Grouping Multidimensional Data

Jacob Kogan · Charles Nicholas
Marc Teboulle (Eds.)

Grouping Multidimensional Data

Recent Advances in Clustering

With 53 Figures

Springer

Editors

Jacob Kogan

Department of Mathematics and Statistics
and Department of Computer Science
and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, Maryland 21250, USA
kogan@umbc.edu

Marc Teboulle

School of Mathematical Sciences
Tel-Aviv University
Ramat Aviv, Tel-Aviv 69978, Israel
teboulle@post.tau.ac.il

Charles Nicholas

Department of Computer Science
and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, Maryland 21250, USA
nicholas@umbc.edu

ACM Classification (1998): H.3.1, H.3.3
Library of Congress Control Number: 2005933258

ISBN-10 3-540-28348-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28348-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in The Netherlands

Typesetting by the authors and SPI Publisher Services using Springer LaTeX macro package

Cover design: KünkelLopka, Heidelberg

Printed on acid-free paper SPIN: 11375456 45/3100/SPI Publisher Services 5 4 3 2 1 0

Foreword

Clustering is one of the most fundamental and essential data analysis tasks with broad applications. It can be used as an independent data mining task to disclose intrinsic characteristics of data, or as a preprocessing step with the clustering results used further in other data mining tasks, such as classification, prediction, correlation analysis, and anomaly detection. It is no wonder that clustering has been studied extensively in various research fields, including data mining, machine learning, pattern recognition, and scientific, engineering, social, economic, and biomedical data analysis. Although there have been numerous studies on clustering methods and their applications, due to the wide spectrum that the theme covers and the diversity of the methodology research publications on this theme have been scattered in various conference proceedings or journals in multiple research fields. There is a need for a good collection of books dedicated to this theme, especially considering the surge of research activities on cluster analysis in the last several years.

This book fills such a gap and meets the demand of many researchers and practitioners who would like to have a solid grasp of the state of the art on cluster analysis methods and their applications. The book consists of a collection of chapters, contributed by a group of authoritative researchers in the field. It covers a broad spectrum of the field, from comprehensive surveys to in-depth treatments of a few important topics. The book is organized in a systematic manner, treating different themes in a balanced way. It is worth reading and further when taken as a good reference book on your shelf.

The chapter "A Survey of Clustering Data Mining Techniques" by Pavel Berkhin provides an overview of the state-of-the-art clustering techniques. It presents a comprehensive classification of clustering methods, covering hierarchical methods, partitioning relocation methods, density-based partitioning methods, grid-based methods, methods based on co-occurrence of categorical data, and other clustering techniques, such as constraint-based and graph-partitioning methods. Moreover, it introduces scalable clustering algorithms

and clustering algorithms for high-dimensional data. Such a coverage provides a well-organized picture of the whole research field.

In the chapter "Similarity-Based Text Clustering: A Comparative Study," Joydeep Ghosh and Alexander Strehl perform the first comparative study among popular similarity measures (Euclidean, cosine, Pearson correlation, extended Jaccard) in conjunction with several clustering techniques (random, self-organizing feature map, hypergraph partitioning, generalized k-means, weighted graph partitioning) on a variety of high-dimensional sparse vector data sets representing text documents as bags of words. The comparative performance results are interesting and instructive.

In the chapter "Criterion Functions for Clustering on High-Dimensional Data", Ying Zhao and George Karypis provide empirical and theoretical comparisons of the performance of a number of widely used criterion functions in the context of partitional clustering algorithms for high-dimensional datasets. This study presents empirical and theoretical guidance on the selection of criterion functions for clustering high-dimensional data, such as text documents.

Other chapters also provide interesting introduction and in-depth treatments of various topics of clustering, including a star-clustering algorithm by Javed Aslam, Ekaterina Pelekhov, and Daniela Rus, a study on clustering large datasets with principal direction divisive partitioning by David Littau and Daniel Boley, a method for clustering with entropy-like k-means algorithms by Marc Teboulle, Pavel Berkhin, Inderjit Dhillon, Yuqiang Guan, and Jacob Kogan, two new sampling methods for building initial partitions for effective clustering by Zeev Volkovich, Jacob Kogan, and Charles Nicholas, and "TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections" by Dimitrios Zeimpekis and Efstratios Gallopoulos. These chapters present in-depth treatment of several popularly studied methods and widely used tools for effective and efficient cluster analysis.

Finally, the book provides a comprehensive bibliography, which is a marvelous and up-to-date list of research papers on cluster analysis. It serves as a valuable resource for researchers.

I enjoyed reading the book. I hope you will also find it a valuable source for learning the concepts and techniques of cluster analysis and a handy reference for in-depth and productive research on these topics.

University of Illinois at
Urbana-Champaign
June 29, 2005

Jiawei Han

Preface

Clustering is a fundamental problem that has numerous applications in many disciplines. Clustering techniques are used to discover natural groups in datasets and to identify abstract structures that might reside there, without having any background knowledge of the characteristics of the data. They have been used in various areas including bioinformatics, computer vision, data mining, gene expression analysis, text mining, VLSI design, and Web page clustering to name just a few. Numerous recent contributions to this research area are scattered in a variety of publications in multiple research fields.

This volume collects contributions of computers scientists, data miners, applied mathematicians, and statisticians from academia and industry. It covers a number of important topics and provides about 500 references relevant to current clustering research (we plan to make this reference list available on the Web). We hope the volume will be useful for anyone willing to learn about or contribute to clustering research.

The editors would like to express gratitude to the authors for making their research available for the volume. Without these individuals' help and cooperation this book would not be possible. Thanks also go to Ralf Gerstner of Springer for his patience and assistance, and for the timely production of this book. We would like to acknowledge the support of the United States–Israel Binational Science Foundation through the grant BSF No. 2002-010, and the support of the Fulbright Program.

Karmiel, Israel and Baltimore, USA, Jacob Kogan
Baltimore, USA, Charles Nicholas
Tel Aviv, Israel, Marc Teboulle
July 2005

Contents

List of Contributors

J. A. Aslam
College of Computer and
Information Science
Northeastern University
Boston, MA 02115, USA
jaa@ccs.neu.edu

P. Berkhin
Yahoo!
701 First Avenue
Sunnyvale, CA 94089, USA
pberkhin@yahoo-inc.com

D. Boley
University of Minnesota
Minneapolis, MN 55455, USA
boley@cs.umn.edu

I. Dhillon
Department of Computer Science
University of Texas
Austin, TX 78712-1188, USA
inderjit@cs.utexas.edu

E. Gallopoulos
Department of Computer
Engineering and Informatics
University of Patras
26500 Patras
Greece
stratis@hpclab.ceid.upatras.gr

J. Ghosh
Department of ECE
University of Texas at Austin
1 University Station C0803
Austin, TX 78712-0240, USA
ghosh@ece.utexas.edu

Y. Guan
Department of Computer Science
University of Texas
Austin, TX 78712-1188, USA
yguan@cs.utexas.edu

G. Karypis
Department of Computer Science
and Engineering and Digital
Technology Center and
Army HPC Research Center
University of Minnesota
Minneapolis, MN 55455, USA
karypis@cs.umn.edu

J. Kogan
Department of Mathematics and
Statistics and
Department of Computer Science
and Electrical Engineering
University of Maryland
Baltimore County
Baltimore, MD 21250, USA
kogan@umbc.edu

D. Littau
University of Minnesota
Minneapolis, MN 55455, USA
littau@cs.umn.edu

C. Nicholas
Department of Computer Science
and Electrical Engineering
University of Maryland
Baltimore County
Baltimore, MD 21250, USA
nicholas@csee.umbc.edu

E. Pelekhov
Department of Computer Science
Dartmouth College
Hanover, NH 03755, USA
ekaterina.pelekhov@alum.
dartmouth.org

D. Rus
Computer Science and Artificial
Intelligence Laboratory
Massachusetts Institute of
Technology
Cambridge, MA 02139, USA
rus@csail.mit.edu

A. Strehl
Leubelfingstrasse 110
90431 Nurnberg
Germany
alexander@strehl.com

M. Teboulle
School of Mathematical Sciences
Tel Aviv University
Tel Aviv, Israel
teboulle@post.tau.ac.il

Z. Volkovich
Software Engineering Department
ORT Braude Academic College
Karmiel 21982, Israel
zeev@actcom.co.il

D. Zeimpekis
Department of Computer
Engineering and Informatics
University of Patras
26500 Patras
Greece
dsz@hpclab.ceid.upatras.gr

Y. Zhao
Department of Computer Science
and Engineering
University of Minnesota
Minneapolis, MN 55455, USA
yzhao@cs.umn.edu