Data Mining at the Undergraduate Level

Dian Lopez
Computer Science Department
University of Minnesota, Morris
lopezdr@mrs.umn.edu

Luke Ludwig
Computer Science Department
University of Minnesota, Morris
ludwigl@mrs.umn.edu

Abstract

A Data Mining course at the University of Minnesota, Morris, first taught during Fall 2000, is one of very few undergraduate courses offered on data mining today. Data mining is the process of discovering patterns in large amounts of data for the purpose of solving problems, gaining knowledge, and making predictions. Students in the Data Mining course at UMM have gained considerable knowledge and insight into the processes involved in data mining through hands-on experience with data mining applications. The Data Mining course at UMM has shown that teaching data mining at the undergraduate level is appropriate and can be successful.

Introduction

The ability to store data digitally within databases has revolutionized our society. Data which once filled entire rooms of filing cabinets can now be stored in a desk drawer on various forms of digital media. This technological breakthrough for data storage has resulted in an overwhelming amount of existing data today. Businesses and governments have realized the value of information and have thus collected and stored data about anything that may be useful to them. Yesterday we were measuring data in gigabytes, today we are measuring in terabytes, and sometimes even in pedabytes. Studies have shown that an average dot-com business doubles their data every 90 days, and estimations have been made that the quantity of data in the world's databases doubles every twenty months [1][2]. Continuing at this rate the measuring of data will soon exceed pedabytes and proceed into the next category, exabytes.

Data is being created faster than we are able to understand and use it. There may be patterns hiding within this data with potentially useful information. Data mining attempts to discover these patterns and unravel this information. Data mining can be formally defined as the process of discovering patterns in large amounts of data for the purpose of solving problems, gaining knowledge, or making predictions. This is a relatively new field that is becoming prominent in today's business world and will prove to be a critical factor in the successful operation of businesses, governments, and other organizations in the future. As the amount of data grows, the importance of gleaning valuable information from data also increases. Data mining, the key to acquiring this valuable information, is giving entities a new type of ROI (Return on Investment), a Return on *Information* [2].

One of the first areas that used data mining was marketing. Data mining helps businesses to learn their customers' preferences and buying patterns to determine whom to offer products, thereby helping to maximize profits. Data mining can be used anytime there exists data to be analyzed; for example, credit-risk analysis, fraud detection, pharmaceutical research, student recruiting and retention, and even food-service menu analysis. Data mining has also been useful in the world of science. New knowledge has been gained from patterns found in molecular structures, weather changes, genetics, and astronomical structures. Data mining techniques are also being used to analyze data gathered by radio telescopes to search for signs of intelligent extraterrestrial life.

The numerous applications and successes of data mining clearly show that data mining is an important and useful tool that should be invested in for the future. The demand for information technology specialists trained in data mining is increasing. Educating tomorrow's information technology specialists in the fundamental concepts of data mining should be a goal for educational institutions. Data mining classes have become prevalent at the graduate level, but few are found at the undergraduate level. An undergraduate Data Mining course at the University of Minnesota, Morris, has shown that undergraduate students can learn about data mining and apply what they have learned to solve existing problems.

Development of the Course

The Computer Science Discipline at the University of Minnesota, Morris (UMM) is designed to provide students with a strong foundation in the diverse and rapidly changing field of computing. This involves offering original elective courses each year on new and innovative approaches to Computer Science, such as the field of data mining. A two-credit course on data mining was designed and taught by Dian Lopez, one of the authors of this paper, during Fall 2000. The goal of this course was for students to develop an understanding of the fundamental principles of data mining, and then apply these concepts to real data, thus gaining a working knowledge of data mining techniques. For the class to be a success, a few accessories would be needed: data, a good textbook, and data mining software.

A data warehouse helps to arrange data in a manner useful for data mining. Data warehouses contain large amounts of data from many separate databases. This data is organized for data mining analysis, as opposed to the conventional usage of a database. Dan Flies, a UMM student, through a directed studies course, found real data and arranged it in the form of a data warehouse for the class to use [3]. This proved to be an ideal situation as the students were able to focus on using data mining applications instead of spending time arranging data in a data warehouse.

Data Mining, a textbook written by Ian H. Witten and Eibe Frank [1], was chosen as the textbook for the course. This textbook provided an up-to-date resource for the students to learn the fundamentals and concepts of data mining. A fully functional data mining software program, Weka, accompanies the book. Weka is a set of machine learning algorithms for solving real-world data mining problems [4]. It is an open source application written in Java that is perfect for a classroom setting in which students can look at, change, and add new algorithms to the source code. Weka was instrumental in giving the students hands-on data mining experience.

Implementation of the Course

To motivate the study of data mining, the students researched a data mining topic of their choice and presented their findings to the class. The class contained twelve students, and the presentations were done in groups of two or three. Topics presented were data warehousing, data mining in e-commerce, text mining, Knowledge Seeker (a data mining software program) [5], and predictive modeling. This approach to beginning the class required the students to investigate data mining in journals and on the web and to attain a better understanding of the breadth and depth of the field. Students learned quickly that data mining is an important new addition to Computer Science that is being applied in many settings.

The fundamentals and concepts behind data mining were studied next. Lectures were given that corresponded to chapters one through five of the textbook *Data Mining*. The class compared classifying, associative, clustering, linear regression, and instance based

approaches to data mining. Specific algorithms from each approach were analyzed, such as OneR, Naïve Bayes, ID3, Prism, and Apriori. The students learned how to evaluate the results of data mining algorithms and to predict their performance. Statistical techniques such as cross-validation and the bootstrap method were also studied. Weka proved to be a very useful learning tool. All of the aforementioned algorithms, and many more, are provided by Weka. The students were able to perform these data mining algorithms upon practice data sets supplied with Weka. Actually seeing the algorithms work and analyzing the results reinforced concepts the students had learned.

During the remainder of the course, students in groups of two or three applied what they had learned about data mining to a self-designed project. The focus of the project was to perform data mining upon real data assembled in the form of a data warehouse and to attempt to gain new knowledge from this data. This project modeled industrial data mining, and students learned practical data mining knowledge through hands-on experience.

The data warehouse contained forestry data from the U. S. Forest Service. The data had information on plots of trees and on individual trees in the states of Minnesota, Michigan, and Missouri. Some of the attributes stored for each tree were age, species, latitude, longitude, owner, life status, diameter, and damage. As an example, in Minnesota there are 247,721 instances of individual trees in the data warehouse. This data warehouse would be considered very small on the scale of commercial data warehouses and may better be described as a miniaturized data warehouse.

The self-designed project required the students to develop a proposal for their research. Before this could occur, it was necessary for them to become familiar with the forestry data. An online user's manual [6] accompanied the data, which was very helpful since none of the students were experts in the field of forestry. The manual described all of the attributes and their values, thus giving the forestry data meaning to the users.

As the students began to understand the forestry data, their projects began to take shape. Interesting questions were formulated such as how strong the relationship is between the diameter and age of an aspen tree, and how effective insects and diseases are at killing trees as opposed to just damaging them. Some groups took a different approach and chose to design their own data mining algorithms. For example, one group created a modified One-Rule algorithm that utilized the Weka interface [7]. Another group discovered that Weka's implementation of the Prism algorithm did not work correctly and implemented a corrected algorithm themselves [8].

Results

One group studied a hypothesis put forth by the U.S. Forest Service: There is a positive correlation between the lifespan of aspen trees and the latitude at which they grow [8]. This hypothesis asserts that aspen trees in northern latitudes live longer than aspen trees in southern latitudes because disease and insects have less time to affect aspen trees due

to the shortened growing season. Statistical analysis, the associative Apriori algorithm, and clustering algorithms were applied in an attempt to support or refute this hypothesis. The average latitude of aspen plots in Minnesota were compared across age groups. The analysis from table 1 supports the hypothesis that an aspen's life span increases with latitude.

Table 1**:** Average latitude of non-consecutive five-year age groups [8]

Plot Age	Latitude (degrees)
70 - 75	47.54
80 - 85	47.58
90 - 95	47.76

Also, an almost linear relationship was shown to exist between latitude and the killing effectiveness of disease upon aspen trees. Table 2 suggests that disease is more effective at killing trees the further south they exist, presumably due to the longer growing season.

Table 2: Ratio of dead diseased trees to total diseased trees separated by latitude [8]

Latitude (degrees)	Dead-Diseased/Diseased
46 - 46.5	18.1%
47 - 47.5	16.2%
48 - 48.5	12.4%

While these two examples are just a sampling of the results obtained during this project, the complete results are described as lending support to the U.S. Forest Service's hypothesis. The project concluded that at the very least the results "can provide the U.S. Forest Service with evidence that there is truth to their observations that an aspen's life span does increase with latitude, as well as circumspect evidence that the killing effectiveness of disease decreases as latitude increases." [8]

Conclusion

The undergraduate Data Mining course at UMM was a success. Students developed a concrete understanding of the fundamentals and principles of data mining. They applied these concepts to real data and gained a working knowledge of data mining techniques. In particular the data warehouse gave the students valuable experience with the idiosyncrasies and imperfections of real data. Students learned that data mining is an involved process that doesn't always produce interesting results. This course has shown that undergraduate students can learn about data mining and apply what they have learned to solve existing problems. Educating tomorrow's information technology specialists in the fundamental concepts of extracting knowledge from data should be a

goal for educational institutions. Data mining is a powerful tool for discovering important information, and information is the fuel powering the digital economy.

References

1. Witten, I., Frank E. (2000). *Data Mining*. Academic Press.
2. Sullivan, T. (2000) “EMC's Ruettgers warns of data explosion.” InfoWorld.com http://www.infoworld.com/articles/hn/xml/00/10/04/001004hnruettgers.xml?sponsor=STORAGE
3. Flies, D. (2001) “Designing & Implementing a Classroom Data Warehouse.” MICS.
4. University of Waikato. “Weka 3 – Machine Learning Software in Java.” http://www.cs.waikato.ac.nz/ml/weka/
5. Knowledge Seeker Data Mining Tool. “Angoss Home Page.” http://www.angoss.com/
6. U. S. Forest Service. (1992). “The Eastwide Forest Inventory Database User’s Manual.” http://srsfia.usfs.msstate.edu/ewman.htm
7. Whalin, P. (2000) “Data Mining Research.”
8. Ludwig, L., Flies, D., Wilson, A. (2000) “Data Mining Techniques Applied to the Relationship of Latitude and the Lifespan of Aspen Trees” http://epoxy.mrs.umn.edu ~ludwigl/datamining/research.pdf

Data mining is one component of the exciting area of machine learning and adaptive computation. The goal of building computer systems that can adapt to their envirionments and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics, neuroscience, and cognitive science.Â From a teaching viewpoint the text is intended for undergraduate students at the senior (final year) level, or first or second-year graduate level, who wish to learn about the basic principles of data mining. The text should also be of value to researchers and practitioners who are interested in gaining a better understanding of data mining methods and techniques. Data Mining Specialists are in demand. Learn about a career as a data mining specialist including salary information and steps to get started.Â Obtaining an advanced degree will likely have a positive effect on your salary, as well as keep you at the forefront of new technologies. Regardless of the degree you hold, youâ€™ll need to continue pursuing classes in data science advancements for the entirety of your career. Step 4: Get hired as a data mining specialist. You can find positions as a data mining specialist in many different industries. You may want to begin your career as a data mining specialist with a company that provides opportunities to contribute to a team working at the forefront of data science. Data mining is the process that helps in extracting information from a given data set to identify trends, patterns, and useful data. The objective of using data mining is to make data-supported decisions from enormous data sets. Data mining works in conjunction with predictive analysis, a branch of statistical science that uses complex algorithms designed to work with a special group of problems.Â Data Generalization: Here, the data gets generalized by replacing any low-level data with higher-level conceptualizations. Data Normalization: Here, data is defined in set ranges. Data Attribute Construction: The data sets are required to be in the set of attributes before data mining.