp.492-501

[2] Cook T.D. and Campbell D.T., "Quasi-experimentation. Design and analysis issues for field settings", IL.: Rand McNally, Chicago, 1979

[3] Winne P.H., "A Landscape of Issues in Evaluating Adaptive Learning Systems", Journal of AI in Education, 4(4), 1993, pp.309-332

[4] Evertson C.M. and Green J.L., "Observation as inquiry and method", in "Handbook of research on teaching" M.C. Wittrock (Ed.), NY: Macmillan, 1986, pp.162-213

[5] Major N.P., "Teachers and Intelligent Tutoring Systems", Proceedings of 7th International PEG Conference, Heriot-Watt University, 1993, pp.91-117

[6] Spensley F., Elsom-Cook M., Byerley P., Brooks P., Federici M. and Searoni C., "Using multiple teaching strategies in an ITS", in "ITSs: At the crossroads of AI and Education" C. Frasson & G. Gauthier (Eds.), NJ: Ablex, 1990

[7] Self J.A., "Student models-What use are they?", in :Artificial Intelligence Tools in Education" P. Ercoli & R. Lewis (Eds.), Amsterdam: North-Holland, 1988

[8] Mizoguchi R. and Ikeda M., "A generic framework for ITS and its evaluation", in "Advanced research on computers in education" R. Lewis & S. Otsuki (Eds.), Amsterdam: North-Holland, 1991

[9] Giangrandi P. and Tasso C., "Truth Maintenance Techniques for Modelling Student's Behaviour", Journal of AI in Education, 6(2/3), 1995, pp.153-202

[10] Wenger E., "Artificial intelligence and tutoring systems", Morgan Kaufmann Publishers, 1987

[11] VanLehn, "Student modeling", in "Foundations of Intelligent Tutoring Systems" M.C. Polson & J.J. Richardson (Eds.), 1988, pp.55-78

[12] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press, 1996.

[13] Hill W.R. and Johnson W.L., "Situated Plan Attribution", Journal of AI in Education, 6(1), 1995, pp.35-66