Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study

Mykola Pechenizkiy[1], Toon Calders[1], Ekaterina Vasilyeva[1,2], and Paul De Bra[1]
{t.calders, m.pechenizkiy, e.vasilyeva}@tue.nl, debra@win.tue.nl
[1] Department of Computer Science, Eindhoven University of Technology, the Netherlands
[2] Department of Computer Science and Information Systems, University of Jyväskylä, Finland

Abstract. In this paper we describe an educational data mining (EDM) case study based on the data collected during the online assessment of students who were able to immediately receive tailored and elaborated feedback (EF) after answering each of the questions in the test. Our main interest as domain experts (i.e. educators) is in studying (by employing any kind of analysis) how well the questions in the test and the corresponding EF were designed or tailored towards the individual needs of the students. The case study itself is aimed at showing that even with a modest size dataset and well-defined problems it is still rather hard to obtain meaningful and truly insightful results with a set of traditional data mining (DM) approaches and techniques including clustering, classification and association analysis.

1 Introduction

In this paper we present a fresh small-scale EDM case study aimed at demonstrating the potential and the limitations of the traditional DM approaches and techniques [5], focusing on the problems of data redundancy and inherited correlation between many attributes that complicate the discovery of truly interesting patterns in the data.

The data in our case study comes from a real online exam (partial exam for the Human-Computer Interaction course) taken by 73 students. This exam has been organized in the form of a multiple-choice test aimed at demonstrating to the students (and teachers) what they have learnt or what (common) misconceptions they might have, and to provide yet another possibility and means for student to fill the gaps in their knowledge (or “patch” the possible misconceptions). Hence, elaborated feedback (EF) has been designed for each of the questions in the multiple-choice test.

The design of appropriate feedback is a critical issue in the development of online assessments within Web-Based Learning Systems (WBLSs). In our recent works we demonstrated the possibilities (and benefits) of tailoring the feedback that is presented to a student as a result of his/her response to questions of an online test, taking into account the individual learning styles (LSs), certitude in a response and correctness of this response [3,4]. Here, our goal is to report upon our experiences of applying different DM techniques for assessing how well the questions in the test and the corresponding EF were designed and tailored towards the students.

2 Online Assessment and Feedback Adaptation

Online assessment becomes an important component of modern education. Nowadays it is used not only in e-learning, but also within blended learning, as part of the learning process. Online assessment is utilized both for self-evaluation and for "real" exams as it tends to complement or even replace traditional methods of evaluation of the student's performance.

Feedback is usually a significant part of the assessment as students need to be informed about the results of their (current and/or overall) performance. The great variety of feedback functions and types that current system can actually support make the authoring and design of the feedback in e-learning rather complicated. In spite of the diverse interest in educational research dealing with feedback, the methods and guidelines for designing and implementing feedback in educational practice remain scarce so far [2]. This especially applies to the design of feedback in WBLSs. An important issue is that different types of feedback can have a different effect (positive or negative) on the learning and interaction processes [1].

Our studies demonstrated that knowledge of the response certitude (specifying the student's certainty of the correctness of the answer) together with response correctness helps in understanding the learning behavior and allows for determining what kind of feedback is more preferable and more effective for the students. EF may significantly improve the performance of students within the online tests [3]. We demonstrated also the potential of tailoring feedback towards individual learning styles [4].

3 Data collection and preparation

We have studied different aspects of feedback tailoring during a series of experiments in the form of eight online multiple-choice tests in the *Moodle* learning system organized as an integral part of courses (with traditional in-class lectures and instructions) at the Eindhoven University of Technology, the Netherlands during the academic year 2007-2008.

In this data-mining study we focused on the most recent online assessment (partial exam) of 73 students in the Human-Computer Interaction (HCI) course that was organized in March 2008. In some of the earlier assessments we also used feedback adaptation strategies based on the student's response correctness and certitude, and learning style.

The online test consisted of 15 multiple-choice questions. The questions were aimed at assessing the knowledge of the concepts and the development of the necessary skills (e.g., understanding the basic usability rules and problems such as consistency, mapping (between interface and real world), response time problem, etc.). For each answer students had to provide their certitude (affecting the grade) and had a possibility to request and examine the EF that could potentially help to answer related questions better.

For every student and for each question in the test we collected all the possible information, including correctness, certitude, grade (determined by correctness and certitude), time spent for answering the question, whether feedback was requested on not, and which feedback was shown directly (if any), was recommended with which strength, and finally which one(s) were actually examined (including time spent for examining each type of feedback in seconds). Before passing the actual tests the students were asked (a few days before) to answer to the learning style questionnaire (that was not compulsory); 90% of students filled the questionnaire. The collected data was transformed into a transactional multi-relational presentation with different views for the corresponding DM tasks.

Further details regarding the organization of the test (including an illustrative example of the questions and the EF) and the data preprocessing are made available in an appendix we placed online at http://www.win.tue.nl/~mpechen/edm08/.

4 Mining for interesting patterns

Certainly, we were eager to find interesting and/or unexpected patterns in student feedback preferences or performance, in order to quantify whether feedback was helpful in answering related questions in the test, to determine if the performance and preference patterns of students with different learning styles differ significantly from each other, etc. However, first we would like to see patterns that reflect adaptation rules that effected feedback tailoring and other background knowledge as an evidence that the data confirms the system was performing according to planned behavior. We applied several DM techniques for association analysis, classification, and clustering (with and without dimensionality reduction). We describe the outcomes in the following subsections.

4.1 Classification

Instead of trying to achieve the highest possible accuracy, our goal was finding descriptive or discriminative patterns providing us with an insight and evidence supporting a particular hypothesis. One simple and straightforward approach is building a decision-tree (like C4.5) or a rule-based model (like JRip) and to analyze it.

Defining a concrete classification task we were able to investigate various patterns extracted from the classification models built on different subsets of the attributes that were selected either manually or with feature subset selection techniques. On the one hand this simple approach helps in searching evidence of the potential of hypotheses. We could see for instance from the C4.5 model that reading EF for a particular question raises the chances of answering a related question correctly and thus increases the chance of passing the whole test. On the other hand, the C4.5 classifier performs rather poorly on this dataset thus forcing us to be suspicious about any conclusion drawn from the model.

4.2 Clustering

During the test design, it is always a challenge to come up with questions that would cover the course material reasonably well, that are reasonably independent from each other (so that e.g. answering one question incorrectly would not strongly correlate with answering a large portion of the other questions incorrectly as well), that are able to capture possible misconceptions and that satisfy a number of other desired properties. Similarly, there is a challenge in designing effective feedback and tailoring it towards an individual student and his/her current needs.

Hierarchical clustering of questions according to the student response correctness or grade produces rather intuitive dendrograms. Similarly, intuitive outcome can be achieved by clustering based on the conditional probabilities P(Qc|Qr)/P(Qc), where r is the number of the row, and c the number of the column of the questions similarity matrix. I.e., the cell (r,c) of the matrix gives the conditional probability of answering question c correctly given that the question r was answered correctly. P(Qc) is the probability of answering question c correctly unconditionally. We see, e.g., that students answering Q3 correctly had a higher chance of answering Q4 correctly: 25% vs 15%. Of course, low numbers indicate also low support of the rule, hence few students supporting it. In general, answering Q3 correctly has an influence on almost all following questions; answering Q6 correctly has a big influence on Q10: from 68% to 100%.

However, due the curse of dimensionality, and poor representation spaces (redundant and correlated attributes, many of which may contain little discriminative/descriptive information) the performance of the many data-mining techniques seriously deteriorates. Another serious problem worth mentioning here is the highly unbalanced distribution of the instances in the dataset along some of the key dimensions (such as learning style and feedback requests) caused by the uncontrolled nature of the real assessment of the students (so that, e.g., sampling students into different groups is not ethical when they compete for a grade). We could hardly get any meaningful outcome as for finding groups of students with respect to their performance or feedback preferences in a fully automated manner. Nevertheless, association analysis was of some help for particular tasks.

4.3 Association analysis

Association rules express associations between different attributes. An association rule X=>Y expresses that the occurrence of X has a positive influence on the occurrence of Y. Whenever X is valid, the probability that Y appears is high.

Despite of a popular belief that association analysis often finds something useful and interesting from real datasets, we would like to stress here that in fact real success is fairly rare due to several reasons. In particular, many studies of association rules share one common problem: the output of association and pattern mining is often very large, in the sense that sometimes there are even more patterns than there were data to begin with. Also, the output contains often many redundant patterns, or is even misleading. We also experienced this even with our modest-size dataset, when the rules that define grade as a combination of response correctness and response certitude may not appear in the top *n* rules of the result set (unless we leave just these three sets of attributes) due to the presence of many other (hundreds) inherently redundant patterns.

Therefore, we switched to the semiautomatic discovery of the association rules focusing the search by predefining what should appear on the right hand side and what may appear on the left hand side of a rule. In particular, we were interested to find out what behavior may explain failing or passing the test, how feedback preferences differ for the students with different leaning styles and for similar questions. Association rules discovered in this way helped to find out, for example, that when answering a question correctly and reading example-based EF students often did not request additional theory-based EF; and we found that students with reflective and sequential learning styles who failed the test often did not study any EF. However, it is worth mentioning that the effectiveness of this approach is still rather questionable for this type of assessment data since applying focused (yet "manual") SQL queries and performing statistical analysis of the data might still require less effort than "automatic" association analysis. Certainly, DM approaches still can be potentially much more effective and therefore, favorable.

5 Conclusions and Future work

In this paper we described an EDM case study based on online assessment data where students were able to receive tailored immediate EF after answering each of the questions in the test. Our case study was aimed towards showing that even with a modest size dataset and well-defined problem(s), for researchers having some prior experience in theory and practice of DM it is still rather hard to obtain meaningful results with a set of traditional DM techniques including clustering, classification and association analysis. This outcome calls for further research in the directions of (1) defining appropriate interestingness measures for the patterns to be mined, (2) the integration of prior domain knowledge (not necessarily subject domain, but also knowledge about the adaptation rules implemented in particular WBLS components) into the DM techniques, and (3) tailoring DM technology towards the EDM needs in general.

Our further work in this direction will be devoted to the development of a generic EDM framework that is able to (re)discover background knowledge and incorporate this knowledge into the mining process focusing it on the discovery of the truly interesting patterns, and in particular, the identification of subgroups and emerging patterns.

References

[1] Hatie, J., Timperley, H. The power of feedback, *J. Review of Educational Research*, 87 (1), 2007, p. 81–112.

[2] Mory, E. Feedback research revisited, In: *Jonassen, D. (eds.) Handbook of research on educational communications and technology*. Mahwah, NJ, 2004, p. 745–783.

[3] Vasilyeva, E., De Bra, P., Pechenizkiy, M., Puuronen, S. Tailoring feedback in online assessment: influence of learning styles on the feedback preferences and elaborated feedback effectiveness. *Proceeding of 8th IEEE ICALT 2008*.

[4] Vasilyeva, E., Pechenizkiy, M., De Bra, P. Adaptation of elaborated feedback in e-learning. *Proceeding of 5th AH 2008*.

[5] Witten, I., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, 2005, Morgan Kaufmann.