Work In Progress: Establishing multiple contexts for student's progressive refinement of data mining

David Kwartowitz[1], Sean Brophy[2], Horace Mann[3]

Abstract - **Students' systematic investigation of challenges from different prospectives provides a richer learning experience than lecture and traditional book based learning. Students in a course in Data Mining, which was taught through a computer science department, were presented challenges in the field of Biomedical Engineering in addition to text book problems that targeted a variety of domain areas. The combination of focused inquiry around two contexts provides students with multiple examples to apply and differentiate their knowledge. This paper presents preliminary results from this pilot implementation of the instructional method and next steps for the course offering in the spring semester.**

Index Terms - Interdisciplinary, Crossing Domain Areas, Computer Science, Biomedical Engineering

INTRODUCTION

Engineering and computer science are fields that continually require interdisciplinary support to achieve the goals of their projects. Therefore, instruction needs to include experiences that prepare students with the skill sets necessary to manage a project that contains unfamiliar domain knowledge. While there is some trend to teach certain courses using real world examples, this trend has not spread far enough and some of the strengths of this technique have not yet been realized. Many courses taught using real world examples tend to choose examples from a scattered number of domain areas and approach the examples as singular experiances as opposed to building an understanding of the underlying concepts. We are researching the benefit of anchoring students' learning experience in a cohesive set of example sets that continually help reinforce their ability to make sense of a particular domain area.

Exposing students to a more focused set of challenges which build around a central theme can be shown to help students gain a grasp of concepts with respect to the theme [1]. Application of this model will help students to learn the subject being taught with a context of the model or examples being used to demonstrate the subject. This exposure of students to examples from a different domain area from the general subject of the course can prove beneficial in teaching students not only the domain areas, but techniques in which they can gain knowledge about a field with which they are unfamiliar. This underlying goal of cross domain teaching will help the students in future interaction with domain experts and with other engineers working on the same project.

An important aspect of this multiple context approach is educating students to be able to realize what information they do not know and need to find. For a learner, understanding what they do not know can sometimes be as important as understanding what they do know [2] [3]. The ability to formulate questions to either be asked of a domain expert or found in a reference is paramount to the ability of a student to use their knowledge in contexts outside of their current knowledge. The field of data mining is of specific interest for presentation in this interdisciplinary motif because of the broad focus in which it can be implemented. Data mining is a general field with applications in multiple contexts such as medicine, consumer profiling, and information management. Data Mining involves the implementation of traditional statistical models to abstract data thus meeting the needs of clients.

METHODS

The course was taught for the first time in Fall 2003 and consisted of seven fourth year Computer Science students. These students had minimal background in data mining and data analysis prior to enrollment. Some had already taken statistics or were taking statistics concurrently with the data mining course. Only a few had experience with databases. The course was offered as a technical elective and the students were informed before they registered that a different instructional method was going to be used in the course.

The course was designed to implement the theories presented in a National Academy of Science report called How People Learn [1] [4], in concert with sections taught in a traditional didactic style. The original instructional design planned to use data mining techniques for analyzing medical records (predicting methods) and identifying gene sequences (categorization methods). Students were presented with challenges in which they were given the opportunity to explain what they knew about the problem, then compared and contrasted various data mining methods to identify the strengths and weaknesses of each analyzing the specific problem.

The assessment scheme in the course was to have students complete a pre and post test for each major subject area covered. The pre and post assessments were designed to be similar in order that the pre-assessments would guide the students to the important questions to be answered and the post-assessments would demonstrate to students their growth over time. Students were also presented with a series of challenges which led to a "grand challenge." The "grand challenge" was

chosen from the field of biomedical informatics, examining medical and diagnostic records of patients who were suspected of having breast cancer. The students were provided various information from the Wisconsin Breast Cancer Database [5] which contains records of patients from the University of Wisconsin, Madison. The students were provided with a tool called WEKA [6], an open source data mining application, which provides a collection of different data mining techniques packaged under a common Graphical User Interface (GUI). The GUI is fairly intuitive and well received by the students. This tool was chosen because it allows the students to interact with the various data mining techniques minimal financial and academic overhead. Students were interviewed following the class. Using a think aloud protocol, two students were asked to solve a challenge similar to the "grand challenge" related to predicting breast cancer. Because data mining contains myriad techniques, the ability for students to determine the optimal technique for the problem which was presented to them is essential to the diagnosis of their understanding of the field. These interviews were conducted in the Spring of 2003 after the students were given time in which they were not reviewing the material on a daily basis. The main purpose of this was to examine both what they learned and what they retained.

RESULTS AND DISCUSSION

The first implementation of this course provided an important pilot test of our hypothesis and identification of issues and opportunities for refinement. The pre-tests indicated students awareness of data mining methods used for consumer marketing questions such as predicting consumers potential to spend based on their buying profiles, but were not aware of the application to biomedical engineering situations. Only a couple of the students could differentiate the concepts of data, information, and knowledge. By the end of the course most of the learners could define these terms well in general terms.

Our original goal was to have students use a server delivered version of WEKA to compare and contrast various data mining techniques for biomedical contexts. Numerous technical problems minimized the performance of this system making it unusable for the students. Therefore, many of the assignments had to be completed by a traditional approach of working out small problems by hand. Late in the semester a desktop version of WEKA became available. Students used this version to complete an end of semester project that asked them to compare and contrast three data mining techniques to analyze the Breast Cancer Data set. Students reported having little difficulty understanding how to use the software and spent most of their time making decisions about how to prepare the data for analysis and analyzing the results.

We are now in the process of identifying more case examples within bioengineering that we can use to anchor students experiences with data mining techniques. Since WEKA was not available we needed to rely on the examples provided in the textbook to provide the context for the knowledge the students were learning to apply [7]. The interviews with

the students indicated that these examples do not provide the cohesiveness we are looking for in multiple context of the content. The students found the WEKA example and breast cancer case extremely compelling and wanted more of that sort of exploration early in the semester. Our next offering of the course consists of a series of learning activities that progressively lead students through the process of preparing data for analysis, performing analysis, and evaluating the robustness of the analysis.

ACKNOWLEDGEMENTS

REFERENCES

[1] Cognition and Technology Group at Vanderbilt, *The Jasper Experiment*. Hillsdale, NJ: Lawrence Erlbaum, 1997.
[2] J. Bransford, *Human Cognition: Learning, Understanding, and Remembering*. Belmont, CA: Wadsworth, 1979.
[3] J. Greeno, *Trends in the theory of knowledge for problem solving*. Hillsdale, NJ: Lawrence Erlbaum, 1980.
[4] J. Bransford, *How People Learn*, expanded ed. National Academy Press, 2000.
[5] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html
[6] I. H Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
[7] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.