WOODHEAD PUBLISHING SERIES IN BIOMEDICINE

DATA MINING FOR BIOINFORMATICS APPLICATIONS

ZENGYOU HE

WP
WOODHEAD
PUBLISHING

Woodhead Publishing Series in Biomedicine: Number 76

Data Mining for Bioinformatics Applications

Zengyou He

ELSEVIER

AMSTERDAM • BOSTON • CAMBRIDGE • HEIDELBERG
LONDON • NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Woodhead Publishing is an imprint of Elsevier

WP
WOODHEAD PUBLISHING

Woodhead Publishing Limited is an imprint of Elsevier
80 High Street, Sawston, Cambridge, CB22 3HJ, UK
225 Wyman Street, Waltham, MA 02451, USA
Langford Lane, Kidlington, OX5 1GB, UK

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

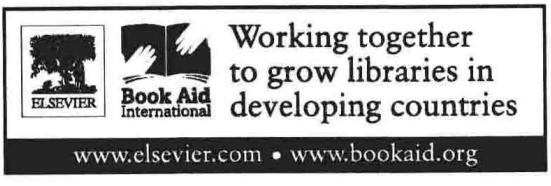
ISBN: 978-0-08-100100-4 (print)
ISBN: 978-0-08-100107-3 (online)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2015934370

For information on all Woodhead Publishing publications visit our website at http://store.elsevier.com/

Printed in the United States of America



Data Mining for Bioinformatics Applications

Related titles

From plant genomics to plant biotechnology
(ISBN: 978-1-907568-29-9)

An introduction to biotechnology
(ISBN: 978-1-907568-28-2)

MATLABs in bioscience and biotechnology
(ISBN: 978-1-907568-04-6)

About the author

Zengyou He is an associate professor in the School of Software at Dalian University of Technology, China.

He received his BS, MS, and PhD degrees in computer science from the Harbin Institute of Technology, China, in 2000, 2002, and 2006, respectively. Prior to becoming an associate professor, he was a research associate in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (2007–2010).

His research interests include computational proteomics and biological data mining. He has published more than 30 papers on leading journals in the field of bioinformatics, including *Bioinformatics*, *BMC Bioinformatics*, *Briefings in Bioinformatics*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, and *Journal of Computational Biology*.

Dedication

Introduction

Data mining methods have been widely used for solving real bioinformatics problems. However, the data mining process is not trivial. It consists of many steps: problem definition, data collection, data preprocessing, modeling, and validation. For each step, different techniques may be applied. Due to the complexity of data mining process and data mining methods, people cannot easily use data mining tools to solve their bioinformatics problems.

In this book, I will use an example-based method to illustrate how to apply data mining techniques to solving real bioinformatics problems. More precisely, I will use six bioinformatics problems that have been investigated in my recent research as examples. For each example, I will describe the entire data mining process, ranging from data preprocessing to modeling and result validation. In addition, I will describe how to use *different* data mining methods to solve the *same* bioinformatics problem in some examples.

In this problem-driven book, I will cover the most commonly used data mining methods, such as frequent pattern mining, discriminative pattern mining, classification, and clustering to show how to select one feasible data mining method to solve a real bioinformatics problem at hand.

Audience

This book will have obvious appeal for a broad audience of computer scientists who are interested in designing new data mining algorithms and biologists who are trying to solve bioinformatics problems using existing data mining tools. To achieve this objective, this book is organized with the following distinct features.

- Providing an example-based description on the whole data mining process for bioinformatics applications. This is distinct from method-based description, in which the chapters are organized according to different data mining techniques. Such an example-based organization is beneficial as it may help the readers to understand how to solve a real problem at hand by choosing proper data mining methods.
- Covering most popular data mining techniques throughout the book. Currently, there are many data mining methods in the literature. This book covers most of them and shows their applications in practical bioinformatics problems.
- Giving detailed illustrations and examples of how to use different data mining techniques to solve the same bioinformatics problem. Due to the complex nature of bioinformatics problems, the same problem can be solved using different data mining techniques. Different

solutions vary from underlying assumptions to algorithmic details. Such kinds of examples will not only enable the reader to understand the target problem more deeply, but also provide hints on how to apply data mining methods in his or her future bioinformatics research. Using frontier bioinformatics problems as examples in each chapter. All the examples discussed in this book will be frontier bioinformatics problems that are under investigation by the author and other researchers. Students who are interested in developing new and better algorithms for these problems may use this book as a starting point.

Acknowledgments

Many articles and books have been referenced in the writing of this book, and citations have been given for these works. To enhance the readability, I have tried to minimize literature references in the text. The sources and some additional literature that may be interesting to the readers are included in the reference section at the end of each chapter. I hope that this organization will provide the authors of the source literature with the appropriate acknowledgments. Moreover, this work was partially supported by the Natural Science Foundation of China under the Grant No. 61003176, the Fundamental Research Funds for the Central Universities of China under the Grant No. DUT14QY07.

This book could not have been written without the help of many people. First of all, most contents in this book are based on the research articles coauthored by my students and myself. Therefore, I would like to first thank my former and current students: Ting Huang, Haipeng Gong, Yang Zhang, Ben Teng, Xiaoqing Liu, Jun Wu, Can Zhao, and Feiyang Gu. In particular, I am grateful to Ben Teng and Xiaoqing Liu, who have carefully read the book and given very detailed comments and suggestions.

In addition, I would like to thank my former colleagues at the Hong Kong University of Science and Technology: Weichuan Yu, Can Yang, Chao Yang, and Xiang Wan, who have assisted me with the study of bioinformatics problems that have contributed to this book.

Furthermore, I am especially thankful to my PhD supervisor, Professor Xiaofei Xu, for his continuous efforts on "transforming" my research style from algorithm-driven research to application-driven research. Without such a transition, I would not have been able to write and finish this book.

Last, but definitely not least, I am indebted to my parents, my wife, and my son for their continuous support and patience.

Zengyou He
Dalian, China

Contents

List of figures

List of tables

An overview of data mining

1.1 What's data mining?

Data mining lies at the intersection of computer science, optimization, and statistics, and often appears in other disciplines. Generally, data mining is the process of searching for knowledge in data from different perspectives. Here knowledge can refer to any kinds of summarized or unknown information that are hidden underlying the raw data. For instance, it can be a set of discriminative rules generated from the data collected on some patients of a certain disease and healthy people. These rules can be used for predicting the disease status of new patients.

In general, data mining tasks can be classified into two categories: *descriptive* and *predictive*. Descriptive mining tasks characterize a target data set in concise, informative, discriminative forms. Predictive mining tasks conduct the induction and inference on the current data to make future predictions.

1.2 Data mining process models

Data mining is an iterative process that consists of many steps. There are already some generic reference models on the data mining process, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) process model. From a data-centric perspective, these models are structured as sequences of steps to transform the raw data into information or knowledge that is practically useful. As shown in Figure 1.1, a data mining process model typically involves the following phases: data collection, data preprocessing, data modeling, model assessment, and model deployment.

1.3 Data collection

The first step in the data mining process is to collect the relevant data according to the analysis goal in the applications. Generally, all the data that are helpful to achieve the objective in the analysis should be included. The key point here is how to define and understand the rather subjective term of "relevant data." Its correct interpretation highly depends on our understanding of the target problem and application background. Although this point will be further illustrated in subsequent chapters, we offer some general remarks here:

- In some cases, people definitely know that some kinds of data are highly relevant to the data mining task at hand. However, the acquisition of such data is very difficult or even impossible due to the device deficiency or cost. For instance, to accurately identify peptides in mass-spectrometry-based shotgun proteomics, it is necessary to generate at least one mass

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00001-6

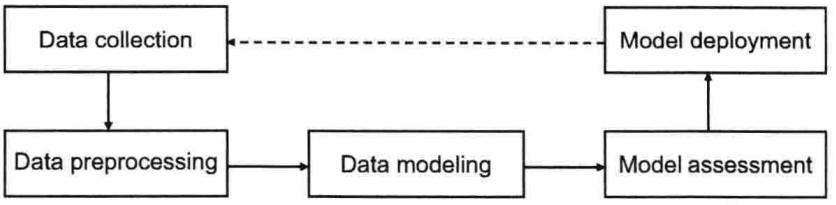


Figure 1.1 Typical phases involved in a data mining process model.

spectrum for each peptide in the sample. However, due to the limitation of current mass spectrometers, it is not always possible to obtain mass spectra data that can cover all peptides present in the sample.

- On the other hand, the inclusion of new relevant data may significantly change the models and methods in the consequent steps of the data mining process. Furthermore, it is necessary to check thoroughly if the use of more relevant data will boost the performance of data mining procedures.

1.4 Data preprocessing

The objective of data preprocessing is twofold: (1) The real-world data are usually low quality; hence preprocessing is used to improve the quality of data, and consequently, the quality of data mining results. (2) In the data modeling step, some specific modeling algorithms cannot operate on the raw data, which should be transformed into some predefined data formats.

There are several general-purpose data preprocessing methods: *data cleaning*, *data integration*, *data reduction*, and *data transformation*.

Data cleaning: Real-world data are usually noisy, inconsistent, and incomplete. Data cleaning procedures aim at removing the noise, correcting inconsistencies, and filling in missing values in the data.

Data integration: In the data collection phase, data sets from different sources are relevant to the analysis problem. Data integration merges data from different sources into an integrated data set for subsequent data mining analysis. The main objective of data integration is to reduce and avoid redundancies and inconsistencies in the resulting data set.

Data reduction: The purpose of data reduction is to generate a new yet smaller representation of the original data set. Generally, the reduced data should contain approximately the same information of the original data that is of primary importance to the analysis target. The most commonly used data reduction technique includes dimension reduction (vertically, reduce the number of features) and sampling (horizontally, reduce the number of samples).

Data transformation: Different data mining algorithms may require different forms of data. Data transformation techniques consolidate the original data into forms appropriate for subsequent mining tasks. For instance, data normalization will transform

the feature values into a predefined range such as [0.0, 1.0]. Data discretization will replace a continuous feature with a discrete one by dividing numeric values into intervals.

1.5 Data modeling

Before discussing the data modeling algorithms, it would be best to explain some terminology. Typically, the data preprocessing step would transform the raw data into a tabular form, in which the columns represent features/variables and rows correspond to samples/instances. For instance, Table 1.1 is a sample data set that has eight samples and five features (class is a special feature for which we are aiming to predict its feature value for a new given sample). The first four features are called *predictive features* and the class feature is the *target feature*. Here the predictive features can be symptoms of some disease, where the value of 1 indicates the existence of a symptom and 0 indicates otherwise. Similarly, the class feature value is 1 if the corresponding person (sample) has the disease.

1.5.1 Pattern mining

Pattern discovery is a core data mining problem, which generates a set of interesting patterns that characterize the data sets in concise and informative forms. Initially, the studies on pattern discovery were dominated by the frequent pattern discovery paradigm, where only frequent patterns were explored. Currently, the issue of frequent pattern discovery has been thoroughly investigated, rendering its limitations well understood. Many alternative pattern discovery formulations are emerging and investigated in the literature. For example, many research efforts impose statistical significance tests over candidate patterns to control the risk of false discoveries.

Table 1.1 An example data set with eight samples and five features

	Feature 1	Feature 2	Feature 3	Feature 4	Class
1	0	0	1	1	0
2	0	1	1	1	0
3	1	0	0	1	0
4	1	0	0	0	0
5	1	1	1	1	1
6	1	1	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1

Here *class* is a special feature representing the category to which each sample belongs.