FEATURE SELECTION FOR KNOWLEDGE DISCOVERY AND DATA MINING

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

Library of Congress Cataloging-in-Publication Data

Liu, Huan.
Feature selection for knowledge discovery and data mining / by Huan Liu and Hiroshi Motoda.
p. cm. -- (Kluwer international series in engineering and computer science ; 454)
Includes bibliographical references and index.
ISBN 978-1-4613-7604-0 ISBN 978-1-4615-5689-3 (eBook)
DOI 10.1007/978-1-4615-5689-3
1. Database management. 2. Data mining. I. Motoda, Hiroshi. II. Title. III. Series : Kluwer international series in engineering and computer science ; SECS 454.
QA76.9.D3L595 1998
006.3--dc21 98-25204
CIP

Originally published by Kluwer Academic Publishers, New York in 1998. Second Printing 2000.
Softcover reprint of the hardcover 1st edition 1998

Printed on acid-free paper.

FEATURE SELECTION FOR KNOWLEDGE DISCOVERY AND DATA MINING

by

Huan Liu
National University of Singapore
SINGAPORE

and

Hiroshi Motoda
Osaka University
Osaka, JAPAN

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Contents

List of Figures

List of Tables

Preface

As computer power grows and data collection technologies advance, a plethora of data is generated in almost every field where computers are used. The computer generated data should be analyzed by computers; without the aid of computing technologies, it is certain that huge amounts of data collected will not ever be examined, let alone be used to our advantages. Even with today's advanced computer technologies (e.g., machine learning and data mining systems), discovering knowledge from data can still be fiendishly hard due to the characteristics of the computer generated data. Taking its simplest form, raw data are represented in feature-values. The size of a dataset can be measured in two dimensions, number of features (N) and number of instances (P). Both N and P can be enormously large. This enormity may cause serious problems to many data mining systems.

Feature selection is one of the long existing methods that deal with these problems. Its objective is to select a minimal subset of features according to some reasonable criteria so that the original task can be achieved equally well, if not better. By choosing a minimal subset of features, irrelevant and redundant features are removed according to the criterion. When N is reduced, the data space shrinks and in a sense, the data set is now a better representative of the whole data population. If necessary, the reduction of N can also give rise to the reduction of P by eliminating duplicates. Simpler data can lead to more concise results and their better comprehensibility. Because of these advantages, feature selection has been the focus of interest for quite some time. Much work has been done from 70's to the present. With the creation of huge databases and the consequent requirements for good data mining programs, new problems arise and novel approaches to feature selection are in high demand. This is a perfect time to look back and see what have been done, and to look forward to the challenges ahead.

This book offers an overview of the various methods developed since 70's, provides a general framework in order to examine many methods and categorize them, employs simple examples to show the essence of representative feature selection methods, compares them using data sets with combinations of intrinsic properties according to the objective of feature selection, suggests guidelines how to use different methods under various circumstances, and points out some new challenges. This book consists of seven chapters, two appendices, and bibliographies after each chapter. Chapter 1 is about the background knowledge on data explosion, knowledge discovery from raw data, machine learning, and feature selection. Many fields related to these topics will be mentioned such as pattern recognition, statistics, visualization, database management systems, and so on. Chapter 2 describes perspectives of feature selection and explains how these different perspectives can be unified. Representative perspectives are pattern recognition, statistics, visualization, and machine learning. Chapter 3 derives a general framework after studying perspectives of feature selection. Chapter 4 starts with aspects of feature selection and illustrates representative feature selection methods to facilitate the reader in constructing his own feature selection method. Chapter 5 introduces the ways in which the methods should be evaluated, based on which empirical studies are performed. Experimental results are organized and presented in answering questions about the methods. It studies different characteristics of the data, and offers guidelines of applying feature selection methods under varied circumstances. Chapter 6 covers related topics such as feature extraction, feature construction, and feature discretization. Chapter 7 elaborates on the underlying philosophy throughout the book - *less is more* and concludes the book with prospects of feature selection in data mining. In the two appendices, we list the on-line sources for machine learning, data mining and knowledge discovery and provide descriptions of data sets and software used in the book. The soft copies of data sets and programs with instructions can be accessed from `http://www.iscs.nus.edu.sg/~liuh/Fsbook`.

This book can be considered as the first book for those, who start working on knowledge discovery but have not been exposed to feature selection, to understand the essence of feature selection and its various concepts. The book can be used by researchers in machine learning, data mining, and knowledge discovery as a toolbox in which they can find relevant tools that help in solving large real-world problems. The book can also be served as a reference book for those who are taking up tomorrow's challenges and conducting the research about feature selection.

To our families for their love,
support and understanding.

Acknowledgments

We are grateful to the editorial staff of Kluwer Academic Publishers, especially Scott Delman, Sharon Fletcher, and Robert Holland for their patience, interest, and helpfulness in bringing this project to a successful conclusion.

We have been fortunate to be assisted and encouraged by our colleagues and research associates who have contributed to this project in many ways. We are deeply indebted to them for their careful reading, thorough checking, and constructive suggestions. Although there is no space to thank them all, our special thanks to Manoranjan Dash, Farhad Hussain, Jian Shu, and Jun Yao for their invaluable assistance and extraordinary patience in this seemingly endless project, Kiansing Ng for helping design and complete the annotations in Appendices and maintaining the well constructed web sites, Rudy Setiono, Takashi Washio and Tadashi Horiuchi for fruitful collaboration without which the book would have been incomplete, Kaitat Foong and Kiansing Ng for their innovative cover design for this book.

Our families have been a constant source of encouragement throughout this project. Huan's greatest debt to his parents Baoyang Liu and Lihua Chen, and his family: Lan, Feitong, and Feige. Hiroshi's deepest gratitude to his family: Kay, Lena, Yuka and Eri.