DATA CLUSTERING

Algorithms and Applications

Edited by

Charu C. Aggarwal
Chandan K. Reddy

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

Contents

Preface

The problem of clustering is perhaps one of the most widely studied in the data mining and machine learning communities. This problem has been studied by researchers from several disciplines over five decades. Applications of clustering include a wide variety of problem domains such as text, multimedia, social networks, and biological data. Furthermore, the problem may be encountered in a number of different scenarios such as streaming or uncertain data. Clustering is a rather diverse topic, and the underlying algorithms depend greatly on the data domain and problem scenario.

Therefore, this book will focus on three primary aspects of data clustering. The first set of chapters will focus on the core methods for data clustering. These include methods such as probabilistic clustering, density-based clustering, grid-based clustering, and spectral clustering. The second set of chapters will focus on different problem domains and scenarios such as multimedia data, text data, biological data, categorical data, network data, data streams and uncertain data. The third set of chapters will focus on different detailed insights from the clustering process, because of the subjectivity of the clustering process, and the many different ways in which the same data set can be clustered. How do we know that a particular clustering is good or that it solves the needs of the application? There are numerous ways in which these issues can be explored. The exploration could be through interactive visualization and human interaction, external knowledge-based supervision, explicitly examining the multiple solutions in order to evaluate different possibilities, combining the multiple solutions in order to create more robust ensembles, or trying to judge the quality of different solutions with the use of different validation criteria.

The clustering problem has been addressed by a number of different communities such as pattern recognition, databases, data mining and machine learning. In some cases, the work by the different communities tends to be fragmented and has not been addressed in a unified way. This book will make a conscious effort to address the work of the different communities in a unified way. The book will start off with an overview of the basic methods in data clustering, and then discuss progressively more refined and complex methods for data clustering. Special attention will also be paid to more recent problem domains such as graphs and social networks.

The chapters in the book will be divided into three types:

- **Method Chapters:** These chapters discuss the *key techniques* which are commonly used for clustering such as feature selection, agglomerative clustering, partitional clustering, density-based clustering, probabilistic clustering, grid-based clustering, spectral clustering, and non-negative matrix factorization.

- **Domain Chapters:** These chapters discuss the specific methods used for different *domains* of data such as categorical data, text data, multimedia data, graph data, biological data, stream data, uncertain data, time series clustering, high-dimensional clustering, and big data. Many of these chapters can also be considered application chapters, because they explore the specific characteristics of the problem in a particular domain.

- **Variations and Insights:** These chapters discuss the *key variations* on the clustering process such as semi-supervised clustering, interactive clustering, multi-view clustering, cluster ensembles, and cluster validation. Such methods are typically used in order to obtain detailed insights from the clustering process, and also to explore different possibilities on the clustering process through either supervision, human intervention, or through automated generation of alternative clusters. The methods for cluster validation also provide an idea of the quality of the underlying clusters.

This book is designed to be comprehensive in its coverage of the entire area of clustering, and it is hoped that it will serve as a knowledgeable compendium to students and researchers.

Editor Biographies

Charu C. Aggarwal is a Research Scientist at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from Massachusetts Institute of Technology in 1996. His research interest during his Ph.D. years was in combinatorial optimization (network flow algorithms), and his thesis advisor was Professor James B. Orlin. He has since worked in the field of performance analysis, databases, and data mining. He has published over 200 papers in refereed conferences and journals, and has applied for or been granted over 80 patents. He is author or editor of nine books, including this one. Because of the commercial value of the above-mentioned patents, he has received several invention achievement awards and has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bioterrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of an IBM Research Division Award (2008) for his scientific contributions to data stream research. He has served on the program committees of most major database/data mining conferences, and served as program vice-chairs of the SIAM Conference on Data Mining (2007), the IEEE ICDM Conference (2007), the WWW Conference (2009), and the IEEE ICDM Conference (2009). He served as an associate editor of the *IEEE Transactions on Knowledge and Data Engineering Journal* from 2004 to 2008. He is an associate editor of the *ACM TKDD Journal*, an action editor of the *Data Mining and Knowledge Discovery Journal*, an associate editor of *ACM SIGKDD Explorations*, and an associate editor of the *Knowledge and Information Systems Journal*. He is a fellow of the IEEE for "*contributions to knowledge discovery and data mining techniques*," and a life-member of the ACM.

Chandan K. Reddy is an Assistant Professor in the Department of Computer Science at Wayne State University. He received his Ph.D. from Cornell University and M.S. from Michigan State University. His primary research interests are in the areas of data mining and machine learning with applications to healthcare, bioinformatics, and social network analysis. His research is funded by the National Science Foundation, the National Institutes of Health, Department of Transportation, and the Susan G. Komen for the Cure Foundation. He has published over 40 peer-reviewed articles in leading conferences and journals. He received the Best Application Paper Award at the ACM SIGKDD conference in 2010 and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a member of IEEE, ACM, and SIAM.

Contributors

Ayan Acharya
University of Texas
Austin, Texas

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, New York

Amrudin Agovic
Reliancy, LLC
Saint Louis Park, Minnesota

Mohammad Al Hasan
Indiana University - Purdue University
Indianapolis, Indiana

Salem Alelyani
Arizona State University
Tempe, Arizona

David C. Anastasiu
University of Minnesota
Minneapolis, Minnesota

Bill Andreopoulos
Lawrence Berkeley National Laboratory
Berkeley, California

James Bailey
The University of Melbourne
Melbourne, Australia

Arindam Banerjee
University of Minnesota
Minneapolis, Minnesota

Sandra Batista
Duke University
Durham, North Carolina

Shiyu Chang
University of Illinois at Urbana-Champaign
Urbana, Illinois

Wei Cheng
University of North Carolina
Chapel Hill, North Carolina

Hongbo Deng
University of Illinois at Urbana-Champaign
Urbana, Illinois

Cha-charis Ding
University of Texas
Arlington, Texas

Martin Ester
Simon Fraser University
British Columbia, Canada

S M Faisal
The Ohio State University
Columbus, Ohio

Joydeep Ghosh
University of Texas
Austin, Texas

Dimitrios Gunopulos
University of Athens
Athens, Greece

Jiawei Han
University of Illinois at Urbana-Champaign
Urbana, Illinois

Alexander Hinneburg
Martin-Luther University
Halle/Saale, Germany

Thomas S. Huang
University of Illinois at Urbana-Champaign
Urbana, Illinois

U Kang
KAIST
Seoul, Korea

George Karypis
University of Minnesota
Minneapolis, Minnesota

Dimitrios Kotsakos
University of Athens
Athens, Greece

Tao Li
Florida International University
Miami, Florida

Zhongmou Li
Rutgers University
New Brunswick, New Jersey

Huan Liu
Arizona State University
Tempe, Arizona

Jialu Liu
University of Illinois at Urbana-Champaign
Urbana, Illinois

Srinivasan Parthasarathy
The Ohio State University
Columbus, Ohio

Guo-Jun Qi
University of Illinois at Urbana-Champaign
Urbana, Illinois

Chandan K. Reddy
Wayne State University
Detroit, Michigan

Andrea Tagarelli
University of Calabria
Arcavacata di Rende, Italy

Jiliang Tang
Arizona State University
Tempe, Arizona

Hanghang Tong
IBM T. J. Watson Research Center
Yorktown Heights, New York

Goce Trajcevski
Northwestern University
Evanston, Illinois

Min-Hsuan Tsai
University of Illinois at Urbana-Champaign
Urbana, Illinois

Shen-Fu Tsai
Microsoft Inc.
Redmond, Washington

Bhanukiran Vinzamuri
Wayne State University
Detroit, Michigan

Wei Wang
University of California
Los Angeles, California

Hui Xiong
Rutgers University
New Brunswick, New Jersey

Mohammed J. Zaki
Rensselaer Polytechnic Institute
Troy, New York

Arthur Zimek
University of Alberta
Edmonton, Canada