The Design and Development of an Expert System Prototype for Enhancing Exam Quality

doi:10.3991/ijac.v3i3.1356

P. DeCarlo, N. Rizk
University of Houston, Houston, Texas, USA

Abstract—**This paper discusses the development of an expert system prototype for use in college institutions. Our aim is to enhance exam quality and student performance by obtaining metrics pertaining to assignments, study materials, textbooks, and lecture quality, then learning dynamically from this information to create a human-readable course evaluation. The goal is to obtain a model which can be applied to courses in which students struggle, so we can identify ways to enhance the most determining factor of their grade: the quality of the exam. This expert system will serve as a prototype for a larger, more comprehensive automated system which will be proposed to enhance curricula.**

Index Terms—**E-learning evaluation, association rule learning, course assessment, expert system development**

I. INTRODUCTION

Exam quality relates to how well an examination of learned material reflects the information provided in course learning materials. Learning materials include study guides, homework, books, lecture information, course videos, and exams themselves, should a comprehensive exam be given. Courses may suffer from lack of coherence between the offered materials. If a material is utilized it should be relevant and effective in preparing students for an exam. In a case of minimal coherence, students could have a potentially difficult time deciding exactly what to prepare for regarding an examination. The idea here is to capture the areas of the course which are deficient in providing preparatory knowledge needed for exam success, so that they can be modified to achieve a better synergistic effect on overall exam success. We also wish to keep or enhance effective course materials. A system which reports these findings to the instructor would benefit the course as a whole by providing the instructor information to support modification of the learning framework, which would guide students to optimal course performance.

Our work achieves this goal using a concept from data mining known as association rule learning. This technique is used to discover relations between variables in large example sets. This concept is particularly interesting because the rules this technique produces can be interpreted as easily as they can be read. For example, a typical rule may take the form {studies daily} => {has a high GPA}. This would mean that the feature 'studies daily’ implies a given example {has a high GPA}. The left most side of a rule is called the antecedent whereas the result on the right hand is known as the consequent. In order to generate meaningful and useful rules, support and confidence thresholds are supplied, which can limit the generation of weak rules. Support is defined as the probability that an example contains a subset X when randomly chosen from the total set of responses. Support of an association rule ‘A=>B’ is defined as the ‘support of (A union B)’. Confidence refers to the likelihood that for a transaction containing A, how likely is it that it also contains B. Confidence of an association rule ‘A=>B’ is defined as the ‘probability that an example contains B given A divided by the probability that an example contains A’. This is the same as the ‘support of (A union B) divided by the support of (A)’. Our study uses the algorithm proposed by Agrawal [1, 2]. This method has become popular in market basket analysis, intrusion detection and bioinformatics. It is also extensible to discovering correlations within data [3]. In the realm of education, association rules have been used to generate efficient learning work-flows [4], aid in academic advising [5], and to provide learning insights in the Moodle Course Management system [6].

This paper reports on a pilot study which was performed as a proof of concept to support the building of an automated system which can be utilized by instructors.

We based our design on the W-CAT model proposed by Rizk [7]. This model is composed of four modules. A base module contains student background information, course material, and teacher methods. This provides input into a reasoning module. Here we use association rule learning to find combinations of attributes contained in the base module which lead to high or low exam performance. These results are then analyzed through an inference module which searches for specific course deficiencies related to communication and pedagogical factors. A final expert module then produces an assessment by recognizing strong and weak course materials as well as study combinations which lead to poor or high exam performance.

This study was undertaken at the University of Houston. The evaluation of the results’ study focused on students’ exam success as determined by the letter grade received. The study investigated student perceptions of study material in terms of their preparatory significance and effectiveness in order to generate the rule sets. It discusses the findings and limitations of using association rule mining in a computer science course to determine the factors which contribute to exam quality, as well as providing evidence to support the creation of an automated course evaluation system.

The first purpose of our study is to identify students’ perceptions of course materials, so as to support decision making with regards to whether adoption would likely enhance exam quality and thus affect student performance positively. This focus may help leaders and faculties determine whether to pursue a particular solution in improving a specific course curriculum. The second purpose of this study is to acquire knowledge for the development of an automated system to aid instructors in future course development by dynamic evaluation. The next section describes the method used to gather our data.

II. METHOD

The first phase of analysis began with data collection. Participants in this study include 52 students enrolled in an introduction to computer organization and design course (COSC2410). The majority of students were full time and aged between 18 and 23 years old. The samples are comprised of students of different levels (freshman, sophomore, junior, senior) and different degrees.

We utilized the open-source Limesurvey software to generate an expert survey to collect information from students in COSC2410. The idea was to use a pre-survey before examination 2, a post-survey after completing examination 2, and then evaluate these as compared to the students actual exam results. The responses can be considered valid, as invitations to the survey were distributed using a secure token system which would allow only a finite number of tokens to be used in the survey, which were e-mailed to students individually. Students were encouraged to participate in the surveys, as we offered extra credit on the exam in return. Anonymity of responses was achieved by matching data to student identification numbers, with no need to ever ask for identifying information beyond that.

The surveys assessed the students' perception of the efficacy of certain course materials in preparing for the topics on the examination. The evaluation of survey information was based on statements regarding the course materials, as well as yes/no and multiple choice questions. The primary dependent variable is the exam performance measured by the letter grade received. Additionally, we asked students what type of what materials they would prefer to see more of.

A. *Students' perceptions*

This study considers including students as stakeholders in the evaluation process [8]. Thus, information regarding students' attitudes and preferred learning style should be known (Table I).

B. *Course Material Usage*

A series of questions about materials employed in the course was asked to gain insight into which materials students were using to prepare for the exam (Table II).

C. *Course Material Effectiveness*

Students rated the effectiveness of course material items in regards to how well they prepared them for the exam. This was done using a 5-point scale (Table III).

D. *Student Suggestion*

Students were asked if they recommended more programming assignments concerning the information on exam 2. 62% supported this recommendation.

E. *Exam Performance*

The information in A-C was to be evaluated against actual exam performance, indicated by letter grade received as determined by raw score (Table II).

To generate our rulesets we used the open-source data mining tool RapidMiner. The following describes the steps used in our RapidMiner process tree. The survey data was cleansed by converting the numerical grades to nominals A-F. These were then converted to binomial data. Questions which used a 5 point ranking scale were discretized into bins and processed as binomial data. We processed the data to find the frequent itemsets and proceeded to generate one-to-one association rules.

TABLE I.
QUESTIONS REGARDING STUDENTS PERCEPTION & ATTITIUDES

Question	Pre-Survey	Post-Survey
Adequately prepared for exam	24%	62%
Believed questions on exam were familiar and not surprising	NA	85%
Student who believed they would receive a B or above	81%	65%
Students who believed they would receive a C	17%	21%
Students who believed they would receive a D or below	2%	14%

TABLE II.
QUESTIONS REGARDING COURSE MATERIALS

Question	Pre-Survey	Post-Survey
Do you own the course textbook	NA	62%
Did you use the course textbook to study for the exam	NA	12%
Did you view online lectures to study for the exam (not review video)	55%	54%
Did you view the exam review video?	NA	90%
Did lectures correlate to knowledge expected for exam	90%	87%

TABLE III.
COURSE MATERIAL EFFECTIVENESS

Material	% who found material effective
Textbook	8%
Lecture videos (not review video)	34%
Exam review video	77%
Written homework (exercises)	67%
Programming Assignment	80%

TABLE IV.
ACTUAL EXAM PERFORMANCE

Letter Grade Bin	Exam1	Exam2
B or above	71%	46%
C	20%	20%
D or below	9%	34%

III. RESULTS

We used two strategies to obtain our results in the rule generation step. The first sought to lower the support threshold, so that rules could be generated which would contain a specific letter grade in the consequent. This would give us a rule that would apply to a final letter grade received and thus insight to exam performance.

TABLE V.
USING A SUPPORT THRESHOLD OF .10 AND MINIMUM CONFIDENCE OF .2 IN RULE GENERATION, TO FIND RULES WITH A SPECIFIC LETTER GRADE IN THE CONSEQUENT.

Generated Rule	Support	Confidence
Students who received an A on the first midterm => received A	.154	.375
Students who rated the exam review as Great + in post-survey => received A	.103	.500
Students who completed programming Assignment 2 => received B	.358	.429
Students who watched the exam review video -> received B	.359	.368
Students who were adequately prepared in the post-survey => received C	.103	.333
Students who studied primarily using the exam review video => received F	.179	.412
No results for D's were obtained because there were only 4 D's and not enough to meet the threshold	N/A	N/A

The second method sought to find strong association rules using high values for support and confidence. This would produce meaningful one to one associations among the independent variables or intercorrelations.

TABLE VI.
USING A SUPPORT THRESHOLD OF .80 AND MINIMUM CONFIDENCE OF .8 IN RULE GENERATION, TO FIND RULES WITH HIGH SUPPORT AND CONFIDENCE.

Generated Rule	Support	Confidence
Students who thought the lectures correlated to exam info on pre-survey => viewed Review Video	.846	1.0
Students who viewed the review video on post-survey => expected to do well on Exam 2	.846	.868
Students who thought the lectures correlated to exam info on the post-survey => thought the lectures correlated to exam info on pre-survey	.846	.868

IV. LIMITATIONS

Our findings from the pilot study have aided in discovering inherent limitations which we will attempt to control better in the design of our automated expert system.

We must first acknowledge the limitations present in our choice of using association rules. In many cases where association rules are used, very large numbers of rules can be generated which can prove to be fruitless and/or time consuming when trying to discover valid correlations. We did not have this issue as our small sample size, thresholds, and our single pass generation of one-to-one rules limited this effect. This explanation raises more issues though. We have not used a large sample size. Given that our method applies to a specific course; it may never become large unless we perform a longitudinal study, which, due to the nature of course curriculum changes, would likely prove to be irrelevant over time. Our design is unable to check for negative correlations, i.e. a rule which has the form 'does not contain A=>B'. Mining for such rules necessitates the examination of an exponentially large search space. Although potentially useful, we have neglected to use an algorithm which extends to generate these rule types [3, 9]. Due to our use of low support and confidence threshold in *results* [A], we must guard against reporting contradictory rules. This did not occur in this study but does remain a possibility so long as support and confidence thresholds below .51 are used. In the pilot study we did not generate rules greater than one-to-one. There may be useful antecedent pairings present in our data but we have not reported them.

We do not feel that a pre-survey will be necessary in future endeavors due to two factors. First, we wish to limit the total number of questions and surveys to a bare minimum. We would expect higher survey completion rates and more honest data from students if we reduce the frequency of survey delivery. Second, the data appears to be most relevant after the exam has been taken due to the fact that students have now seen the exam and are not answering questions based on expectation, but rather actual experience. In the future we could bring back the pre-survey to predict success based on responses from a previous course semester.

The teaching method used by courses which adopt this method for evaluation may have an impact on whether the results are useful. We assume that most courses with a pre-determined syllabus would be able to benefit from an association -rule based system which allows the instructor to specify the learning materials present in the course. Some courses may only have a textbook, lecture, and final project. In this case, we could evaluate against the final project grade rather than exam grades, but may not capture other factors that would lead to success on the project. For example, team size, student background, and student classification may prove to be better indicators. We may need to develop a different survey for different pedagogical styles.

Furthermore, the validity of the questions used in the survey has not been determined. We feel this should be done in the future when other courses have applied our system. We could ask instructors to rate the perceived usefulness of the results and identify the teaching style used in a final survey to obtain validity measures. We could also obtain suggestions from users to better aid in capturing relevant success indicators.

V. CONCLUSION

The results of this pilot study have aided in the development of an automated course/exam evaluation system. We performed this study with the intent of gaining the domain knowledge necessary to build an expert system of this type. A description of this system follows:

Instructors will be able to login to a system and define a course to evaluate. They simply need to provide information concerning materials present in the course and student names / identification numbers. The system then generates dynamic surveys and a link is sent to students to collect responses. Upon completion of the surveys, the software will generate the association rules. From these rules, we still need to develop an inference engine which can provide dynamic feedback to course instructors. Finally, we parse rule sets into a human-readable paragraph which will provide the patterns followed by students who are earning high or low exam marks and a summary of the course material efficacy.

An example of what instructor feedback would look like in this pilot study:

1) *Regarding exam performance*

Students receiving high marks on exam 2 completed the programming assignment. They study using the exam review video and find its content relevant to the exam material. Students with low marks on exam 2 study primarily using the exam review video.

2) *Intercorrelations*

Students who thought the lectures correlate to the knowledge expected for exam 2 watched the review video. Those who viewed the video expected to do well on exam 2. Before and after the exam students thought the lectures correlated to the exam information.

3) *Summary*

Students are not using the textbook to study. Only 62% of students have the textbook. The homework exercises, review video, and programming assignments are considered effective learning tools.

4) *Recommendations*

Students support the recommendation of assigning more programming assignments based on the exam 2 information.

From what we have obtained in our results, it is indeed possible to create an automated system which could produce results similar to this for any course. *Regarding exam performance* is taken directly from the results in A. Provided data is captured and parsed properly we would expect this to be repeatable. *Intercorrelations* are taken from B. Again, this could also be repeated assuming proper data acquisition and handling. *Summary* generates information directly from the survey. It is based on response frequency in the survey. *Recommendation*s are also determined directly from response frequency.

It should be noted that inter-correlations may not always be useful, as the results may be too vague or unrelated. For example, in this study we would generate ‘students who though the lectures correlate to the knowledge expected for exam 2 -> watched the review video’. This information does not seem very useful. On the other hand, inter-correlations may give support for conclusions arrived at in A. For example, ‘students who viewed the video->expected to do well on exam2’ and students who received low grades->studied primarily using the review video. An instructor may infer that the review video inspires false confidence in students.

We took knowledge from this study survey and applied it to the course in the following semester. We noticed students were not using the book and did not think it was effective. This prompted further inquiry from the professor to the students. Students explained that we do programming and logic design in this course, and that the programming is not covered well in the current textbook, and also that the logic design portion is contained in an appendix. We replaced the old textbook with a smaller one focused on MIPS programming and another textbook dedicated to logic design. With a book that better models the information on the exams; we expect to see students using it more—and more effectively—as a study-aid. We have also added more programming assignments to supplement exam2. We believe this will increase the frequency of high marks on the exam, as it was one of the indicators in our study.

We currently have a web interface where instructors can enter the information necessary to create dynamic surveys for their courses and no longer rely on LimeSurvey. Responses have already been collected from two computer science courses this semester using this system. We are currently developing the inference engine and parser to produce human readable feedback. We seek to involve more courses to adopt our software when it is complete so that we may begin evaluating its efficacy. We believe the final product will be able to actively evaluate physical and electronic classrooms which are based on examinations, with the benefit of improving teaching and student learning.

REFERENCES

[1] R. Agrawal, T. Imielinaki, and A. Swami, “Mining association rules between sets of items in large databases”, *Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.

[2] R. Srikant and R. Agrawal, “Mining generalized association rules”, *Future Generation Computer Systems*, v13 n2-3 Nov. 1997.

[3] S. Brin, R. Motwani, C. Silverstein, "Beyond market baskets: generalizing association rules to correlations". *ACM SIGMOD* Record 26(2): 265-276, 1997. doi:10.1145/253262.253327

[4] S. Enclieva, S. Tumin, “Application of association rules for efficient learning work-flow”. *IFIP International Federation for Information Processing*, Volume 228, *Intelligent Information Processing III*, pp. 499-504, 2006.

[5] C. Romero, S. Ventura, E. Garcia, “Data mining in course management systems: Moodle case study and tutorial”, *Computers & Education*, v.51 n.1, pp.368-384, August, 2008. doi:10.1016/j.compedu.2007.05.016

[6] A.Bykat, “Knowledge-based academic advising systems: survey, evaluation, and recommendations”, Cybernetics and Systems: An International Journal, 1087-6553, Volume 28, Issue 7, 1997, pp. 571 – 590.

[7] N. Rizk, “Witty Curriculum Assessment Tool,” unpublished.

[8] L. Towery, and R. Oliveri, “Engaging stakeholders in professional development and its evaluation,” *the Evaluation Exchange*, vol. 11, no.4, 2006.

[9] X. Yuan, B. Buckles , Z. Yuan , J. Zhang, “Mining negative association rules”, *Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)*, p.623, July 01-04, 2002

AUTHORS

Paul J. DeCarlo is a graduate student with the Department of Computer Science, University of Houston, Houston, Texas, 77204 USA (e-mail:pjdecarlo@uh.edu).

Nouhad J. Rizk was with Notre Dame University. She is now with the Department of Computer Science, University of Houston, Houston, Texas, 77204 USA (e-mail:njrizk@uh.edu).

This article was modified from a presentation at the ICELW2010 Conference in New York, USA in June 2010. Submitted July 21st, 2010. Published as resubmitted by the authors July 29th, 2010.