Exploratory Data Mining and Data Cleaning

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Exploratory Data Mining and Data Cleaning

TAMRAPARNI DASU

THEODORE JOHNSON

AT&T Labs, Research Division
Florham Park, NJ

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data:
Dasu, Tamraparni.
Exploratory data mining and data cleaning / Tamraparni Dasu, Theorodre Johnson.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-26851-8 (cloth)
1. Data mining. 2. Electronic data processing—Data preparation. 3. Electronic data processing—Quality control. I. Johnson, Theodore. II. Title.
QA76.9.D343 D34 2003
006.3—dc21

2002191085

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

Preface

As data analysts at a large information-intensive business, we often have been asked to analyze new (to us) data sets. This experience was the original motivation for our interest in the topics of exploratory data mining and data quality. Most data mining and analysis techniques assume that the data have been joined into a single table and cleaned, and that the analyst already knows what she or he is looking for. Unfortunately, the data set is usually dirty, composed of many tables, and has unknown properties. Before any results can be produced, the data must be cleaned and explored—often a long and difficult task.

Current books on data mining and analysis usually focus on the last stage of the analysis process (getting the results) and spend little time on how data exploration and cleaning is done. Usually, their primary aim is to discuss the efficient implementation of the data mining algorithms and the interpretation of the results. However, the true challenges in the task of data mining are:

- Creating a data set that contains the relevant and accurate information, and
- Determining the appropriate analysis techniques.

In our experience, the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the ultimate data mining results. Data mining books (a good one is [56]) provide a great amount of detail about the analytical process and advanced data mining techniques. However they assume that the data has already been gathered, cleaned, explored, and understood.

As we gained experience with exploratory data mining and data quality issues, we became involved in projects in which data quality improvement was the goal of the project (i.e., for operational databases) rather than a prerequisite. Several books recently have been published on the topic of ensuring data quality (e.g., the books by Loshin [84], by Redman [107]), and by English [41]). However, these books are written for managers and take a

managerial viewpoint. While the problem of ensuring data quality requires a significant managerial support, there is also a need for technical and analytic tools. At the time of this writing, we have not seen any organized exposition of the technical aspects of data quality management. The most closely related book is Pyle [102], which discusses data preparation for data mining. However, this text has little discussion of data quality issues or of exploratory data mining—pre-requisites even to preparing data for data mining.

Our focus in this book is to develop a systematic process of data exploration and data quality management. We have found these seemingly unrelated topics to be inseparable. The exploratory phase of any data analysis project inevitably involves sorting out data quality problems, and any data quality improvement project inevitably involves data exploration. As a further benefit, data exploration sheds light on appropriate analytic strategies.

Data quality is a notoriously messy problem that refuses to be put into a neat container, and therefore is often viewed as technically intractable. We have found that data quality problems can be addressed, but doing so requires that we draw on methods from many disciplines: statistics, exploratory data mining (EDM), databases, management, and metadata. Our focus in this book is to present an integrated approach to EDM and data quality. Because of the very broad nature of the subject, the exposition tends to be a summarization of material discussed in great detail elsewhere (for which we provide references), with an emphasis on how the techniques relate to each other and to EDM and data quality. Some topics (such as data quality metrics and certain aspects of EDM) have no other good source, so we discuss them in greater detail.

EXPLORATORY DATA MINING (EDM)

Data sets of the twenty-first century are different from the ones that motivated analytical techniques of statistics, machine learning and others. Earlier data sets were reasonably small and relatively homogeneous so that the structure in them could be captured with compact models that had large but a manageable number of parameters. Many researchers have focused on scaling the methods to run efficiently and quickly on the much larger data sets collected by automated devices. In addition, methods have been developed specifically for massive data (i.e., data mining techniques). However, there are two fundamental issues that need to be addressed before these methods can be applied.

- A "data set" is often a patchwork of data collected from many sources, which might not have been designed for integration. One example of this problem is when two corporate entities providing different services to a common customer base merge to become a single entity. Another is when different divisions of a "federation enterprise" need to merge their data

stores. In such situations, approximate matching heuristics are used to combine the data. The resulting patchwork data set will have many data quality issues that need to be addressed. The data are likely to contain many other data glitches, and these need to be treated as well.

- Data mining methods often do not focus on the "appropriateness of the model for the data," namely, goodness-of-fit. While finding the best model in a given class of models is desirable, it is equally important to determine the class of models that best fits the data.

There is no simple or single method for analyzing a complex, unfamiliar data set. The task typically requires the sequential application of disparate techniques, leveraging the additional information acquired at each stage to converge to a powerful, accurate and fast method. The end-product is often a "piecewise technique" where at each stage we might have had to adapt or extend, to improvise on an existing method. The importance of such an approach has been emphasized by statisticians such as John Tukey [123] and more recently in the machine learning community, for instance, in the AutoClass project [19].

DATA QUALITY

A major confounding factor in EDM is the presence of data quality issues. These are often unearthed as "interesting patterns" but on closer examination prove to be artifacts. We emphasize this aspect in our case study, since typically data analysts spend a significant portion of their time weeding-out data quality problems. No matter how sophisticated the data mining techniques, bad data will lead to misleading findings.

While most practitioners of data analysis are aware of the pitfalls of data quality issues, it is only recently that there has been an emphasis on the systematic detection and removal of data problems. There have been efforts directed at managing processes that generate the data, at cleaning up databases (e.g. merging/purging of duplicates), and at finding tools and algorithms for the automatic detection of data glitches. Statistical methods for process control (predominantly univariate) that date back to quality control charts developed for detecting batches of poorly produced lots in industrial manufacturing are often adapted to monitor fluctuations in variables that populate databases.

For operations databases, data quality is an end in itself. Most business (and governmental, etc.) processes involve complex interactions between many databases. Data quality problems can have very expensive manifestations (e.g., "losing" a cross-country cable, forgetting to bill customers). In this electronic age, many businesses (and governmental organizations, etc.) would like to "e-enable" their customers—that is, let them examine the relevant parts of the

operational databases to manage their own accounts. Depending on the state of the underlying databases, this can be embarrassing or even impossible.

SUMMARY

In this book, we intend to:

- Focus on developing a modeling strategy through an iterative data exploration loop and incorporation of domain knowledge;
- Address methods for dealing with data quality issues that can have a significant impact on findings and decisions, using commercially available tools as well as new algorithmic approaches;
- Emphasize application in real-life scenarios throughout the narrative with examples;
- Highlight new approaches and methodologies, such as the *DataSphere* space partitioning and summary-based analysis techniques, and approaches to developing data quality metrics.

The book is intended for serious data analysts everywhere that need to analyze large amounts of unfamiliar, potentially noisy data, and for managers of operations databases. It can also serve as a text on data quality to supplement an advanced undergraduate or graduate level course in large-scale data analysis and data mining. The book is especially appropriate for a cross-disciplinary course in statistics and computer science.

ACKNOWLEDGMENTS

We wish to thank the following people who have contributed to the material in this book: Deepak Agarwal, Dave Belanger, Bob Bell, Simon Byers, Corinna Cortes, Ken Church, Christos Faloutsos, Mary Fernandez, Joel Gottlieb, Andrew Hume, Nick Koudas, Elefteris Koutsofios, Bala Krishnamurthy, Ken Lyons, David Poole, Daryl Pregibon, Matthew Roughan, Gregg Vesonder, and Jon Wright.