Exploring Practical Data Mining Techniques at Undergraduate Level

ERIC P. JIANG
University of San Diego
5998 Alcala Park, San Diego, CA 92110
UNITED STATES OF AMERICA
jiang@sandiego.edu

Abstract: Data mining is referred to as the process of analyzing and extracting patterns embedded in large amounts of data by using various methods from machine learning, pattern recognition, statistics and database management. With the rapid proliferation of the Internet and advances of computing technology, data mining has become an increasingly important tool of transforming large quantities of digital data into previously unknown and meaningful information and has been applied in many areas that include business and finance, health care, telecommunication, science and higher education. Data mining is also a relatively new field of computer science, and there are only a few undergraduate data mining courses are currently offered in institutions of higher education. In this paper, we describe the design, implementation and evaluation of a data mining course that we have developed and offered as an undergraduate computer science upper-division elective at the University of San Diego. The course combines lectures on a number of key data mining principles and applications, mini student lecture sessions, programming projects and research activities to engage students in active learning. Our experience has shown that data mining can be taught successfully at the undergraduate level and students can learn a great deal of data mining techniques and are able to apply them to solve many real world problems.

Key-Words: Data mining, machine learning, computer science curriculum, classification, clustering, association analysis, anomaly detection

1 Introduction

Data mining is a relatively new field of computer science and it represents an interdisciplinary process of extracting and analysing patterns embedded in large amounts of data by using various methods from machine learning, statistics and database management. With the rapid proliferation of the Internet and advances of computing technology, data mining has become an increasingly important tool of transforming large quantities of digital data into previously unknown and meaningful information and has been used in many areas ranging from business and finance, health care, telecommunication, to science, engineering and higher education. Many computer science programs offer a graduate data mining course but only a few of them offer such a course at the undergraduate level.

Recently we have developed and taught a data mining course as an undergraduate upper-division computer science elective at the University of San Diego. It was a very successful academic experience.

There were several motives for the development of the course. First, as aforementioned, data mining has become one of the most important and promising fields of computer science and has so many applications in this digital information age. Data mining is also an area of active research and the field offers many great employment opportunities. A course that introduces data mining and applications helps prepare our students for graduate study or for a variety of professional careers. Secondly, since data mining has been serving as a key analytical modelling tool of many real world problems, offering a data mining course helps facilitate undergraduate research. There are many data related applications that can be formulated into appropriate undergraduate research projects. Over the past years we have worked with students in various research projects in information retrieval, off-topic search, neural networks, Web link analysis and mining, and students would generally need to spend first few weeks of the projects to learn basic data mining concepts, algorithms and performance evaluation metrics. In addition, our program now requires all graduating seniors to do a capstone project. A data mining course is definitely beneficial to the students interested in learning data analysis and conducting data mining based research projects. Lastly, a well-designed data mining course provides students with a great and integrated opportunity to apply what they have learned from other (previous) computer science courses such as data structures, algorithm analysis and database management to address many practical issues. For instance, students can apply their knowledge of algorithm analysis to understand the practical need of reducing the number of candidate item-sets in the context of association analysis and can also use their knowledge of advanced data structures to follow the idea of partitioning and storing candidate item-sets in a hash tree.

For us, there were also a number of challenges in designing an appropriate undergraduate data mining course, in particular for a group of students with diverse backgrounds. We aimed to develop the course as a practically useful elective to attract more computer science students and to reinforce the course material that students have learned elsewhere. Since we have a small program, the course prerequisites needed to be somewhat minimal in order to have a sufficient enrolment for the course. In theory, it would make sense to require the students who are enrolled in the class to have some background in statistics and linear algebra. However, this requirement would not be easy to be satisfied in our case, as these courses are currently not required in the curriculum. To make up for that, we have added a review on the background material to the class lectures. As the result, we require a sophomore data structures course as the only prerequisite for the data mining course. When the course was offered in the spring of 2011 for the first time, we had a total of 11 students enrolled in the class and among them, there were eight computer science majors, two computer science minors and one foreign exchange student who majors in system engineering.

Selecting a good textbook for this undergraduate course wasn't an easy task as well. Many data mining textbooks we have reviewed either assume a quite strong mathematical background or business-oriented and do not contain enough technical content. After reviewing multiple books, we selected the book by Tan, et. al [8] as our textbook, which we felt is a relatively suitable choice for such a course because it is written with a modest prerequisite in statistics or mathematics. In addition, the book offers a comprehensive introduction to data mining and its major chapters on the key topics, namely, classification, association analysis and clustering, are self-contained and the order in which topics can be covered is also flexible.

2 Related Work

Over the recent years, several undergraduate data mining courses have been developed and also reported in computer science educational journals and conferences. As an example, a data mining course based on the Weka data mining software package and the Weka book [10] was discussed in [5]. The course emphasizes on hands-on experience with various data mining algorithms provided by Weka. Some of the other reports have proposed data mining courses for both undergraduate and graduate students ([3][7]). These courses were planned and implemented to adequately accommodate the needs and backgrounds of the two different groups of students. Another interesting data mining course for undergraduate was discussed in [6] and the course uses a set of research publications as its primary reading material. This approach would allow students to gain better understanding of data mining research and the intricacies of data mining algorithms. But it may or may not work well for students with considerably diverse backgrounds, which often is the case in our program.

The course we describe in this paper focuses on motivating and engaging students in learning data mining methodologies and applications. By requiring students to read relevant material and research papers and then present their findings in student lecture sessions, implement some of data mining algorithms and complete a final research project, the course allows students to better understand data mining key ideas and algorithms and be able to to identify problems that can be potentially solved by data mining techniques.

3 Course Overall Structure

The data mining course we developed is a 3-unit computer science upper-division elective. The course combines lectures on key ideas and algorithms in data mining, mini student lecture sessions on supplemental material, homework and programming projects in some learning algorithms, in-class quizzes and a midterm exam, and a final research project.

3.1 Lecture Topics

Data mining is a multidisciplinary field that straddle multiple areas of computer science and the amount of material can be covered in a undergraduate course is unfortunately limited. We had to choose some most popular and practically useful data mining concepts and algorithms for the class.

3.1.1 Overview of Data Mining, Datasets and Data Preprocessing

In this part of lecture, we introduced to students the concepts, tasks, challenges and applications of data mining, with an emphasis on the importance and relevance of this information technology tool to many critical problems in business, finance, healthcare, science and engineering. This helps motivate students to learn the course contents and identify potential applications of individual algorithms covered later in the class. Then we covered the concepts and properties of data, and the general procedures for preprocessing and visualizing data.

3.1.2 Classification

Classification, a task of assigning objects to one of the predefined categories, is a pervasive problem that encompasses many diverse applications. For this widespread data mining domain, we covered decision tree induction algorithms, nearest neighbour classifiers, naïve Bayes classifier, and briefly introduced perceptron neural networks and support vector machines. We spent a great deal of time in discussing the construction of decision trees including various measures for selecting the best split in a tree node, and we believe that this topic coverage would be appropriate if we assume the audience has the knowledge of tree data structures. Naïve Bayes is another classification method we covered extensively in the class. It is a probabilistic learning model that can be implemented very efficiently with a linear complexity. It should be pointed out that most of our students in the class didn't know much about probability and never heard about the Bayes theorem. In order to present the naïve Bayes algorithm in an understandable manner, we first added a lecture on basic probability and rules, and then gently introduced the well-known Bayes theorem. As is the case with naïve Bayes classifier, the naïve Bayes theorem has so many practical applications. We feel the time we used to cover this probabilistic modelling tool was well spent as it likely provided the only opportunity for many of our computer science students for knowing one of the most popularly used probability theorems.

3.1.3 Clustering

Clustering or cluster analysis partitions data into meaningful groups and has been widely applied in biology, business, psychology, medicine and climate research. For the course, we covered both partitional and hierarchical clustering algorithms with an emphasis on the former type. That included the famous k-means algorithm and its several variants such as bisecting k-means, basic agglomerative hierarchical clustering algorithm and a representative density-based algorithm (DBSCAN).

3.1.4 Association Analysis

Association analysis, also known as market basket analysis, is a methodology for discovering interesting relationships that may be hidden in large datasets. After introducing a few simple examples of association rules, students would understand that this analysis tool has a variety of business-related applications such as marketing promotions, inventory management and customer relationship management. The primary algorithm we covered in this category for the course is the most popular Apriori model. The model should be well connected to some data structures (e.g., trees, hashing, etc.) and require some background in algorithm complexity analysis.

3.1.5 Anomaly Detection

The goal of anomaly (or outlier) detection is to find objects that are significantly different from most others. It can be used in fraud and intrusion detection, public health and medicine. There is a variety of anomaly detection approaches and for the course, we provided an overview of algorithms that are statistical based, proximity based and model based.

3.2 Assignments and Exams

To help students better understand course material, there are several assignments associated with the course: homework, book and paper reading, and programming projects. Homework problems closely tied to some of the learning algorithms were assigned on a regular basis. All the assignments were collected and graded by the instructor. Programming projects that implement some of the data mining algorithms were also required for the course and they would surely help to learn in depth the intricacies underlying the algorithms. Due to different backgrounds among students in the class, we selected the implementations of the k-nearest neighbour classifier and the bisecting k-means clustering algorithm as student programming projects and used some datasets from the popular UCI Machine Learning Repository [9] for project model training and testing. For the projects, students could use any modern programming languages of their choice.

Several selective research papers were used in the course as additional reading assignments to help students gain first hand exposure to real data mining research. For instance, a paper that applies neural networks and association rules to classify medical images for tumour detection [1] was assigned in the middle of the semester as a supplemental reading material. The paper is well written and quite readable to undergraduate students who have had some basic training in data mining. The major reading assignment throughout the entire semester is a supplemental book, Super Crunchers [2] and we will describe it in detail in the next subsection.

There were a number of quizzes and one midterm exam that were designed to evaluate the students on course material covered in the class, in particular their understanding of data mining ideas and basic structures of various learning algorithms.

3.3 Mini Student Lecture Sessions

We developed this special learning component for the purpose of stimulating interest, sharing ideas and motivating students to learn course contents. We used a New York Times bestseller, Super Crunchers, as the major supplemental reading material for the course that was discussed periodically in our mini student lecture sessions. More specifically, we partitioned the book chapters among the students in the class; one or two students for a chapter, depending upon the length of the chapter. In the order of the book chapters, the students, acting as guest speakers, read and prepared and then lectured their assigned chapters to the class. We arranged one student lecture session for each week of the semester and each session took about 15 minutes.

Super Crunchers is a book about data mining and statistical analysis. The book does not cover complex data mining techniques but it beautifully and successfully introduces the basic ideas behind data mining for a general audience. In particular, the book presents many small yet very interesting real stories about people crunching data and shows how crunching numbers really help people make intelligent decisions.

We believe that the student lecture sessions specifically planned for this data mining course was a successful learning experience to all students, and by actively engaging the students in discussing the reading material, we would expect them to develop a broader interest in learning data mining and a deeper insight into its applications in real life situations.

3.4 Final Research Project

For the course final research project, which was assigned in the last month of the semester, students can develop a data mining tool, solve a real world problem using data mining techniques, or implement an interested learning algorithm from a recent research paper and demonstrate its effectiveness. Students could choose to do the project as solo one or as a team project of two students. Since this course was counted as a computer science upper-division elective, some programming implementation was strongly expected for the project and this can be done in any programming languages of student's choice. In addition, students were required to present their findings from the project at the end of the semester.

For the course, the project served as an important research activity where students can gain better understanding of the course material and deeper insights into the challenges faced by data miners. It also provided an opportunity for students to learn from each other about some latest advances in data mining research and applications that might not have been covered in lectures and to share ideas and implementation techniques with fellow students. Below are some of the final research projects, proposed by students, from our course in the spring of 2011:

Multi-dimensional scaling for data visualization

Data mining cup 2011: an e-commerce recommendation system

Mining Web navigation patterns with a path traversal graph

Improved probabilistic C-means clustering algorithms

Spam email detection and text classification

Data mining for public education funding decision

We would like to point out that two students in the class chose to participate in the Data Mining Cup (DMC) competition 2011 [4] as their final project. DMC is an annual event attracting many college student teams from over 40 countries and it aims to promote intelligent data analysis. The challenge of that year was to develop an efficient product recommendation system that predicts the products customers would be interested in based on over 9.5 million customer transactional data. Our students were able to successfully apply various association analysis ideas and data preprocessing schemes to the given task and they actually achieved the 2nd place in the contest

4 Conclusion

This paper has presented an academic experience of developing a data mining course for undergraduate students. The course was offered in the spring of 2011 as a computer science upper-division for the first time in our program. The course combines lectures on key data mining principles and applications, mini student lecture sessions, programming projects and research activities to engage students in active learning. Our experience has shown that, with carefully planned course material and learning activities, data mining can be taught successfully at the undergraduate level and students can learn a great deal of data mining techniques and are capable of applying them to many real world applications. As jobs in data mining and related fields have been growing rapidly over recent years, a course in data mining can help prepare our students for many great career opportunities. In fact, one student from the data mining class had a job interview with Google in late April of 2011. He emailed us right after the interview and stated "… my data mining knowledge might have sealed-the-deal to a job offer [with Google] …" and he actually did it.

References:

[1] Antonie M, Zaiane, O. and Coman, A., Application of Data Mining Techniques for Medical Image Classification, *Proceedings of the 2nd International Workshop on Multimedia Data Mining,* 2001.

[2] Ayres, I., *Super Crunchers,* Random House, 2008.

[3] Chawla, N., Teaching Data Mining by Coalescing Theory and Applications, *Proceedings of the 35th ASEE/IEEE Frontiers in Education Conference*, 2005.

[4] Data Mining Cup 2011, *http://www.data-mining-cup.de/en/dmc-competition*.

[5] Lopez, D. and Ludwig, L, Data Mining at the Undergraduate Level, *Proceedings of the Midwest Instruction and Computing Symposium,* 2001.

[6] Musicant, D., A Data Mining Course for Computer Science: Primary Sources and Implementations*, Proceedings of SIGCSE,* 2006.

[7] Sequer, J., A Data Mining Course for Computer Science and Non-Computer Science Students, *Journal of Computer Sciences in Colleges*, Vol.22, No.4, 2007, pp. 109-114.

[8] Tan, P., Steinbach, M. and Kumar, V., *Introduction to Data Mining*, Addison Wesley, 2006.

[9] University of California at Irvine Machine Learning Repository, *http://archive.ics.uci.edu/ml/.*

[10] Witten, I. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2nd Edition, 2005.