Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Cluster Analysis

Edited by

Christian Hennig

Marina Meila

Fionn Murtagh

Roberto Rocci

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Handbooks of Modern
Statistical Methods

Handbook of Cluster Analysis

Edited by

Christian Hennig
University College London, UK

Marina Meila
University of Washington, Seattle, USA

Fionn Murtagh
University of Derby, UK
Goldsmiths, University of London, UK

Roberto Rocci
University of Rome Tor Vergata, Italy

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed and bound by CPI UK on sustainably sourced paper
Version Date: 20151012

International Standard Book Number-13: 978-1-4665-5188-6 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Hennig, Christian M., editor. | Meilæa, Marina, 1962- editor |
Murtagh, Fionn, editor. | Rocci, Roberto, editor.
Title: Handbook of cluster analysis / Christian M Hennig, Marina Meila, Fionn
Murtagh, Roberto Rocci, editors.
Description: Boca Raton : Taylor & Francis, 2016. | Series: Chapman &
Hall/CRC handbooks of modern statistical methods ; 9 | "A CRC title." |
Includes bibliographical references and index.
Identifiers: LCCN 2015021473 | ISBN 9781466551886 (alk. paper)
Subjects: LCSH: Cluster analysis. | Spatial analysis (Statistics)
Classification: LCC QA278 .H3456 2016 | DDC 519.5/3--dc23
LC record available at http://lccn.loc.gov/2015021473

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Handbook of Cluster Analysis

Chapman & Hall/CRC
Handbooks of Modern Statistical Methods

Series Editor

Garrett Fitzmaurice
Department of Biostatistics
Harvard School of Public Health
Boston, MA, U.S.A.

Aims and Scope

The objective of the series is to provide high-quality volumes covering the state-of-the-art in the theory and applications of statistical methodology. The books in the series are thoroughly edited and present comprehensive, coherent, and unified summaries of specific methodological topics from statistics. The chapters are written by the leading researchers in the field, and present a good balance of theory and application through a synthesis of the key methodological developments and examples and case studies using real data.

The scope of the series is wide, covering topics of statistical methodology that are well developed and find application in a range of scientific disciplines. The volumes are primarily of interest to researchers and graduate students from statistics and biostatistics, but also appeal to scientists from fields where the methodology is applied to real problems, including medical research, epidemiology and public health, engineering, biological science, environmental science, and the social sciences.

Published Titles

Handbook of Mixed Membership Models and Their Applications
Edited by Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg

Handbook of Markov Chain Monte Carlo
Edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng

Handbook of Discrete-Valued Time Series
Edited by Richard A. Davis, Scott H. Holan, Robert Lund, and Nalini Ravishanker

Handbook of Design and Analysis of Experiments
Edited by Angela Dean, Max Morris, John Stufken, and Derek Bingham

Longitudinal Data Analysis
Edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs

Handbook of Spatial Statistics
Edited by Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, and Peter Guttorp

Handbook of Cluster Analysis
Edited by Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci

Handbook of Survival Analysis
Edited by John P. Klein, Hans C. van Houwelingen, Joseph G. Ibrahim, and Thomas H. Scheike

Handbook of Missing Data Methodology
Edited by Geert Molenberghs, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke

Preface

This Handbook intends to give a comprehensive, structured, and unified account of the central developments in current research on cluster analysis.

The book is aimed at researchers and practitioners in statistics, and all the scientists and engineers who are involved in some way in data clustering and have a sufficient background in statistics and mathematics. Recognizing the interdisciplinary nature of cluster analysis, most parts of the book were written in a way that is accessible to readers from various disciplines. How much background is required depends to some extent on the chapter. Familiarity with mathematical and statistical reasoning is very helpful, but an academic degree in mathematics or statistics is not required for most parts. Occasionally some knowledge of algorithms and computation will help to make most of the material.

Since we wanted this book to be immediately useful in practice, the clustering methods we present are usually described in enough detail to be directly implementable. In addition, each chapter comprises the general ideas, motivation, advantages and potential limits of the methods described, signposts to software, theory and applications, and a discussion of recent research issues.

For those already experienced with cluster analysis, the book offers a broad and structured overview. For those starting to work in this field, it offers an orientation and an introduction to the key issues. For the many researchers who are only temporarily or marginally involved with cluster analysis problems, the book chapters contain enough algorithmic and practical detail to give them a working knowledge of specific areas of clustering. Furthermore, the book should help scientists, engineers, and other users of clustering methods to make informed choices of the most suitable clustering approach for their problem, and to make better use of the existing cluster analysis tools.

Cluster analysis, also sometimes known as unsupervised classification, is about finding groups in a set of objects characterized by certain measurements. This task has a very wide range of applications such as delimitation of species in biology, data compression, classification of diseases or mental illnesses, market segmentation, detection of patterns of unusual Internet use, delimitation of communities, or classification of regions or countries for administrative use. Unsupervised classification can be seen as a basic human learning activity, connected to issues as basic as the development of stable concepts in language.

Formal cluster analysis methodology has been developed, among others, by mathematicians, statisticians, computer scientists, psychologists, social scientists, econometrists, biologists, and geoscientists. Some of these branches of development existed independently for quite some time. As a consequence, cluster analysis as a research area is very heterogeneous. This makes sense, because there are also various different relevant concepts of what constitutes a cluster. Elements of a cluster can be connected by being very similar to each other and distant from nonmembers of the cluster, by having a particular characterization in terms of few (potentially out of many) variables, by being appropriately represented by the same centroid object, by constituting some kind of distinctive shape or pattern, or by being generated from a common homogeneous probabilistic process.

Cluster analysis is currently a very popular research area and its popularity can be expected to grow more connected to the growing availability and relevance of data collected in all areas of life, which often come in unstructured ways and require some processing in order to become useful. Unsupervised classification is a central technique to structure such data.

Research on cluster analysis faces many challenges. Cluster analysis is applied to ever new data formats; many approaches to cluster analysis are computer intensive and their application to large databases is difficult; there is little unification and standardization in the field of cluster analysis, which makes it difficult to compare different approaches in a systematic manner. Even the investigation of properties such as statistical consistency and stability of traditional elementary cluster analysis techniques is often surprisingly hard.

Cluster analysis as a research area has grown so much in recent years that it is all but impossible to cover everything that could be considered relevant in a handbook like this. We have chosen to organize this book according to the traditional core approaches to cluster analysis, tracing them from the origins to recent developments. The book starts with an overview of approaches (Chapter 1), followed by a quick journey through the history of cluster analysis (Chapter 2). The next four sections of the book are devoted to four major approaches toward cluster analysis, all of which go back to the beginnings of cluster analysis in the 1950s and 1960s or even further. (Probably Pearson's paper on fitting Gaussian mixtures in 1894, see Chapter 2, was the first publication of a method covered in this Handbook, although Pearson's use of it is not appropriately described as "cluster analysis.")

Section I is about methods that aim at optimizing an objective function that describes how well data is grouped around centroids. The most popular of these methods and probably the most popular clustering method in general is K-means. The efficient optimization of the K-means and other objective functions of this kind is still a hard problem and a topic of much recent research.

Section II is concerned with dissimilarity-based methods, formalizing the idea that objects within clusters should be similar and objects in different clusters should be dissimilar. Chapters treat the traditional hierarchical methods such as single linkage, and more recent approaches to analyze dissimilarity data such as spectral clustering and graph-based approaches.

Section III covers the broad field of clustering methods based on probability models for clusters, that is, mixture models and partitioning models. Such models have been analyzed for many different kinds of data, including standard real-valued vector data, categorical, ordinal and mixed data, regression-type data, functional and time-series data, spatial data, and network data. A related issue, treated in Chapter 15, is to test for the existence of clusters.

Section IV deals with clustering methods inspired by nonparametric density estimation. Instead of setting up specific models for the clusters in the data, these approaches identify clusters with the "islands" of high density in the data, no matter what shape these have, or they aim at finding the modes of the data density, which are interpreted as "attractors" or representatives for the rest of the points. Most of these methods also have a probabilistic background, but their nature is nonparametric; they formalize a cluster by characterizing in terms of the density or distribution of points instead of setting it up.

Section V collects a number of further approaches to cluster analysis, partly analyzing specific data types such as symbolic data and ensembles of clusterings, partly presenting specific problems such as constrained and semi-supervised clustering and two-mode and multipartitioning, fuzzy and rough set clustering.

By and large, Sections I through V are about methods for clustering. But having a clustering method is not all that is needed in cluster analysis. Section VI treats further relevant issues, many of which can be grouped under the headline "cluster validation," evaluating the quality of a clustering. Aspects include indexes to measure cluster validity (which are often also used for choosing the number of clusters), comparing different clusterings, measuring cluster stability and robustness of clustering methods, cluster visualization, and the general strategy in carrying out a cluster analysis and the choice of an appropriate method.

Given the limited length of the book, there are a number of topics that some readers may expect in the *Handbook of Cluster Analysis*, but that are not covered. We see most of the presented material as essential; some decisions were motivated by individual preferences and the chosen focus, some by the difficulty of finding good authors for certain topics. Much of what is missing are methods for further types of data (such as text clustering), some more recent approaches that are currently used by rather limited groups of users, some of the recent progress in computational issues for large data sets including some of the clustering methods, of which the main motivation is to be able to deal with large amounts of data, and some hybrid approaches that piece together various elementary ideas from clustering and classification. We have weighted an introduction to the elementary approaches (on which there is still much research and that still confronts us with open problems) higher than the coverage of as many branches as possible of current specialized cutting-edge research, although some of this was included by chapter authors, all of whom are active and well-distinguished researchers in the area.

It has been a long process to write this book and we are very grateful for the continuous support by Chapman & Hall/CRC and particularly by Robert Calver, who gave us a lot of encouragement and pushed us when necessary.

Editors

Christian Hennig is senior lecturer at the Department of Statistical Science, University College London. Previous affiliations were the Seminar für Statistik, ETH Zürich and the Faculty of Mathematics, University of Hamburg. He is currently secretary of the International Federation of Classification Societies. He is associate editor of *Statistics and Computing, Computational Statistics and Data Analysis, Advances in Data Analysis and Classification,* and *Statistical Methods and Applications.* His main research interests are cluster analysis, philosophy of statistics, robust statistics, multivariate analysis, data visualization, and model selection.

Marina Meila is professor of statistics at the University of Washington. She earned an MS in electrical engineering from the Polytechnic University of Bucharest, and a PhD in computer science and electrical engineering from the Massachusetts Institute of Technology. She held appointments at the Bucharest Research Institute for Computer Technology, the Polytechnic University of Bucharest, and the Robotics Institute of Carnegie Mellon University. Her long-term interests are in machine learning and reasoning in uncertainty, and how these can be performed efficiently on large complex data sets.

Fionn Murtagh earned degrees in engineering science, mathematics, computer science, a PhD in mathematical statistics, and habilitation in computational astronomy. He works in the field of data science and big data analytics. He served the Space Science Department of the European Space Agency for 12 years. He also held professorial chairs in computer science in a number of universities in the United Kingdom. He currently is a professor of data science. He is a fellow of the International Association for Pattern Recognition, a fellow of the British Computer Society, and an elected member of the Royal Irish Academy and of Academia Europaea. He is a member of the editorial boards of many journals, and has been editor-in-chief of the *Computer Journal* for more than 10 years.

Roberto Rocci is full professor of statistics at the Department of Economics and Finance, University of Rome Tor Vergata. He earned his PhD in statistics in 1994 at the Department of Statistical Science, Probability and Applied Probability, University of Rome La Sapienza. The topic of his dissertation was on multilinear models for multiway data. His field of interests are cluster analysis, mixture models, and latent variable models. He is the author of many papers published in international journals. Recently, he was the secretary of the Italian Statistical Society (SIS). Currently, he is associate editor of the *Statistical Methods and Applications Journal* and board member of SIS-CLADAG (SIS-CLassification and Data Analysis Group).

Contributors

Ayan Acharya
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas

Marco Alfó
Department of Statistical Sciences
Sapienza University of Rome
Rome, Italy

Pranjal Awasthi
Department of Computer Science
Rutgers University
New Brunswick, New Jersey

Adelchi Azzalini
Senior Scholar
University of Padua
Padua, Italy

Maria Florina Balcan
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania

Paula Brito
Faculdade de Economia and LIAAD-INESC TEC
Universidade do Porto
Porto, Portugal

Jorge Caiado
CEMAPRE/ISEG
University of Lisbon
Lisbon, Portugal

Miguel Á. Carreira-Perpiñán
Electrical Engineering and Computer Science
University of California, Merced
Merced, California

G. Celeux
Orsay
Île-de-France, France

Radha Chitta
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan

Pedro Contreras
Thinking Safe Limited
Egham, United Kingdom

Sébastien Déjean
Institut de Mathématiques UMR CNRS et Université de Toulouse
Université Paul Sabatier
Toulouse, France

Ivo Düntsch
Department of Computer Science
Brock University
St. Catharines, Ontario, Canada

Pierpaolo D'Urso
Department of Social Sciences and Economics
Sapienza University of Rome
Rome, Italy

L.A. García-Escudero
Departamento de Estadística e Investigación Operativa and IMUVA
Universidad de Valladolid
Valladolid, Spain

Günther Gediga
Fachbereich Psychologie
Universität Münster
Münster, Germany

Joydeep Ghosh
Department of ECE
University of Texas at Austin
Austin, Texas

A. Gordaliza
Departamento de Estadística e Investigación Operativa and IMUVA
Universidad de Valladolid
Valladolid, Spain

Gérard Govaert
Saclay
Île-de-France, France

and

CNRS and Université Technologique de Compiègne
Compiègne, France

Mark C. Greenwood
Department of Mathematical Sciences
Montana State University
Bozeman, Montana

Maria Halkidi
Department of Digital Systems
University of Piraeus
Piraeus, Greece

Julia Handl
Manchester Business School
University of Manchester
Manchester, United Kingdom

Lisa Handl
Institute of Stochastics
Ulm University
Ulm, Germany

David Neil Hayes
Lineberger Comprehensive Cancer Center
Department of Internal Medicine
University of North Carolina
Chapel Hill, North Carolina

Christian Hennig
Department of Statistical Science
University College London
London, United Kingdom

Christian Hirsch
Institute of Stochastics
Ulm University
Ulm, Germany

David B. Hitchcock
Department of Mathematical Sciences
Montana State University
Bozeman, Montana

Hanwen Huang
Department of Epidemiology and Biostatistics
University of Georgia
Athens, Georgia

Anil Jain
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan

Rong Jin
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan

Joshua Knowles
School of Computer Science
University of Manchester
Manchester, United Kingdom

Friedrich Leisch
Institute of Applied Statistics and Computing
University of Natural Resources and Life Sciences
Vienna, Austria

Yufeng Liu
Department of Statistics and Operations Research
Carolina Center for Genome Sciences
Lineberger Comprehensive Cancer Center
University of North Carolina
Chapel Hill, North Carolina

Elizabeth Ann Maharaj
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia

J.S. Marron
Department of Statistics and Operations Research
Lineberger Comprehensive Cancer Center
University of North Carolina
Chapel Hill, North Carolina

C. Matrán
Departamento de Estadística e Investigación Operativa and IMUVA
Universidad de Valladolid
Valladolid, Spain

A. Mayo-Iscar
Departamento de Estadística e Investigación Operativa and IMUVA
Universidad de Valladolid
Valladolid, Spain

Geoffrey J. McLachlan
Department of Mathematics
University of Queensland
St. Lucia, Australia

Marina Meila
Department of Statistics
University of Washington
Seattle, Washington

Boris Mirkin
Department of Computer Science
Birkbeck University of London
London, United Kingdom

and

Department of Data Analysis and Machine Intelligence
National Research University Higher School of Economics
Moscow, Russia

Josiane Mothe
Ecole Supérieure du Professorat et de L'éducation Académie de Toulouse
Institut de Recherche en Informatique de Toulouse
Université de Toulouse
Toulouse, France

Thomas Brendan Murphy
School of Mathematical Sciences
Complex and Adaptive Systems Laboratory and Insight Research Centre
University College Dublin
Dublin, Ireland

Fionn Murtagh
Department of Computing
Goldsmiths University of London
London, United Kingdom

and

Department of Computing and Mathematics
University of Derby
Derby, United Kingdom

Andrew Nobel
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, North Carolina

Vinayak Rao
Department of Statistics
Purdue University
West Lafayette, Indiana

Suren I. Rathnayake
Department of Mathematics
University of Queensland
St. Lucia, Australia

Roberto Rocci
Department of Economics and Finance
University of Tor Vergata
Rome, Italy

Volker Schmidt
Institute of Stochastics
Ulm University
Ulm, Germany

Douglas Steinley
Department of Psychological Sciences
University of Missouri
Columbia, Missouri

Michalis Vazirgiannis
Department of Informatics
Athens University of
Economics and Business
Athens, Greece

Maurizio Vichi
Department of Statistical Sciences
Sapienza University of Rome
Rome, Italy

Sara Viviani
Department of Statistical Sciences
Sapienza University of Rome
Rome, Italy