Preface

Clustering, the unsupervised classification of patterns into groups, is one of the most important tasks in exploratory data analysis. Primary goals of clustering include gaining insight into, classifying, and compressing data. Clustering has a long and rich history that spans a variety of scientific disciplines including anthropology, biology, medicine, psychology, statistics, mathematics, engineering, and computer science. As a result, numerous clustering algorithms have been proposed since the early 1950s. Among these algorithms, partitional (nonhierarchical) ones have found many applications, especially in engineering and computer science.

The goal of this volume is to summarize the state of the art in partitional clustering. The intended audience includes researchers and practitioners, who are increasingly using partitional clustering algorithms to analyze their data.

The volume opens with a chapter on model-based clustering entitled "Recent Developments in Model-Based Clustering with Applications." In this chapter, Melnykov et al. review the latest developments in this field including semi-supervised clustering, nonparametric mixture modeling, initialization strategies, merging mixture components, and handling spurious solutions. In "Accelerating Lloyd's Algorithm for *k*-Means Clustering," Hamerly and Drake present a survey of triangle inequality-based acceleration techniques for the celebrated *k*-means clustering algorithm. Based on extensive experiments, the authors conclude that a suitable application of the triangle inequality can provide dramatic speedups of up to 40x over a naive implementation of the standard Lloyd's algorithm. In another *k*-means related chapter entitled "Linear, Deterministic, and Order-Invariant Initialization Methods for the *k*-Means Clustering Algorithm," Celebi and Kingravi investigate the empirical performance of six linear, deterministic, and order-invariant *k*-means initialization methods on a large and diverse collection of data sets from the UCI Machine Learning Repository. Their results demonstrate that two relatively unknown hierarchical initialization methods outperform the remaining four methods with respect to two objective effectiveness criteria. They also show that one of the most recent initialization methods performs surprisingly poorly. In "Nonsmooth

Optimization Based Algorithms in Cluster Analysis," Bagirov and Mohebi approach the problem of partitional clustering using nonsmooth and nonconvex optimization. Based on this formulation, they design an efficient incremental algorithm similar to k-means that can handle ℓ_1 and ℓ_∞ norms besides the commonly used ℓ_2 norm.

In "Fuzzy Clustering Algorithms and Validity Indices for Distributed Data," Vendramin et al. present a framework to generalize several fuzzy clustering algorithms to handle distributed data without resorting to approximations. This framework also allows the exact calculation of a variety of relative validity indices to evaluate the quality of fuzzy partitions. The authors also describe a procedure based on this framework for the estimation of the number of clusters in parallel and distributed settings.

In "Density Based Clustering: Alternatives to DBSCAN," Braune et al. propose two algorithms similar to the celebrated DBSCAN algorithm. Unlike DBSCAN, both algorithms require only one input parameter. One of these algorithms gives similar results to DBSCAN while being able to assign multiple cluster labels to a data point, whereas the other one is significantly faster than DBSCAN.

In "Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering," Kuang et al. propose a novel formulation of the nonnegative matrix factorization (NMF) problem based on the block coordinate descent algorithm. The authors present a clustering algorithm based on this formulation and prove its convergence. In addition to extending this framework to sparse and weakly supervised clustering, the authors describe a method to determine the number of clusters based on random sampling and consensus clustering. Experiments on various real-world document data sets demonstrate the advantage of the proposed NMF clustering algorithm in terms of clustering quality, convergence behavior, sparseness, and consistency.

In "Overview of Overlapping Partitional Clustering Methods," Ben N'Cir et al. review the fundamental concepts of partitional overlapping clustering and present a survey of widely known partitional overlapping clustering algorithms as well as techniques to evaluate the quality of non-disjoint partitions. Furthermore, the authors investigate the ability of various clustering algorithms to generate overlapping partitions from multi-labeled real-world data sets.

In "On Semi-Supervised Clustering," Bongini et al. present a survey of semi-supervised clustering (SSC). The authors first give a conceptual overview of the field and then discuss some of the most important algorithms for SSC including COP-COBWEB, COP-kMeans, HMRF k-means, seeded k-means, constrained k-means, and active fuzzy constrained clustering. The authors conclude with a discussion of future directions for this relatively new field.

In "Consensus of Clusterings Based on High-Order Dissimilarities," Aidos and Fred describe a novel dynamic clustering algorithm based on a recently proposed dissimilarity measure called "dissimilarity increments." Starting from an initial partition, this algorithm incorporates a merging strategy based on either a likelihood-ratio test or a test based on the minimum description length principle. The authors address the initialization dependence of their algorithm using a consensus function-based combination strategy. Finally, the best partition is selected using a

criterion based on dissimilarity increments. Experimental results demonstrate the effectiveness of the proposed algorithm on a variety of synthetic as well as real-world data sets.

In "Hubness-Based Clustering of High-Dimensional Data," Tomašev et al. investigate the hubness phenomenon observed in k-nearest neighbor graphs of high-dimensional data. The authors demonstrate that hubness complicates the cluster discovery process by reducing the separability of clusters. The authors then review some recent clustering algorithms that exploit the hubness phenomenon and then introduce a kernel-based clustering algorithm that does not restrict the shape of the clusters to hyperspheres.

A chapter entitled "Clustering for Monitoring Distributed Data Streams" by Barouti et al. completes the volume. The authors consider an application of clustering to monitoring data streams in a distributed system. Unlike conventional clustering algorithms that group similar data points into clusters, monitoring requires that clusters with dissimilar points cancel each other as much as possible. The authors devise a novel clustering algorithm to tackle this problem and demonstrate that it yields a reduction in communication load.

We hope that this volume focused on partitional clustering will demonstrate the significant progress that has occurred in this field in recent years. We also hope that the developments reported in this volume will motivate further research in this exciting field.

Shreveport, LA, USA M. Emre Celebi

http://www.springer.com/978-3-319-09258-4

Partitional Clustering Algorithms
Celebi, M.E. (Ed.)
2015, X, 415 p. 78 illus., 45 illus. in color., Hardcover
ISBN: 978-3-319-09258-4