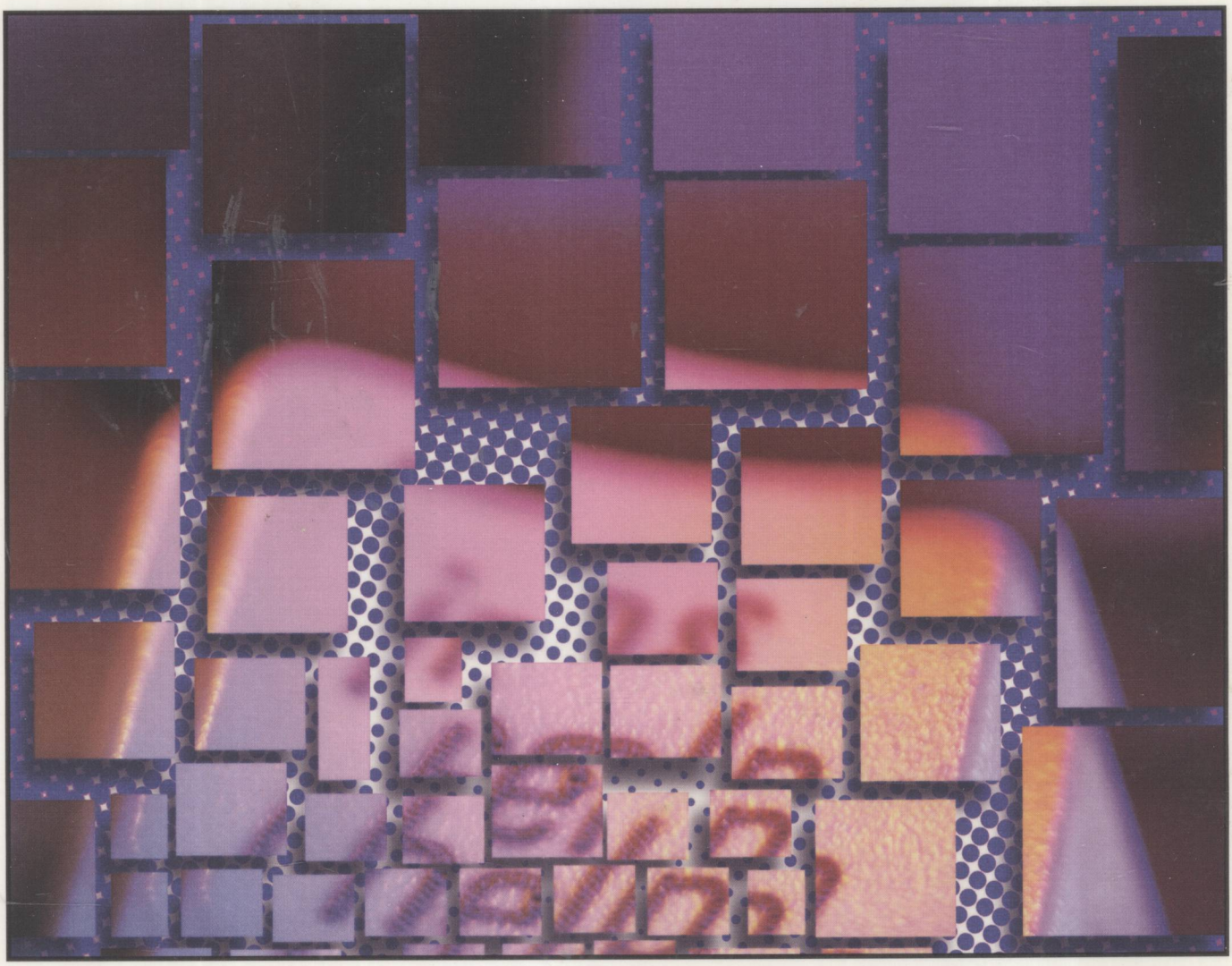
# INTELLIGENT DATA ANALYSIS

## Developing New Methodologies Through Pattern Discovery and Recovery



HSIAO-FAN WANG

# Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery

Hsiao-Fan Wang
*National Tsing Hua University, Taiwan, ROC*


Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

| | |
|---|---|
| Director of Editorial Content: | Kristin Klinger |
| Managing Development Editor: | Kristin M. Roth |
| Senior Managing Editor: | Jennifer Neidig |
| Managing Editor: | Jamie Snavely |
| Assistant Managing Editor: | Carole Coulson |
| Copy Editor: | Jennifer Young |
| Typesetter: | Chris Hrobak |
| Cover Design: | Lisa Tosheff |
| Printed at: | Yurchak Printing Inc. |

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com



Library of Congress Cataloging-in-Publication Data

Intelligent data analysis : developing new methodologies through pattern discovery and recovery / Hsiao-Fan Wang, Editor.

p. cm.

Summary: "This book tackles those data sets and covers a variety of issues in relation to intelligent data analysis so that patterns from frequent or rare events in spatial or temporal spaces can be revealed. This book brings together current research, results, problems, and applications from both theoretical and practical approaches"--Provided by publisher.

ISBN-13: 978-1-59904-982-3 (hardcover)

ISBN-13: 978-1-59904-983-0 (e-book)

1. Pattern perception--Data processing. 2. Artificial intelligence. I. Wang, Hsiao-Fan.

Q327.1545 2008

003'.52--dc22

2007046947

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

# Editorial Advisory Board

Dilip Kumar Pratihar
*Indian Institute of Technology, India*

Witold Pedrycz,
*University of Alberta, Canada*

Heather Ruskin,
*Dublin City University, Ireland*

Tony Smith
*University of Waikato, Hamilton, New Zealand*

Eulalia Szmidt
*Systems Research Institute, Polish Academy of Sciences, Poland*

M.L. Wang
*Ming-Hsin University of Science & Technology, Taiwan, ROC*

Sławomir T. Wierzchoń
*Institute of Komputer Scence, Polish Academy of Sciences, Poland*

W. C. Yeh
*National Tsing Hua University, Taiwan, ROC*

Mohammed J. Zaki
*Rensselaer Polytechnic Institute*

Dr. Francisco Torrens Zaragoza
*Institut Universitari de Ciencia Molecular Universitat de Valencia Edifici d-Instituts, Spain*

# Foreword

Professor Hsiao-Fan Wang has prepared a technically excellent and timely book on a new, important, and rapidly developing area of research and applications: *Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery.*

This book includes two categories of non-classical data analysis of Data Mining and Data Construction. It is perhaps the first book that provides comprehensive methodologies and real case applications towards huge data warehouses and small data nests. In particular, the book takes the viewpoint of pattern recognition and demonstrates that the tools developed can be applied to studies in learning and decision making of general human activities as a whole.

Professor Hsiao-Fan Wang should be congratulated for an outstanding job on this edited version of the book. Her vision on placing the theme on this rapidly developed yet very basic agenda of data analysis is valuable and significant. Professor Wang's balanced approach between theories and applications is a reflection of her own extensive research experience. Her way to tackle small data samples demonstrates very well her deep understanding of the needs and the trends of development in the real engineering world. While data mining has become a matured research area, developing data construction techniques for such rare events will undoubtedly draw more and more research attention in the future.

As we are entering an era in which machine intelligence and artificial intelligence will play significant roles in the design of intelligent and complex systems, the problem of recognizing patterns from either huge or small data set will continue to challenge the researchers and practitioners. And I am sure this book will play an important role in motivating a leading further development and applications in the area of intelligent data analysis.

*Chung Laung Liu*
*Honorary Chair Professor*
*National Tsing Hua University*

***C.L. Liu*** *received his advanced degrees from the Massachusetts Institute of Technology. He taught at MIT, the University of Illinois at Urbana Champaign, and the National Tsing Hua University. From 1996 to 1998, he served as associate provost at UIUC. From 1988 to 2002, he was president of National Tsing Hua University (NTHU). Dr. Liu is a member of Academia Sinica, and also, fellow of IEEE and ACM. After his term as president of NTHU, Dr. Liu continues his teaching and research activities. He also serves as consultant to high tech companies, works for a charitable foundation in Hong Kong, and, in the last two years, hosts a weekly radio show on technology and humanities in the radio station IC975 in Hsinchu, Taiwan, ROC.*

# Preface

*Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery* provides learning tools of finding data patterns based on artificial intelligence. Pattern Recognition has a long history of applications to data analysis in business, military, and social economic activities. While the aim of pattern recognition is to discover the pattern of a data set, the size of the data set is closely related to the methodology one adopts for analysis. The classic approach is using certain statistical techniques to deal with data sets of more than 30 samples and by dimension reduction to reveal the pattern. With the rapid increase of Internet development and usage, the amount of data has been enormous. The term "Data Warehouse" has been used to describe such quantities of data and the corresponding methodologies for analysis are under the title of "data mining."

In contrast to the huge amount of data sets, there is another type of data set which is small (less than 30), but still is significant in terms of socioeconomic cost. Consider severe earthquakes, random terrorist attacks, and nuclear plant explosions; the occurrences of such events are relatively few that the conventional statistic assumptions cannot be verified and thus the methods fail to apply. The ability to predict such kinds of events remains a challenge for the researchers. This leads to the necessity of recovering a pattern by constructing data.

Apart from these two extreme cases related to the amount of data which affect the method of analysis to be adopted, the types of the data are another major factor needed to be considered. Since in reality, the collected data are never complete, a certain degree of uncertainty is always embedded. The classical approach to coping with uncertainty is based on Probability Theory in random nature. Along with different methodologies and observations being investigated, data types other than randomness are studied and explored. Among these, fuzzy data, grey data, and coarse data with their hybrid forms are studied most extensively. The results pave a way to find data patterns from binary groupings to degree of belongings in more accurate and precise manner.

For all of these data types in quantity and quality, apart from Probability Inference being adopted for analysis, a group of heuristic approaches, namely soft computing (or computational intelligence), has been developed and employed for different areas of applications. Fuzzy logic, evolutionary computing, neural net analysis, and so on have shown their capability in coping with such kinds of data sets. It is an art and science for intelligent data analysis

Since pattern recognition has been a learning process ever since living beings began, classical approaches to classifying data into binary groups have been enormous in the literature. Due to the increasing impact of extreme events on the socio-economic costs, 38 authors from 10 different countries contributed their findings to 18 chapters in this book, each addresses different issues of intelligent pattern discovery and recovery from both theoretical and practical viewpoints. The readers will benefit from the integration of these two extreme cases in a comparative manner.

The book is categorized into four sections. After an introduction of the up-to-date development and research on methodologies and data properties in Section I, issues and resolution of pattern discovery

from huge data set are discussed and applied respectively in Sections II and III. Finally, in Section IV, methodology developments and the possible applications of pattern recovery from small data sets are presented. It can be noted from the unbalanced numbers of chapters related to huge data sets and small data sets, methods related to pattern recovery from small data set require the devotion of more researchers. The outline of each chapter is given below.

In **Section I**, five chapters are included, as outlined below:

Chapter I provides a software platform for automatic data analysis that uses a fuzzy knowledge base for automatically selecting and executing data analysis methods. The authors show that a system based on a fuzzy pattern base that stores heuristic expert knowledge from data analysis can successfully lead to automatic intelligent data analysis. Therefore, the system is able to support business users in running data analysis projects more efficiently

Chapter II provides a rigorous theory of random fuzzy sets in its most general form. We focus on imprecise data which are both random and fuzzy. Critical issues in relation to such kinds of data with hybrid natures are discussed and a framework based on Probability Theory is proposed for analysis.

Chapter III highlights meaningful pattern discovery techniques for gene expression data. The properties of gene expression data themselves are examined and the possible patterns are suggested. The classes of clustering techniques in the context of their application to gene expression data are investigated and a comparative analysis of standard and non-standard methods is given with the suggestion of areas for possible future development.

Chapter IV describes the use of fast, data-mining algorithms such as TreeNet and Random Forests (Salford Systems Ltd) to identify ecologically meaningful patterns and relationships in subsets of data that carry various degrees of outliers and uncertainty. An example of using satellite data from a wintering golden eagle shows that the proposed approach has provided a promising tool for wildlife ecology and conservation management.

Chapter V applies Atanassov's theory of intuitionistic fuzzy sets to analyze imbalanced and overlapping classes by defining both the membership and non-membership degrees for each member. Since imbalanced and overlapping classes are a real challenge for the standard classifiers, the method proposed in this chapter is crucial not only in theory but also on many different types of real tasks.

**Section II** of methodologies regarding pattern discovery from huge data set contains five chapters; each is introduced as below:

Chapter VI introduces fuzzy neural network models as a means for knowledge discovery from databases. It not only describes architectures and learning algorithms for fuzzy neural networks, but also proposes an algorithm for extracting and optimizing classification rules from a trained fuzzy neural network. An example of multispectral satellite images is given and it shows that the presented models and the methodology for generating classification rules from data samples provide a valuable tool for knowledge discovery.

Chapter VII discusses the paradigm of genetic algorithms and their incorporation into machine learning. Special attention is given to three issues: (a) the ways of initialization of a population for a genetic algorithm, (b) representation of chromosomes in genetic algorithms, and (c) discretization and fuzzification of numerical attributes for genetic algorithms. Furthermore, this chapter surveys new trends of dealing with the variable-length chromosomes and other issues related to the genetic learners.

Chapter VIII introduces the evolutionary computing as a whole and discusses specifically in detail two sub-areas of nature-inspired computing in Evolutionary Computing; namely, evolutionary algorithms

and swarm intelligence. The theoretical background of these sub-areas is illustrated with demonstration of some real-world applications. The chapter also points out future trends and directions in these areas.

Chapter IX proposes two composite approaches which combine conventional data fitting with peak-matching to cope with "noise" data in solving an inverse light scattering problem for single, spherical, homogeneous particles using least squares global optimization and show that they lead to a more robust identification procedure.

Chapter X introduces an approach called *Markov chain Monte Carlo* for the *exact* simulation of sample values from complex distributions. The proposed algorithm facilitates the implementation of a Markov chain that has a given distribution as its stationary distribution. The applications of these algorithms in probabilistic data analysis and inference are given.

**Section III** of the applications of pattern discovery from huge data set contains five cases from different industrial sectors of manufactory, transportation, and services:

Chapter XI provides suitable knowledge bases (KBs) for carrying out forward and reverse mappings of the Tungsten Inert Gas (TIG) welding process. Both the forward as well as reverse mappings are required for an effective online control of a process. Although conventional statistical regression analysis is able to carry out the forward mapping efficiently, it may not be always able to solve the problem of reverse mapping. Fuzzy logic (FL)-based approaches are adopted to conduct the forward and reverse mappings of the TIG welding process and they have shown to solve the above problem efficiently.

Chapter XII concerns a problem of road travel in the US, namely the discernment of the levels of traffic fatalities across the individual states. Based on the cognitive uncertainties evident in the imprecision inherent with the data values, a fuzzy approach to decision tree is adopted for inference. The results show that the inference from the tree structure takes advantage of the ability of humans to distinguish between patterns and observable characteristics.

Chapter XIII provides a method to resolve the major problem of time discontinuity resulting from the transactional character of events in telecom market. By gradually enriching the data information content from the prior lifetime expectancy through standard static events data up to decay-weighted data sequences, the proposed sequential processing of appropriately preprocessed data streams is shown to be able to have better performance of customer churn prediction.

Chapter XIV applies Dempster-Shafer Theory to object classification and ranking. Based on this theory, a method called CaRBS is proposed and an application to cope with uncertain reasoning on Moody's Bank Financial Strength Rating (BFSR) process is demonstrated. The value of this chapter is placed on the measures of ignorance such that during a series of classification and ranking analyses, decision on adopting or abandoning the existing evidence can be determined.

Chapter XV illustrates how to describe the individual's preference structure and utilize its properties to define an individual's risk level for the confronted risk. Then, a response evaluation model was proposed to develop the appropriate response strategy. These two stages of risk analysis and a risk response contribute to a complete individual risk management process (IRM). A case of A-C court was demonstrated and the results showed that the proposed method is able to provide more useful and pertinent information than the traditional method of decision tree which is based on the expected monetary value (EMV).

**Section IV** contains three chapters of current methodologies developed for analyzing small sample sets with illustrations of their applications.

Chapter XVI introduces the use of the bootstrap in a nonlinear, nonparametric regression framework with dependent errors. The AR-Sieve bootstrap and the Moving Block bootstrap which are used to gen-

erate bootstrap replicates with a proper dependence structure are used to avoid the inconsistent choice inherent in conventional Bootstrap method. In the framework of neural network models which are often used as an accurate nonparametric estimation and prediction tool, both procedures have shown to have satisfactory results.

Chapter XVII proposes a methodology based on Hilbert-EMD-based support vector machine (SVM) to predict financial crisis events for early-warning purpose. A typical financial indicator currency exchange rate reflecting economic fluctuation conditions is first chosen. Then the Hilbert-EMD algorithm is applied to the economic indicator series. This chapter also applies the proposed method to two real-world cases of South Korea and Thailand, which suffered from the 1997-1998 disastrous financial crisis experience. The results show that the proposed Hilbert-EMD-based SVM methodology is capable of predicting the financial crisis events effectively.

Chapter XVIII proposed an alternative approach named Data Construction Method (DCM) to overcome the problems derived from insufficient data; in particular, the defects existent in one commonly used approach called *Intervalized Kernel method of Density Estimation* (IKDE). Comparative studies have shown that the proposed DCA is not only resolve the insufficient data in general; but also improve the prediction accuracy in both degrees and stability of IKDE.

From the content described above, it can be noted that this book will be useful for both researchers and practitioners who are interested in receiving comprehensive views and insights from the variety of issues covered in this book in relation to pattern discovery and recovery. In particular, those who have been working on data analysis will have an overall picture of the existing and potential developments on the issues related to intelligent pattern recognition.

*Hsiao-Fan Wang*
*National Tsing Hua University*

# Acknowledgment

The editor would like to acknowledge the help of all involved in the collation and review process of the book, without whose support the project could not have been satisfactorily completed. Deep appreciation and gratitude is due to two organizations of the Department of Management, University of Canterbury, NZ and National Science Council, Taiwan, ROC for their facility and financial support this year as visiting professor.

Deep gratitude first is sent to Dr. Chung Laung Liu, the formal president of National Tsing Hua University, for his Foreword, full of encouragement and kind support. We also would like to thank all authors for their excellent contributions to this volume. In particular, most of the authors of chapters included in this book also served as referees for chapters written by other authors. Thanks go to all those who provided comprehensive reviews and constructive comments. Special thanks also go to the publishing team at IGI Global, in particular to Ross Miller and Jessica Thompson, who continuously prodded via e-mail to keep the project on schedule and to Mehdi Khosrow-Pour, whose enthusiasm motivated me to initially accept his invitation for taking on this project.

Special thanks go to National Tsing Hua University and the colleagues of the Department of Industrial Engineering and Engineering Management in Taiwan, without their understanding and support, this volume won't be possible.

Finally, I wish to thank my boys, I-Fan (Daniel) and Tao-Fan (Ray), for their understanding and immense love during this project.

*Editor,*
*Hsiao-Fan Wang*
*Tsing Hua Chair Professor*
*Taiwan, Republic of China*
*October 2007*

# Section I
# Introduction

# Table of Contents

## Section III
## Pattern Discovery from Huge Data Set: Applications



## Section IV
## Pattern Recovery from Small Data Set: Methodologies and Applications

# Detailed Table of Contents