DISCOVERING KNOWLEDGE IN DATA

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

DANIEL T. LAROSE
Director of Data Mining
Central Connecticut State University

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.
Discovering knowledge in data : an introduction to data mining / Daniel T. Larose
p. cm.
Includes bibliographical references and index.
ISBN 0-471-66657-2 (cloth)
1. Data mining. I. Title.
QA76.9.D343L38 2005
006.3′12—dc22 2004003680

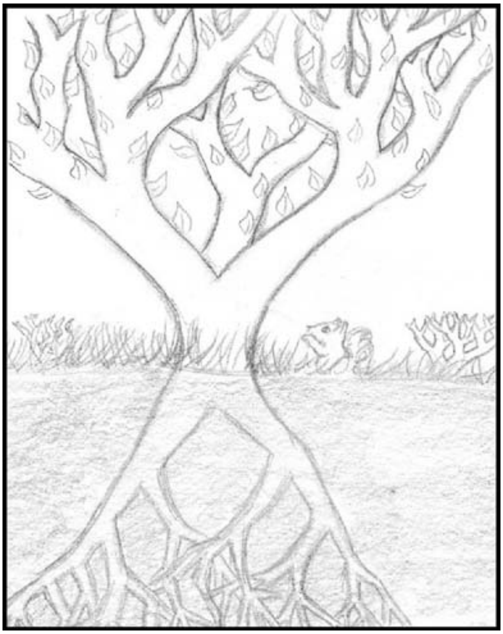
Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Dedication

To my parents,
And their parents,
And so on...

For my children,
And their children,
And so on...



© 2004 Chantal Larose

CONTENTS

PREFACE

WHAT IS DATA MINING?

Data mining is predicted to be "one of the most revolutionary developments of the next decade," according to the online technology magazine *ZDNET News* (February 8, 2001). In fact, the *MIT Technology Review* chose data mining as one of ten emerging technologies that will change the world. According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

Because data mining represents such an important field, Wiley-Interscience and Dr. Daniel T. Larose have teamed up to publish a series of volumes on data mining, consisting initially of three volumes. The first volume in the series, *Discovering Knowledge in Data: An Introduction to Data Mining*, introduces the reader to this rapidly growing field of data mining.

WHY IS THIS BOOK NEEDED?

Human beings are inundated with data in most fields. Unfortunately, these valuable data, which cost firms millions to collect and collate, are languishing in warehouses and repositories. *The problem is that not enough trained human analysts are available who are skilled at translating all of the data into knowledge*, and thence up the taxonomy tree into wisdom. This is why this book is needed; it provides readers with:

- Models and techniques to uncover hidden nuggets of information
- Insight into how data mining algorithms work
- The experience of actually performing data mining on large data sets

Data mining is becoming more widespread every day, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies that do not apply these techniques are in danger of falling behind and losing market share, because their competitors are using data mining and are thereby gaining the competitive edge. In *Discovering Knowledge in Data*, the step-by-step hands-on solutions of real-world business problems using widely available data mining techniques applied to real-world data sets will appeal to managers, CIOs, CEOs, CFOs, and others who need to keep abreast of the latest methods for enhancing return on investment.

DANGER! DATA MINING IS EASY TO DO BADLY

The plethora of new off-the-shelf software platforms for performing data mining has kindled a new kind of danger. The ease with which these GUI-based applications can manipulate data, combined with the power of the formidable data mining algorithms embedded in the black-box software currently available, make their misuse proportionally more hazardous.

Just as with any new information technology, *data mining is easy to do badly*. A little knowledge is especially dangerous when it comes to applying powerful models based on large data sets. For example, analyses carried out on unpreprocessed data can lead to erroneous conclusions, or inappropriate analysis may be applied to data sets that call for a completely different approach, or models may be derived that are built upon wholly specious assumptions. If deployed, these errors in analysis can lead to very expensive failures.

"WHITE BOX" APPROACH: UNDERSTANDING THE UNDERLYING ALGORITHMIC AND MODEL STRUCTURES

The best way to avoid these costly errors, which stem from a blind black-box approach to data mining, is to apply instead a "white-box" methodology, which emphasizes an understanding of the algorithmic and statistical model structures underlying the software. *Discovering Knowledge in Data* applies this white-box approach by:

- Walking the reader through the various algorithms
- Providing examples of the operation of the algorithm on actual large data sets
- Testing the reader's level of understanding of the concepts and algorithms
- Providing an opportunity for the reader to do some real data mining on large data sets

Algorithm Walk-Throughs

Discovering Knowledge in Data walks the reader through the operations and nuances of the various algorithms, using small-sample data sets, so that the reader gets a true appreciation of what is really going on inside the algorithm. For example, in Chapter 8, we see the updated cluster centers being updated, moving toward the center of their respective clusters. Also, in Chapter 9 we see just which type of network weights will result in a particular network node "winning" a particular record.

Applications of the Algorithms to Large Data Sets

Discovering Knowledge in Data provides examples of the application of various algorithms on actual large data sets. For example, in Chapter 7 a classification problem is attacked using a neural network model on a real-world data set. The resulting neural network topology is examined along with the network connection weights, as reported by the software. These data sets are included at the book series Web site, so that readers may follow the analytical steps on their own, using data mining software of their choice.

Chapter Exercises: Checking to Make Sure That You Understand It

Discovering Knowledge in Data includes over 90 chapter exercises, which allow readers to assess their depth of understanding of the material, as well as to have a little fun playing with numbers and data. These include conceptual exercises, which help to clarify some of the more challenging concepts in data mining, and "tiny data set" exercises, which challenge the reader to apply the particular data mining algorithm to a small data set and, step by step, to arrive at a computationally sound solution. For example, in Chapter 6 readers are provided with a small data set and asked to construct by hand, using the methods shown in the chapter, a C4.5 decision tree model, as well as a classification and regression tree model, and to compare the benefits and drawbacks of each.

Hands-on Analysis: Learn Data Mining by Doing Data Mining

Chapters 2 to 4 and 6 to 11 provide the reader with hands-on analysis problems, representing an opportunity for the reader to apply his or her newly acquired data mining expertise to solving real problems using large data sets. Many people learn by doing. *Discovering Knowledge in Data* provides a framework by which the reader can learn data mining by doing data mining. The intention is to mirror the real-world data mining scenario. In the real world, dirty data sets need cleaning; raw data needs to be normalized; outliers need to be checked. So it is with *Discovering Knowledge in Data*, where over 70 hands-on analysis problems are provided. In this way, the reader can "ramp up" quickly and be "up and running" his or her own data mining analyses relatively shortly.

For example, in Chapter 10 readers are challenged to uncover high-confidence, high-support rules for predicting which customer will be leaving a company's service. In Chapter 11 readers are asked to produce lift charts and gains charts for a set of classification models using a large data set, so that the best model may be identified.

DATA MINING AS A PROCESS

One of the fallacies associated with data mining implementation is that data mining somehow represents an isolated set of tools, to be applied by some aloof analysis department, and is related only inconsequentially to the mainstream business or research endeavor. Organizations that attempt to implement data mining in this way will see their chances of success greatly reduced. This is because data mining should be view as a *process*.

Discovering Knowledge in Data presents data mining as a well-structured *standard process*, intimately connected with managers, decision makers, and those involved in deploying the results. Thus, this book is not only for analysts but also for managers, who need to be able to communicate in the language of data mining. The particular standard process used is the CRISP–DM framework: the Cross-Industry Standard Process for Data Mining. CRISP–DM demands that data mining be seen as an entire process, from communication of the business problem through data collection and management, data preprocessing, model building, model evaluation, and finally, model deployment. Therefore, this book is not only for analysts and managers but also for data management professionals, database analysts, and decision makers.

GRAPHICAL APPROACH, EMPHASIZING EXPLORATORY DATA ANALYSIS

Discovering Knowledge in Data emphasizes a graphical approach to data analysis. There are more than 80 screen shots of actual computer output throughout the book, and over 30 other figures. Exploratory data analysis (EDA) represents an interesting and exciting way to "feel your way" through large data sets. Using graphical and numerical summaries, the analyst gradually sheds light on the complex relationships hidden within the data. *Discovering Knowledge in Data* emphasizes an EDA approach to data mining, which goes hand in hand with the overall graphical approach.

HOW THE BOOK IS STRUCTURED

Discovering Knowledge in Data provides a comprehensive introduction to the field. Case studies are provided showing how data mining has been utilized successfully (and not so successfully). Common myths about data mining are debunked, and common pitfalls are flagged, so that new data miners do not have to learn these lessons themselves.

The first three chapters introduce and follow the CRISP–DM standard process, especially the data preparation phase and data understanding phase. The next seven chapters represent the heart of the book and are associated with the CRISP–DM modeling phase. Each chapter presents data mining methods and techniques for a specific data mining task.

- Chapters 5, 6, and 7 relate to the *classification* task, examining the *k*-nearest neighbor (Chapter 5), decision tree (Chapter 6), and neural network (Chapter 7) algorithms.
- Chapters 8 and 9 investigate the *clustering* task, with hierarchical and *k*-means clustering (Chapter 8) and Kohonen network (Chapter 9) algorithms.
- Chapter 10 handles the *association* task, examining association rules through the a priori and GRI algorithms.
- Finally, Chapter 11 covers model evaluation techniques, which belong to the CRISP–DM evaluation phase.

DISCOVERING KNOWLEDGE IN DATA AS A TEXTBOOK

Discovering Knowledge in Data naturally fits the role of textbook for an introductory course in data mining. Instructors may appreciate:

- The presentation of data mining as a *process*
- The "white-box" approach, emphasizing an understanding of the underlying algorithmic structures:
  - ○ algorithm walk-throughs
  - ○ application of the algorithms to large data sets
  - ○ chapter exercises
  - ○ hands-on analysis
- The graphical approach, emphasizing exploratory data analysis
- The logical presentation, flowing naturally from the CRISP–DM standard process and the set of data mining tasks

Discovering Knowledge in Data is appropriate for advanced undergraduate or graduate courses. Except for one section in Chapter 7, no calculus is required. An introductory statistics course would be nice but is not required. No computer programming or database expertise is required.

ACKNOWLEDGMENTS

Discovering Knowledge in Data would have remained unwritten without the assistance of Val Moliere, editor, Kirsten Rohsted, editorial program coordinator, and Rosalyn Farkas, production editor, at Wiley-Interscience and Barbara Zeiders, who copyedited the work. Thank you for your guidance and perserverance.

I wish also to thank Dr. Chun Jin and Dr. Daniel S. Miller, my colleagues in the Master of Science in Data Mining program at Central Connecticut State University; Dr. Timothy Craine, the chair of the Department of Mathematical Sciences; Dr. Dipak K. Dey, chair of the Department of Statistics at the University of Connecticut; and Dr. John Judge, chair of the Department of Mathematics at Westfield State College. Your support was (and is) invaluable.

Thanks to my children, Chantal, Tristan, and Ravel, for sharing the computer with me. Finally, I would like to thank my wonderful wife, Debra J. Larose, for her patience, understanding, and proofreading skills. But words cannot express....

Daniel T. Larose, Ph.D.
Director, Data Mining @CCSU
www.ccsu.edu/datamining