David L. Olson
Dursun Delen

Advanced Data Mining Techniques

Springer

David L. Olson · Dursun Delen

Advanced Data Mining Techniques

Springer

Dr. David L. Olson
Department of Management Science
University of Nebraska
Lincoln, NE 68588-0491
USA
dolson3@unl.edu

Dr. Dursun Delen
Department of Management
Science and Information Systems
700 North Greenwood Avenue
Tulsa, Oklahoma 74106
USA
dursun.delen@okstate.edu

ISBN: 978-3-540-76916-3 e-ISBN: 978-3-540-76917-0

Library of Congress Control Number: 2007940052

Cover design: WMX Design, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Advanced Data Mining Techniques

I dedicate this book to my grandchildren.

David L. Olson

I dedicate this book to my children, Altug and Serra.

Dursun Delen

Preface

The intent of this book is to describe some recent data mining tools that have proven effective in dealing with data sets which often involve uncertain description or other complexities that cause difficulty for the conventional approaches of logistic regression, neural network models, and decision trees. Among these traditional algorithms, neural network models often have a relative advantage when data is complex. We will discuss methods with simple examples, review applications, and evaluate relative advantages of several contemporary methods.

Book Concept

Our intent is to cover the fundamental concepts of data mining, to demonstrate the potential of gathering large sets of data, and analyzing these data sets to gain useful business understanding. We have organized the material into three parts. Part I introduces concepts. Part II contains chapters on a number of different techniques often used in data mining. Part III focuses on business applications of data mining. Not all of these chapters need to be covered, and their sequence could be varied at instructor design.

The book will include short vignettes of how specific concepts have been applied in real practice. A series of representative data sets will be generated to demonstrate specific methods and concepts. References to data mining software and sites such as www.kdnuggets.com will be provided.

Part I: Introduction

Chapter 1 gives an overview of data mining, and provides a description of the data mining process. An overview of useful business applications is provided.

Chapter 2 presents the data mining process in more detail. It demonstrates this process with a typical set of data. Visualization of data through data mining software is addressed.

Part II: Data Mining Methods as Tools

Chapter 3 presents memory-based reasoning methods of data mining. Major real applications are described. Algorithms are demonstrated with prototypical data based on real applications.

Chapter 4 discusses association rule methods. Application in the form of market basket analysis is discussed. A real data set is described, and a simplified version used to demonstrate association rule methods.

Chapter 5 presents fuzzy data mining approaches. Fuzzy decision tree approaches are described, as well as fuzzy association rule applications. Real data mining applications are described and demonstrated

Chapter 6 presents Rough Sets, a recently popularized data mining method.

Chapter 7 describes support vector machines and the types of data sets in which they seem to have relative advantage.

Chapter 8 discusses the use of genetic algorithms to supplement various data mining operations.

Chapter 9 describes methods to evaluate models in the process of data mining.

Part III: Applications

Chapter 10 presents a spectrum of successful applications of the data mining techniques, focusing on the value of these analyses to business decision making.

University of Nebraska-Lincoln David L. Olson

Oklahoma State University Dursun Delen

Contents

Part I INTRODUCTION

Part II DATA MINING METHODS AS TOOLS

Part III APPLICATIONS

Part I
INTRODUCTION

1 Introduction

Data mining refers to the analysis of the large quantities of data that are stored in computers. For example, grocery stores have large amounts of data generated by our purchases. Bar coding has made checkout very convenient for us, and provides retail establishments with masses of data. Grocery stores and other retail stores are able to quickly process our purchases, and use computers to accurately determine product prices. These same computers can help the stores with their inventory management, by instantaneously determining the quantity of items of each product on hand. They are also able to apply computer technology to contact their vendors so that they do not run out of the things that we want to purchase. Computers allow the store's accounting system to more accurately measure costs, and determine the profit that store stockholders are concerned about. All of this information is available based upon the bar coding information attached to each product. Along with many other sources of information, information gathered through bar coding can be used for data mining analysis.

Data mining is not limited to business. Both major parties in the 2004 U.S. election utilized data mining of potential voters.[1] Data mining has been heavily used in the medical field, to include diagnosis of patient records to help identify best practices.[2] The Mayo Clinic worked with IBM to develop an online computer system to identify how that last 100 Mayo patients with the same gender, age, and medical history had responded to particular treatments.[3]

Data mining is widely used by banking firms in soliciting credit card customers,[4] by insurance and telecommunication companies in detecting

[1] H. Havenstein (2006). IT efforts to help determine election successes, failures: Dems deploy data tools; GOP expands microtargeting use, *Computerworld* 40: 45, 11 Sep 2006, 1, 16.

[2] T.G. Roche (2006). Expect increased adoption rates of certain types of EHRs, EMRs, *Managed Healthcare Executive* 16:4, 58.

[3] N. Swartz (2004). IBM, Mayo clinic to mine medical data, *The Information Management Journal* 38:6, Nov/Dec 2004, 8.

[4] S.-S. Weng, R.-K. Chiu, B.-J. Wang, S.-H. Su (2006/2007). The study and verification of mathematical modeling for customer purchasing behavior, *Journal of Computer Information Systems* 47:2, 46–57.

fraud,[5] by telephone companies and credit card issuers in identifying those potential customers most likely to churn,[6] by manufacturing firms in quality control,[7] and many other applications. Data mining is being applied to improve food and drug product safety,[8] and detection of terrorists or criminals.[9] Data mining involves statistical and/or artificial intelligence analysis, usually applied to large-scale data sets. Traditional statistical analysis involves an approach that is usually directed, in that a specific set of expected outcomes exists. This approach is referred to as *supervised* (hypothesis development and testing). However, there is more to data mining than the technical tools used. Data mining involves a spirit of knowledge discovery (learning new and useful things). Knowledge discovery is referred to as *unsupervised* (knowledge discovery) Much of this can be accomplished through automatic means, as we will see in decision tree analysis, for example. But data mining is not limited to automated analysis. Knowledge discovery by humans can be enhanced by graphical tools and identification of unexpected patterns through a combination of human and computer interaction.

Data mining can be used by businesses in many ways. Three examples are:

1. *Customer profiling*, identifying those subsets of customers most profitable to the business;
2. *Targeting*, determining the characteristics of profitable customers who have been captured by competitors;
3. *Market-basket analysis*, determining product purchases by consumer, which can be used for product positioning and for cross-selling.

These are not the only applications of data mining, but are three important applications useful to businesses.

[5] R.M. Rejesus, B.B. Little, A.C. Lovell (2004). Using data mining to detect crop insurance fraud: Is there a role for social scientists? *Journal of Financial Crime* 12:1, 24–32.

[6] G.S. Linoff (2004). Survival data mining for customer insight, *Intelligent Enterprise* 7:12, 28–33.

[7] C. Da Cunha, B. Agard, A. Kusiak (2006). Data mining for improvement of product quality, *International Journal of Production Research* 44:18/19, 4041–4054.

[8] M. O'Connell (2006). Drug safety, the U.S. Food and Drug Administration and statistical data mining, *Scientific Computing* 23:7, 32–33.

[9] ___., Data mining: Early attention to privacy in developing a key DHS program could reduce risks, *GAO Report 07-293*, 3/21/2007.

What is Data Mining?

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic-specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. However, systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data.

A variety of analytic computer models have been used in data mining. The standard model types in data mining include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees. These techniques are well known. This book focuses on less used techniques applied to specific problem types, to include association rules for initial data exploration, fuzzy data mining approaches, rough set models, support vector machines, and genetic algorithms. The book will also review some interesting applications in business, and conclude with a comparison of methods.

But these methods are not the only tools available for data mining. Work has continued in a number of areas, which we will describe in this book. This new work is generated because we generate ever larger data sets, express data in more complete terms, and deal with more complex forms of data. Association rules deal with large scale data sets such as those generated each day by retail organizations such as groceries. Association rules seek to identify what things go together. Research continues to enable more accurate identification of relationships when coping with massive data sets. Fuzzy representation is a way to more completely describe the uncertainty associated with concepts. Rough sets is a way to express this uncertainty in a specific probabilistic form. Support vector machines offer a way to separate data more reliably when certain forms of complexity are present in data sets. And genetic algorithms help identify better solutions for data that is in a particular form. All of these topics have interesting developments that we will try to demonstrate.

What is Needed to Do Data Mining

Data mining requires identification of a problem, along with collection of data that can lead to better understanding, and computer models to provide statistical or other means of analysis. This may be supported by visualization tools, that display data, or through fundamental statistical analysis, such as correlation analysis.

Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation. Versatile refers to the ability of the tool to apply a wide variety of models. Scalable tools imply that if the tools works on a small data set, it should also work on larger data sets. Automation is useful, but its application is relative. Some analytic functions are often automated, but human setup prior to implementing procedures is required. In fact, analyst judgment is critical to successful implementation of data mining. Proper selection of data to include in searches is critical. Data transformation also is often required. Too many variables produce too much output, while too few can overlook key relationships in the data. Fundamental understanding of statistical concepts is mandatory for successful data mining.

Data mining is expanding rapidly, with many benefits to business. Two of the most profitable application areas have been the use of customer segmentation by marketing organizations to identify those with marginally greater probabilities of responding to different forms of marketing media, and banks using data mining to more accurately predict the likelihood of people to respond to offers of different services offered. Many companies are using this technology to identify their blue-chip customers so that they can provide them the service needed to retain them.[10]

The casino business has also adopted data warehousing and data mining. Harrah's Entertainment Inc. is one of many casino organizations who use incentive programs.[11] About 8 million customers hold Total Gold cards, which are used whenever the customer plays at the casino, or eats, or stays, or spends money in other ways. Points accumulated can be used for complementary meals and lodging. More points are awarded for activities which provide Harrah's more profit. The information obtained is sent to the firm's corporate database, where it is retained for several years. Trump's Taj Card is used in a similar fashion. Recently, high competition has led to the use of data mining. Instead of advertising the loosest slots in town, Bellagio and Mandalay Bay have developed the strategy of promoting luxury visits. Data mining is used to identify high rollers, so that these valued customers can be cultivated. Data warehouses enable casinos to estimate the lifetime value of players. Incentive travel programs, in-house

[10] R. Hendler, F. Hendler (2004). Revenue management in fabulous Las Vegas: Combining customer relationship management and revenue management to maximize profitability, *Journal of Revenue & Pricing Management* 3:1, 73–79.

[11] G. Loveman (2003). Diamonds in the data mine, *Harvard Business Review* 81:5, 109–113.