Contents

List of Figures

List of Tables

List of Algorithms

Preface to the Second Edition

The monograph *Data Clustering: Theory, Algorithms, and Applications* was published in 2007. Starting with the common ground and knowledge for data clustering, the monograph focuses on several popular clustering algorithms and groups them according to some specific baseline methodologies, such as hierarchical, center-based, and search-based methods. Since the publication of this monograph, development in the subject area has exploded in many different directions, especially in clustering algorithms for big data and open-source software for cluster analysis.

This second edition aims to reflect some of these new developments. In addition to expanding some existing chapters, we added a few new chapters (e.g., Chapters 16, 19, 20, 21, and 22). Chapter 16 introduces scalable clustering algorithms for big data. Chapters 19 discusses open-source software that can be used to perform data clustering. Chapter 20 introduces a lightweight Java framework that can be used by researchers and practitioners to develop and test clustering algorithms. Chapters 21 and 22 demonstrate additional applications of data clustering in different areas. We also removed Chapters 19 and 20 from the first edition. We removed Chapter 19, which is about data clustering in MATLAB, for two reasons: first, MATLAB is a scripting language and is not efficient for developing clustering algorithms; second, the R programming language has become popular since the first edition. We removed Chapter 20, which is about data clustering in C/C++ because there is a whole book (i.e., Gan, 2011) devoted to this topic.

As in the first edition, we made no attempt to provide a comprehensive coverage of the subject area. The second edition follows the same structure as the first edition and covers a few clustering algorithms of different types. The criteria for selecting these algorithms are personal rather than scientific, and it is highly possible that some of those famous algorithms deserving an entry are absent. We apologize sincerely for any such unintentional omissions and for having to leave out many contributions. In this second edition, we also removed the author index of the first edition because the names of authors after the first or the second author in multi-author articles do not appear in the referenced context.

Java code and relevant data sets used in this new edition are available online from `https://bookstore.siam.org/mn05/bonus`. We hope that this new edition continues to provide researchers, students, and practitioners from a variety of disciplines with a useful reference to theory, algorithms, and applications of data clustering.

We would like to acknowledge Paula M. Callaghan from the Society for Industrial and Applied Mathematics (SIAM) for encouraging us to work on the second edition. We are deeply grateful for the researchers and readers who have published reviews for the first edition of this book: Dick Burkhart, David J. Hand, Fatih Kurugollu, Kenneth Joseph Ryan, Jin-Hong Park, and Hao Helen Zhang. We apologize again if we missed someone here. We thank the anonymous reviewers who provided valuable feedback and suggestions for the second edition.

We would also like to acknowledge the following financial support: the CAE (Centers of Actuarial Excellence) grant awarded to the University of Connecticut by the Society of Actuaries, the Key Projects of the National Natural Science Foundation of China (grant 71431008, the

Modeling of Financial Data with High Dimensional, Nonlinear, Nonstationary, Time-Varying and Their Application), the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada's Discovery Grant Program and Collaborative Research Development Program, and Mathematics for Information Technology and Complex Systems.

May 31, 2020
Guojun, Chaoqun, and Jianhong
Storrs, CT, USA, Changsha, Hunan, China, and Toronto, ON, Canada

Preface to the First Edition

Cluster analysis is an unsupervised process that divides a set of objects into homogeneous groups. There have been many clustering algorithms scattered in publications in very diversified areas such as pattern recognition, artificial intelligence, information technology, image processing, biology, psychology, and marketing. As such, readers and users often find it very difficult to identify an appropriate algorithm for their applications and/or to compare novel ideas with existing results.

In this monograph, we shall focus on a small number of popular clustering algorithms and group them according to some specific baseline methodologies, such as hierarchical, center-based, and search-based methods. We shall, of course, start with the common ground and knowledge for cluster analysis, including the classification of data and the corresponding similarity measures, and we shall also provide examples of clustering applications to illustrate the advantages and shortcomings of different clustering architectures and algorithms.

This monograph is intended not only for statistics, applied mathematics, and computer science senior undergraduates and graduates, but also for research scientists who need cluster analysis to deal with data. It may be used as a textbook for introductory courses in cluster analysis or as source material for an introductory course in data mining at the graduate level. We assume that the reader is familiar with elementary linear algebra, calculus, and basic statistical concepts and methods.

The book is divided into four parts: basic concepts (clustering, data, and similarity measures), algorithms, applications, and programming languages. We now briefly describe the content of each chapter.

Chapter 1. Data clustering. In this chapter, we introduce the basic concepts of clustering. Cluster analysis is defined as a way to create groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct. Some working definitions of clusters are discussed, and several popular books relevant to cluster analysis are introduced.

Chapter 2. Data types. The type of data is directly associated with data clustering, and it is a major factor to consider in choosing an appropriate clustering algorithm. Five data types are discussed in this chapter: categorical, binary, transaction, symbolic, and time series. They share a common feature that nonnumerical similarity measures must be used. There are many other data types, such as image data, that are not discussed here, though we believe that once readers get familiar with these basic types of data, they should be able to adjust the algorithms accordingly.

Chapter 3. Scale conversion. Scale conversion is concerned with the transformation between different types of variables. For example, one may convert a continuous measured variable to an interval variable. In this chapter, we first review several scale conversion techniques and then discuss several approaches for categorizing numerical data.

Chapter 4. Data standardization and transformation. In many situations, raw data should be normalized and/or transformed before a cluster analysis. One reason to do this is that objects

in raw data may be described by variables measured with different scales; another reason is to reduce the size of the data to improve the effectiveness of clustering algorithms. Therefore, we present several data standardization and transformation techniques in this chapter.

Chapter 5. Data visualization. Data visualization is vital in the final step of data-mining applications. This chapter introduces various techniques of visualization with an emphasis on visualization of clustered data. Some dimension reduction techniques, such as multidimensional scaling (MDS) and self-organizing maps (SOMs), are discussed.

Chapter 6. Similarity and dissimilarity measures. In the literature of data clustering, a similarity measure or distance (dissimilarity measure) is used to quantitatively describe the similarity or dissimilarity of two data points or two clusters. Similarity and distance measures are basic elements of a clustering algorithm, without which no meaningful cluster analysis is possible. Due to the important role of similarity and distance measures in cluster analysis, we present a comprehensive discussion of different measures for various types of data in this chapter. We also introduce measures between points and measures between clusters.

Chapter 7. Hierarchical clustering techniques. Hierarchical clustering algorithms and partitioning algorithms are two major clustering algorithms. Unlike partitioning algorithms, which divide a data set into a single partition, hierarchical algorithms divide a data set into a sequence of nested partitions. There are two major hierarchical algorithms: agglomerative algorithms and divisive algorithms. Agglomerative algorithms start with every single object in a single cluster, while divisive ones start with all objects in one cluster and repeat splitting large clusters into small pieces. In this chapter, we present representations of hierarchical clustering and several popular hierarchical clustering algorithms.

Chapter 8. Fuzzy clustering algorithms. Clustering algorithms can be classified into two categories: hard clustering algorithms and fuzzy clustering algorithms. Unlike hard clustering algorithms, which require that each data point of the data set belong to one and only one cluster, fuzzy clustering algorithms allow a data point to belong to two or more clusters with different probabilities. There is also a huge number of published works related to fuzzy clustering. In this chapter, we review some basic concepts of fuzzy logic and present three well-known fuzzy clustering algorithms: fuzzy k-means, fuzzy k-modes, and c-means.

Chapter 9. Center-based clustering algorithms. Compared to other types of clustering algorithms, center-based clustering algorithms are more suitable for clustering large data sets and high-dimensional data sets. Several well-known center-based clustering algorithms (e.g., k-means, k-modes) are presented and discussed in this chapter.

Chapter 10. Search-based clustering algorithms. A well-known problem associated with most of the clustering algorithms is that they may not be able to find the globally optimal clustering that fits the data set, since these algorithms will stop if they find a local optimal partition of the data set. This problem led to the invention of search-based clustering algorithms to search the solution space and find a globally optimal clustering that fits the data set. In this chapter, we present several clustering algorithms based on genetic algorithms, tabu search algorithms, and simulated annealing algorithms.

Chapter 11. Graph-based clustering algorithms. Graph-based clustering algorithms cluster a data set by clustering the graph or hypergraph constructed from the data set. The construction of a graph or hypergraph is usually based on the dissimilarity matrix of the data set under consideration. In this chapter, we present several graph-based clustering algorithms that do not use the spectral graph partition techniques, although we also list a few references related to spectral graph partition techniques.

Chapter 12. Grid-based clustering algorithms. In general, a grid-based clustering algorithm consists of the following five basic steps: partitioning the data space into a finite number of cells (or creating grid structure), estimating the cell density for each cell, sorting the cells according to their densities, identifying cluster centers, and traversal of neighbor cells. A major

advantage of grid-based clustering is that it significantly reduces the computational complexity. Some recent works on grid-based clustering are presented in this chapter.

Chapter 13. Density-based clustering algorithms. The density-based clustering approach is capable of finding arbitrarily shaped clusters, where clusters are defined as dense regions separated by low-density regions. Usually, density-based clustering algorithms are not suitable for high-dimensional data sets, since data points are sparse in high-dimensional spaces. Five density-based clustering algorithms (DBSCAN, BRIDGE, DBCLASD, DENCLUE, and CUBN) are presented in this chapter.

Chapter 14. Model-based clustering algorithms. In the framework of model-based clustering algorithms, the data are assumed to come from a mixture of probability distributions, each of which represents a different cluster. There is a huge number of published works related to model-based clustering algorithms. In particular, there are more than 400 articles devoted to the development and discussion of the expectation-maximization (EM) algorithm. In this chapter, we introduce model-based clustering and present two model-based clustering algorithms: COOLCAT and STUCCO.

Chapter 15. Subspace clustering. Subspace clustering is a relatively new concept. After the first subspace clustering algorithm, CLIQUE, was published by the IBM group, many subspace clustering algorithms were developed and studied. One feature of the subspace clustering algorithms is that they are capable of identifying different clusters embedded in different subspaces of the high-dimensional data. Several subspace clustering algorithms are presented in this chapter, including the neural network–inspired algorithm PART.

Chapter 16. Miscellaneous algorithms. This chapter introduces some clustering algorithms for clustering time series, data streams, and transaction data. Proximity measures for these data and several related clustering algorithms are presented.

Chapter 17. Evaluation of clustering algorithms. Clustering is an unsupervised process and there are no predefined classes and no examples to show that the clusters found by the clustering algorithms are valid. Usually one or more validity criteria, presented in this chapter, are required to verify the clustering result of one algorithm or to compare the clustering results of different algorithms.

Chapter 18. Clustering gene expression data. As an application of cluster analysis, gene expression data clustering is introduced in this chapter. The background and similarity measures for gene expression data are introduced. Clustering a real set of gene expression data with the fuzzy subspace clustering (FSC) algorithm is presented.

Chapter 19. Data clustering in MATLAB. In this chapter, we show how to perform clustering in MATLAB in the following three aspects. Firstly, we introduce some MATLAB commands related to file operations, since the first thing to do about clustering is to load data into MATLAB, and data are usually stored in a text file. Secondly, we introduce MATLAB M-files, MEX-files, and MAT-files in order to demonstrate how to code algorithms and save current work. Finally, we present several MATLAB codes, which can be found in Appendix C.

Chapter 20. Clustering in C/C++. C++ is an object-oriented programming language built on the C language. In this last chapter of the book, we introduce the Standard Template Library (STL) in C++ and C/C++ program compilation. C++ data structure for data clustering is introduced. This chapter assumes that readers have basic knowledge of the C/C++ language.

This monograph has grown and evolved from a few collaborative projects for industrial applications undertaken by the Laboratory for Industrial and Applied Mathematics at York University, some of which are in collaboration with Generation 5 Mathematical Technologies, Inc. We would like to thank the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada's Discovery Grant Program and Collaborative Research Development Program, and Mathematics for Information Technology and Complex Systems for their support.

Index