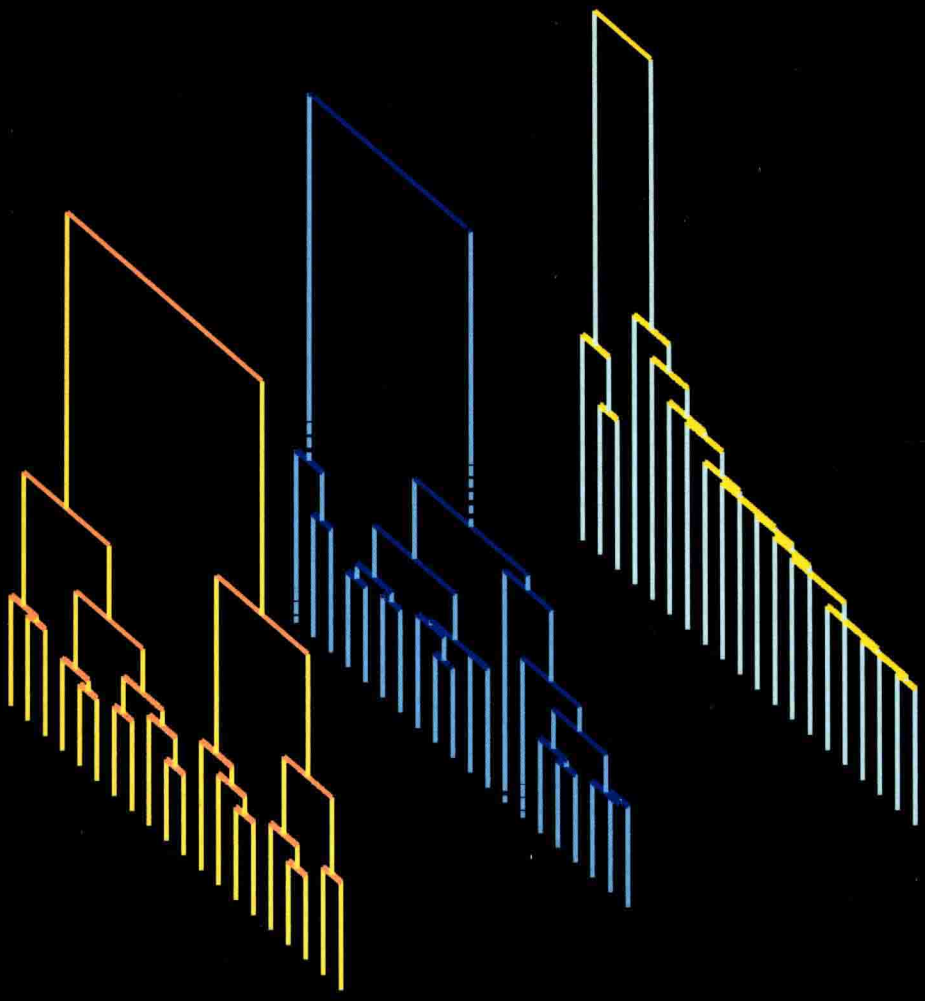
Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Data Clustering in C++

An Object-Oriented Approach



Guojun Gan

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Data Clustering in C++

An Object-Oriented Approach

Guojun Gan

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

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

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4398-6223-0 (Hardback)

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

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

Data Clustering in C++

An Object-Oriented Approach

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

SERIES EDITOR

Vipin Kumar
University of Minnesota
Department of Computer Science and Engineering
Minneapolis, Minnesota, U.S.A

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

UNDERSTANDING COMPLEX DATASETS:
DATA MINING WITH MATRIX DECOMPOSITIONS
David Skillicorn

COMPUTATIONAL METHODS OF FEATURE SELECTION
Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: ADVANCES IN ALGORITHMS, THEORY, AND APPLICATIONS
Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT
David Skillicorn

MULTIMEDIA DATA MINING: A SYSTEMATIC INTRODUCTION TO CONCEPTS AND THEORY
Zhongfei Zhang and Ruofei Zhang

NEXT GENERATION OF DATA MINING
Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

DATA MINING FOR DESIGN AND MARKETING
Yukio Ohsawa and Katsutoshi Yada

THE TOP TEN ALGORITHMS IN DATA MINING
Xindong Wu and Vipin Kumar

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, SECOND EDITION
Harvey J. Miller and Jiawei Han

TEXT MINING: CLASSIFICATION, CLUSTERING, AND APPLICATIONS
Ashok N. Srivastava and Mehran Sahami

BIOLOGICAL DATA MINING
Jake Y. Chen and Stefano Lonardi

INFORMATION DISCOVERY ON ELECTRONIC HEALTH RECORDS
Vagelis Hristidis

TEMPORAL DATA MINING
Theophano Mitsa

RELATIONAL DATA CLUSTERING: MODELS, ALGORITHMS, AND APPLICATIONS
Bo Long, Zhongfei Zhang, and Philip S. Yu

KNOWLEDGE DISCOVERY FROM DATA STREAMS
João Gama

STATISTICAL DATA MINING USING SAS APPLICATIONS, SECOND EDITION
George Fernandez

INTRODUCTION TO PRIVACY-PRESERVING DATA PUBLISHING: CONCEPTS AND TECHNIQUES
Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu

HANDBOOK OF EDUCATIONAL DATA MINING
Cristóbal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker

DATA MINING WITH R: LEARNING WITH CASE STUDIES
Luís Torgo

MINING SOFTWARE SPECIFICATIONS: METHODOLOGIES AND APPLICATIONS
David Lo, Siau-Cheng Khoo, Jiawei Han, and Chao Liu

DATA CLUSTERING IN C++: AN OBJECT-ORIENTED APPROACH
Guojun Gan

Dedication

To my grandmother and my parents

Preface

Data clustering is a highly interdisciplinary field whose goal is to divide a set of objects into homogeneous groups such that objects in the same group are similar and objects in different groups are quite distinct. Thousands of papers and a number of books on data clustering have been published over the past 50 years. However, almost all papers and books focus on the theory of data clustering. There are few books that teach people how to implement data clustering algorithms.

This book was written for anyone who wants to implement data clustering algorithms and for those who want to implement new data clustering algorithms in a better way. Using object-oriented design and programming techniques, I have exploited the commonalities of all data clustering algorithms to create a flexible set of reusable classes that simplifies the implementation of any data clustering algorithm. Readers can follow me through the development of the base data clustering classes and several popular data clustering algorithms.

This book focuses on how to implement data clustering algorithms in an object-oriented way. Other topics of clustering such as data pre-processing, data visualization, cluster visualization, and cluster interpretation are touched but not in detail. In this book, I used a direct and simple way to implement data clustering algorithms so that readers can understand the methodology easily. I also present the material in this book in a straightforward way. When I introduce a class, I present and explain the class method by method rather than present and go through the whole implementation of the class.

Complete listings of classes, examples, unit test cases, and GNU configuration files are included in the appendices of this book as well as in the CD-ROM of the book. I have tested the code under Unix-like platforms (e.g., Ubuntu and Cygwin) and Microsoft Windows XP. The only requirements to compile the code are a modern C++ compiler and the Boost C++ libraries.

This book is divided into three parts: Data Clustering and C++ Preliminaries, A C++ Data Clustering Framework, and Data Clustering Algorithms. The first part reviews some basic concepts of data clustering, the unified modeling language, object-oriented programming in C++, and design patterns. The second part develops the data clustering base classes. The third part implements several popular data clustering algorithms. The content of each chapter is described briefly below.

Chapter 1. Introduction to Data Clustering. In this chapter, we review some basic concepts of data clustering. The clustering process, data types, similarity and dissimilarity measures, hierarchical and partitional clustering algorithms, cluster validity, and applications of data clustering are briefly introduced. In addition, a list of survey papers and books related to data clustering are presented.

Chapter 2. The Unified Modeling Language. The Unified Modeling Language (UML) is a general-purpose modeling language that includes a set of standardized graphic notation to create visual models of software systems. In this chapter, we introduce several UML diagrams such as class diagrams, use-case diagrams, and activity diagrams. Illustrations of these UML diagrams are presented.

Chapter 3. Object-Oriented Programming and C++. Object-oriented programming is a programming paradigm that is based on the concept of objects, which are reusable components. Object-oriented programming has three pillars: encapsulation, inheritance, and polymorphism. In this chapter, these three pillars are introduced and illustrated with simple programs in C++. The exception handling ability of C++ is also discussed in this chapter.

Chapter 4. Design Patterns. Design patterns are reusable designs just as objects are reusable components. In fact, a design pattern is a general reusable solution to a problem that occurs over and over again in software design. In this chapter, several design patterns are described and illustrated by simple C++ programs.

Chapter 5. C++ Libraries and Tools. As an object-oriented programming language, C++ has many well-designed and useful libraries. In this chapter, the standard template library (STL) and several Boost C++ libraries are introduced and illustrated by C++ programs. The GNU build system (i.e., GNU Autotools) and the Cygwin system, which simulates a Unix-like platform under Microsoft Windows, are also introduced.

Chapter 6. The Clustering Library. This chapter introduces the file system of the clustering library, which is a collection of reusable classes used to develop clustering algorithms. The structure of the library and file name convention are introduced. In addition, the GNU configuration files, the error handling class, unit testing, and compilation of the clustering library are described.

Chapter 7. Datasets. This chapter introduces the design and implementation of datasets. In this book, we assume that a dataset consists of a set of records and a record is a vector of values. The attribute value class, the attribute information class, the schema class, the record class, and the dataset class are introduced in this chapter. These classes are illustrated by an example in C++.

Chapter 8. Clusters. A cluster is a collection of records. In this chapter, the cluster class and its child classes such as the center cluster class and the subspace cluster class are introduced. In addition, partitional clustering class and hierarchical clustering class are also introduced.

Chapter 9. Dissimilarity Measures. Dissimilarity or distance measures are an important part of most clustering algorithms. In this chapter, the design of the distance base class is introduced. Several popular distance measures such as the Euclidean distance, the simple matching distance, and the mixed distance are introduced. In this chapter, we also introduce the implementation of the Mahalanobis distance.

Chapter 10. Clustering Algorithms. This chapter introduces the design and implementation of the clustering algorithm base class. All data clustering algorithms have three components: arguments or parameters, clustering method, and clustering results. In this chapter, we introduce the argument class, the result class, and the base algorithm class. A dummy clustering algorithm is used to illustrate the usage of the base clustering algorithm class.

Chapter 11. Utility Classes. This chapter, as its title implies, introduces several useful utility classes used frequently in the clustering library. Two template classes, the container class and the double-key map class, are introduced in this chapter. A CSV (comma-separated values) dataset reader class and a multivariate Gaussian mixture dataset generator class are also introduced in this chapter. In addition, two hierarchical tree visitor classes, the join value visitor class and the partition creation visitor class, are introduced in this chapter. This chapter also includes two classes that provide functionalities to draw dendrograms in EPS (Encapsulated PostScript) figures from hierarchical clustering trees.

Chapter 12. Agglomerative Hierarchical Algorithms. This chapter introduces the implementations of several agglomerative hierarchical clustering algorithms that are based on the Lance-Williams framework. In this chapter, single linkage, complete linkage, group average, weighted group average, centroid, median, and Ward's method are implemented and illustrated by a synthetic dataset and the Iris dataset.

Chapter 13. DIANA. This chapter introduces a divisive hierarchical clustering algorithm and its implementation. The algorithm is illustrated by a synthetic dataset and the Iris dataset.

Chapter 14. The k-means Algorithm. This chapter introduces the standard k-means algorithm and its implementation. A synthetic dataset and the Iris dataset are used to illustrate the algorithm.

Chapter 15. The c-means Algorithm. This chapter introduces the fuzzy c-means algorithm and its implementation. The algorithm is also illustrated by a synthetic dataset and the Iris dataset.

Chapter 16. The k-prototype Algorithm. This chapter introduces the k-prototype algorithm and its implementation. This algorithm was designed to cluster mixed-type data. A numeric dataset (the Iris dataset), a categorical dataset (the Soybean dataset), and a mixed-type dataset (the heart dataset) are used to illustrate the algorithm.

Chapter 17. The Genetic k-modes Algorithm. This chapter introduces the genetic k-modes algorithm and its implementation. A brief introduction to the genetic algorithm is also presented. The Soybean dataset is used to illustrate the algorithm.

Chapter 18. The FSC Algorithm. This chapter introduces the fuzzy subspace clustering (FSC) algorithm and its implementation. The algorithm is illustrated by a synthetic dataset and the Iris dataset.

Chapter 19. The Gaussian Mixture Model Clustering Algorithm. This chapter introduces a clustering algorithm based on the Gaussian mixture model.

Chapter 20. A Parallel k-means Algorithm. This chapter introduces a simple parallel version of the k-means algorithm based on the message passing interface and the Boost MPI library.

Chapters 2–5 introduce programming related materials. Readers who are already familiar with object-oriented programming in C++ can skip those chapters. Chapters 6–11 introduce the base clustering classes and some utility classes. Chapter 12 includes several agglomerative hierarchical clustering algorithms. Each one of the last eight chapters is devoted to one particular clustering algorithm. The eight chapters introduce and implement a diverse set of clustering algorithms such as divisive clustering, center-based clustering, fuzzy clustering, mixed-type data clustering, search-based clustering, subspace clustering, mode-based clustering, and parallel data clustering.

A key to learning a clustering algorithm is to implement and experiment the clustering algorithm. I encourage readers to compile and experiment the examples included in this book. After getting familiar with the classes and their usage, readers can implement new clustering algorithms using these classes or even improve the designs and implementations presented in this book. To this end, I included some exercises and projects in the appendix of this book.

This book grew out of my wish to help undergraduate and graduate students who study data clustering to learn how to implement clustering algorithms and how to do it in a better way. When I was a PhD student, there were no books or papers to teach me how to implement clustering algorithms. It took me a long time to implement my first clustering algorithm. The clustering programs I wrote at that time were just C programs written in C++. It has taken me years to learn how to use the powerful C++ language in the right way. With the help of this book, readers should be able to learn how to implement clustering algorithms and how to do it in a better way in a short period of time.

I would like to take this opportunity to thank my boss, Dr. Hong Xie, who taught me how to write in an effective and rigorous way. I would also like to thank my ex-boss, Dr. Matthew Willis, who taught me how to program in C++ in a better way. I thank my PhD supervisor, Dr. Jianhong Wu, who brought me into the field of data clustering. Finally, I would like to thank my wife, Xiaoying, and my children, Albert and Ella, for their support.

Guojun Gan
Toronto, Ontario
December 31, 2010

Contents