Development of a Data Mining Course for Undergraduate Students

Terri L. Lenox
and
Carolyn Cuff
Department of Mathematics and Computer Science, Westminster College
New Wilmington, PA 16172 USA

ABSTRACT

The goal of this project is to develop an introductory course for junior/senior computer science and mathematics undergraduate students on the topic of Data Mining. The course will be team developed and team taught at Westminster College by faculty in the Mathematics and Computer Science department. A detailed course description and day-by-day outline including textbooks, supplementary texts used as background, laboratory exercises, homework assignments and summary of lectures will be developed. This paper discusses some of the necessary steps to develop this course.

Keywords: data mining, computer science curriculum

1. INTRODUCTION

Current collegiate courses in data mining primarily appear at the graduate level. A carefully designed undergraduate course in data mining would introduce students to concepts in data mining by extending the existing database course and applying statistical concepts within the computer science curriculum. This data-mining course will be designed to integrate current research into the curriculum and promote the student's critical thinking and problem-solving skills. This course would provide an upper-level elective, offering a glimpse into innovative research. Research in computer science is often portrayed as very theoretical or hardware-related. Students become reluctant to pursue graduate-level research due to this stereotype. The authors hope that this course would demonstrate a more applied type of research and make graduate study seem more approachable to graduates of liberal arts colleges.

This paper discusses some of the necessary steps to develop an undergraduate data mining course including : 1) development of a detailed course description and day-by-day outline; 2) refinement of course outcomes and expectations; 3) selection of textbooks and supplementary texts used for background; 4) selection of appropriate software tools; 5) development of laboratory exercises and homework assignments; and 6) creation of lectures.

1.1 Introduction to Data Mining

Data mining consists of a set of quantitative techniques that help to extract patterns of information or rules from large amounts of data. It is used in business to profile customers with greater accuracy, and to detect fraud. In scientific disciplines, it can be used to analyze the large set sets produced by applications such as medical imaging or remote sensing (Elmasri & Navathe, 2000). Data mining requires access to large amounts of quality data, the ability to use commercial data mining tools, and the sufficient computational power. Large data sets are now being produced and are readily available. Students have access to powerful and fast computers that could be used to mine these data seeking to extract knowledge. However, few courses have been developed at the undergraduate level to lead prepared students to understand the purpose and general methodology of data mining

1.2 Rationale

Westminster College is a liberal arts residential college with an enrollment of approximately 1,500 full-time undergraduates. The College offers majors in Mathematics, Computer Science and Computer Information Systems. These majors are housed in a combined Mathematics and Computer Science department which has five full-time mathematicians and three full-time computer scientists. All eight of these faculty hold Ph.D.s in respective fields. The total number of majors in the department has been steady for the last ten years averaging approximately 25 graduates per year. On average two are double majors in

mathematics and computer science, two mathematics majors have minors in computer science and four computer science majors have mathematics minors. Although the course offerings within each major are strong, few upper-level courses integrate mathematics and computer science.

Westminster College's curriculum has a track record of innovation. In Fall, 1998, a four-credit hour discrete analysis course replaced Calculus I as the gateway course to all majors within the department. This course, depending on the major, is followed by at least two additional required credit hours in discrete analysis. Mathematics, Computer Science (CS) and Computer Information Systems (CIS) majors are required to take the introductory statistics course and single variable calculus. Mathematics majors are required to take the introductory computer science course and a two semester sequence mathematics intensive course. Often mathematics majors elect to fulfill this requirement by taking the second introductory computer science course and Data Structures. All majors at Westminster are required to take a Capstone course. Within the Mathematics and Computer Science department, the Capstone course has the form of a major individual or group project. The graduating class of 2001 was the first class required to complete the Capstone course.

The faculty believes that the integration of mathematics within computer science and the option of integrating the computer science within mathematics is strong at the freshman and sophomore level. The authors seek to strengthen these ties at the upper level by the introduction of the Data Mining course. The authors believe that an undergraduate Data Mining course is accessible to strong computer science majors, will incorporate mathematics in the upper level computer science courses, provides students with exposure to state-of-the-art applications and is suitable in a liberal arts environment. A strength of the computer science discipline is in the application of foundational concepts to particular domains. A data mining course demonstrates how these foundational concepts can be built upon in new and innovative ways.

An additional consideration in offering a data mining course is that data mining jobs are becoming more common. A quick search of several on-line job banks indicates entry level positions for programmers, analysts, and engineers (jobsearch.monster.com). These positions appear to require a mathematics or computer science degree with some data mining and statistics exposure.

Finally as the number of students who obtain master's degrees in computer science and/or information sciences slows (U.S. Department of Education, 2001), the authors hope to encourage students to consider graduate studies and believe that innovative topics such as data mining may motivate students.

1.3 Data Mining Courses at Other Schools

Although data mining is of increasing interest to industry, it does not appear to be a common undergraduate course at this time. However, it is likely that some institutions offer data mining and/or data warehousing subjects under a special or advanced topics listing which makes it difficult to survey. The list below from a brief Internet survey of over 125 institutes found the following undergraduate courses in data mining:

1. Data Warehousing and Mining (IFMG 455); Indiana University of Pennsylvania (Pierce, 1999).
2. Data Mining (ECMM 420); Christopher Newport University; e-commerce program.
3. Data Mining (INFO 329); Xavier University; cross-listed as a marketing course (MKTG 329).
4. Advanced Quantitative Methods in Business (BUS 491); Cal Poly San Luis Obispo; business course.
5. Emerging Database Technologies and Applications (CS 4440); Georgia Institute of Technology.
6. Data Mining (CS 373); LaSalle University.
7. Elementary Data Mining (Stat 321); Central Connecticut State University; on-line statistics course.

The following lists some graduate courses in data mining and/or data warehousing:

1. Knowledge Discovery and Data Mining Masters program; Carnegie Mellon University
2. Data Warehousing and Data Mining (CS 753); Bentley College.
3. Data Warehousing and Data Mining (IS 549); DePaul University; data warehousing concentration.
4. Data Mining and Knowledge Discovery (INFSY 566); Penn State University.
5. Information Systems Data Warehousing and Decision Support ISC 571); University of South Alabama.
6. Data Warehousing and Data Mining (ISM 6930); University of South Florida; MBA program
7. Data Mining (COSC 7397); University of Houston.
8. Advanced Topics in Spatial Knowledge Discovery and Data Mining Systems (CS 783); North Dakota State University.
9. Advanced Topics in Data Mining Architectures (CS 785); North Dakota State University.
10. Introduction to Data Mining (Stat 521, CS 580); Central Connecticut State University on-line.
11. Data Mining (INFO 780-08); Drexel University.
12. Database Systems Planning (IS 304); Claremont Graduate University.

In addition to traditional courses, several institutions offer certificate programs in data mining and/or data warehousing:

1. Data Mining Techniques and Applications; UCLA's Engineering, computer science and technical management short courses.
2. Data Mining; SAS Institute.
3. Data Mining; Central Connecticut State University; on-line.
4. Data Mining : Level I; The Modeling Agency
5. Penn State Data Mining Certificate Program.

Because Westminster College is a traditional liberal arts college, our students are not currently interested in web-based courses.

2. DEVELOPING A DATA MINING COURSE

2.1 Prerequisite Knowledge

The prerequisite database course at Westminster College provides the student with an introduction to the theory and practice of applying database technology to the solution of business- and information-related problems. Database terminology and concepts, data and file structures, and a comparison of the relational database with other models (hierarchical, network and object-oriented) are addressed. Experience is provided by database design and implementation based on a thorough requirements analysis and information modeling. An introduction to relational database technology is provided, highlighting the use of structured query language (SQL).

The focus of traditional database courses is to provide students with a solid foundation on which to build and support operational databases. Data warehousing focuses on the creation of large consolidated databases to aid in the strategic decision making functions of an organization. Data mining typically focuses on analysis of these databases in an attempt to find previously unknown associations and patterns in the data. Since a purpose of this proposed data mining class is to strengthen the relationship between computer science and mathematics concepts, the statistical and analytical aspects of the subject are emphasized.

2.2 Development Plan

The planned tasks for this project include development of the data mining course, field testing materials and post-hoc surveys. The authors will also team teach this course twice with an evaluation and dissemination period between sessions. Two institutions will be selected to field test this data mining course. Their evaluations will be analyzed and the course modified where necessary.

The data mining course will have the following six prerequisites: Principles of Computer Science I and II, Theoretical Foundations of Computer Science, Data Structures, Database Theory and Design, and Statistics. This course will to complement the existing Neural Networks and Artificial Intelligence courses, and provide a basis for additional projects for the Capstone course.

A brief outline of the material that will be covered in this data mining course follows. Pruning this list is expected to take the most effort during the development of this data mining course.

1. Additional background in statistics: likelihood functions, analysis of variance, clustering, and Bayesian networks.
2. Background in machine learning: concept learning, version spaces, decision trees, overfitting, Occam's razor, neural networks, single-pass learning algorithms and reinforcement learning.
3. Review of text-based database management systems.
4. Extension of traditional database topics:
   - Multimedia databases including: time sequences, photographs and medical images, video clips, feature extraction, continuous media storage and delivery.
   - Database methods including: massive datasets and association rules, frequent sets, sampling, visualization of large data sets.
   - Data warehousing and on-line analytical processing (OLAP).
5. Data mining topics may include (Han & Kamber, 2000):
   - Data preparation.
   - Association rule mining (market basket analysis, basic concepts, and association rule mining).
   - Classification by decision tree induction (decision tree induction, tree pruning, extracting classification rules from decision trees, enhancements to basic decision tree induction, scalability and decision tree induction, integrating data warehousing techniques and decision tree induction).
   - Bayesian classification (Bayes theorem, naïve Bayesian classification, and belief networks).
   - Classification by backpropagation (multiplayer feed-forward neural network, defining a network topology, backpropagation interpretability, classification based on concepts from association rule mining, k-nearest neighbor classifiers, and case-based reasoning).
   - Prediction (linear and multiple regression, nonlinear regression, and other regression models).
   - Cluster analysis (types of data in clustering analysis, interval-scaled variables, binary variables, nominal, ordinal, and ratio-scaled variables).
   - Model-based clustering methods (statistical approach and neural network approach).
   - Outlier analysis (statistical-based outlier detection, distance-based outlier detection, and deviation-based outlier detection).

2.3 Software Tools

At this time, the first two software packages from the list below are being evaluated. Westminster College students have experience with either S-Plus of SPSS, which makes both Clementine and S-Plus logical choices for this data mining course.

1. Clementine from SPSS.
2. S-Plus (desktop) or Insightful Miner from Insightful Corporation.
3. PowerPlay and related tools from Cognos.
4. VisualMine and Daisy from AI Softw@re.
5. Data Mining Suite or Darwin from Oracle Corporation.
6. Intelligent Miner from IBM.*
7. Weka from the University of Waikato. *
8. TMiner Personal Edition from Department of Computer Sciences and Artificial Intelligence of Granada University (Spain) .*

A more in-depth list of data mining software tools can be found on the Knowledge Discovery and Data Mining web page (www.kdnuggets.com/software/suites.html).

2.4 Data Mining Text Books

The number of data mining and data warehousing textbooks has increased substantially over the past few years. The seminal textbook on this subject is Inmon's Building the Data Warehouse from 1996. The Han and Kamber Data Mining: Concepts and Techniques textbook is the likely choice for this data mining course. The following lists a few of the many available books:

1. Han, J. and Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
2. Inmon, W. H. (1996). Building the Data Warehouse, (2nd Edition). New York , NY : John Wiley & Sons, Inc.
3. Berry, M. and Linoff, G. (1997). Data Mining Techniques for Marketing, Sales and Customer Support, New York , NY : John Wiley & Sons.
4. Berry, M. and Linoff, G. (1999). Mastering Data Mining, New York , NY : John Wiley & Sons.
5. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi , A. (1998). Discovering Data Mining. Prentice Hall PTR.
6. Weiss, S. and Indurkhya, N. (1997). Predictive Data Mining: A Practical Guide. Morgan Kaufmann Publishers.
7. Westphal, C. and Blaxton, T. (1998). Data Mining Solutions , New York , NY : John Wiley & Sons.

3. OUTCOMES EXPECTED FROM THIS PROJECT

Outcomes from this project are based on measures of student achievement, transferability of this course to another similar institution, and whether this course can be adapted to meet the needs of other undergraduate institutions.

Student outcomes are as follows (ACM Computing Curricula 2001 Ironman Report, 2001):

- Given several data sets, students will be able to understand and discuss ways to prepare (clean) and warehouse data.
- Given a prepared set of data, students will be able to classify data based on cluster analysis, decision trees, Bayesian networks or backpropagation algorithms.
- Given a prepared set of data, students will be able to analyze the data using one of the techniques discussed.
- Given an analysis of a particular data set, students will be able to make appropriate predictions.
- Students will be able to understand, adapt and apply algorithms.
- Students will be able to outline the applications of data mining techniques.
- Some students could be motivated to continue computer science research or graduate studies.

3.1 Evaluation Plan

Student outcomes will be evaluated by some of the traditional methods (projects, written examinations, laboratory exercises and homework assignments) during the two semesters that this course is initially taught. Once the course has been taught, it will be fine tuned and the authors hope to disseminate our findings at both mathematical and information science forms such as American Statistical Association (ASA) Joint Statistical Meetings and Information Systems Education conference (ISECON).

4. REFERENCES

ACM Computing Curricula 2001 Ironman Report February 6, 2001. http://www.computer.org /education/cc2001/ironman/cc2001/index.html).

Elmasri, R. and Navathe, S. B. (2000). Fundamentals of Database Systems. Addison-Wesley Longman, Inc.

Han, J. and Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Knowledge and Data Discovery Organization, http://www.kdnuggets.com /software/suites.html

Pierce, E. M. (1999). "Developing and delivering a data warehousing and mining course, " Communications of the Association for Information Systems, Vol. 2 (16).

* Shareware or freeware.

U.S. Department of Education, National Center for Education Statistics, Higher Education (HEGIS) (2001). Table 2.86 "Earned degrees in computer and information sciences by degree-granting institutions."