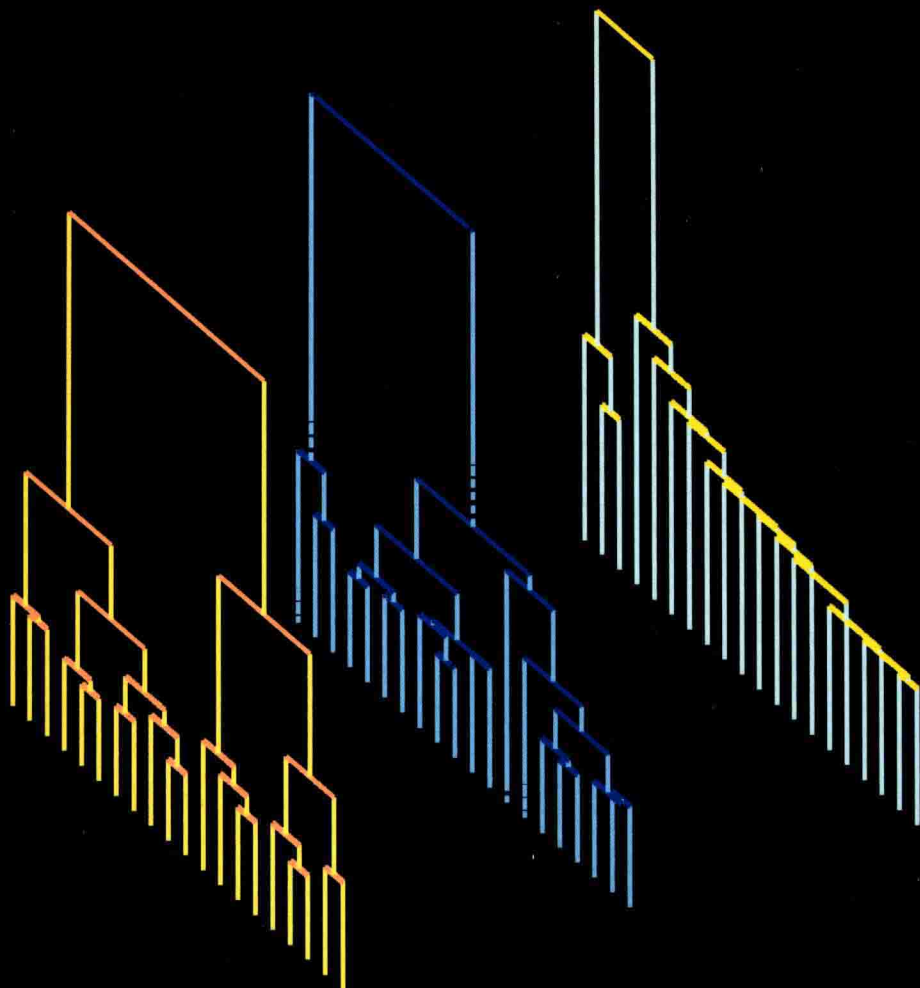


Chapman & Hall/CRC  
Data Mining and Knowledge Discovery Series

# 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

© 2011 by Taylor and Francis Group, LLC  
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

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)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

**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

List of Figures	xv
List of Tables	xix
Preface	xxi
<b>I Data Clustering and C++ Preliminaries</b>	<b>1</b>
<b>1 Introduction to Data Clustering</b>	<b>3</b>
1.1 Data Clustering . . . . .	3
1.1.1 Clustering versus Classification . . . . .	4
1.1.2 Definition of Clusters . . . . .	5
1.2 Data Types . . . . .	7
1.3 Dissimilarity and Similarity Measures . . . . .	8
1.3.1 Measures for Continuous Data . . . . .	9
1.3.2 Measures for Discrete Data . . . . .	10
1.3.3 Measures for Mixed-Type Data . . . . .	10
1.4 Hierarchical Clustering Algorithms . . . . .	11
1.4.1 Agglomerative Hierarchical Algorithms . . . . .	12
1.4.2 Divisive Hierarchical Algorithms . . . . .	14
1.4.3 Other Hierarchical Algorithms . . . . .	14
1.4.4 Dendrograms . . . . .	15
1.5 Partitional Clustering Algorithms . . . . .	15
1.5.1 Center-Based Clustering Algorithms . . . . .	17
1.5.2 Search-Based Clustering Algorithms . . . . .	18
1.5.3 Graph-Based Clustering Algorithms . . . . .	19
1.5.4 Grid-Based Clustering Algorithms . . . . .	20
1.5.5 Density-Based Clustering Algorithms . . . . .	20
1.5.6 Model-Based Clustering Algorithms . . . . .	21
1.5.7 Subspace Clustering Algorithms . . . . .	22
1.5.8 Neural Network-Based Clustering Algorithms . . . . .	22
1.5.9 Fuzzy Clustering Algorithms . . . . .	23
1.6 Cluster Validity . . . . .	23
1.7 Clustering Applications . . . . .	24
1.8 Literature of Clustering Algorithms . . . . .	25
1.8.1 Books on Data Clustering . . . . .	25

1.8.2	Surveys on Data Clustering . . . . .	26
1.9	Summary . . . . .	28
<b>2</b>	<b>The Unified Modeling Language</b>	<b>29</b>
2.1	Package Diagrams . . . . .	29
2.2	Class Diagrams . . . . .	32
2.3	Use Case Diagrams . . . . .	36
2.4	Activity Diagrams . . . . .	38
2.5	Notes . . . . .	39
2.6	Summary . . . . .	40
<b>3</b>	<b>Object-Oriented Programming and C++</b>	<b>41</b>
3.1	Object-Oriented Programming . . . . .	41
3.2	The C++ Programming Language . . . . .	42
3.3	Encapsulation . . . . .	45
3.4	Inheritance . . . . .	48
3.5	Polymorphism . . . . .	50
3.5.1	Dynamic Polymorphism . . . . .	51
3.5.2	Static Polymorphism . . . . .	52
3.6	Exception Handling . . . . .	54
3.7	Summary . . . . .	56
<b>4</b>	<b>Design Patterns</b>	<b>57</b>
4.1	Singleton . . . . .	58
4.2	Composite . . . . .	61
4.3	Prototype . . . . .	64
4.4	Strategy . . . . .	67
4.5	Template Method . . . . .	69
4.6	Visitor . . . . .	72
4.7	Summary . . . . .	75
<b>5</b>	<b>C++ Libraries and Tools</b>	<b>77</b>
5.1	The Standard Template Library . . . . .	77
5.1.1	Containers . . . . .	77
5.1.2	Iterators . . . . .	82
5.1.3	Algorithms . . . . .	84
5.2	Boost C++ Libraries . . . . .	86
5.2.1	Smart Pointers . . . . .	87
5.2.2	Variant . . . . .	89
5.2.3	Variant versus Any . . . . .	90
5.2.4	Tokenizer . . . . .	92
5.2.5	Unit Test Framework . . . . .	93
5.3	GNU Build System . . . . .	95
5.3.1	Autoconf . . . . .	96
5.3.2	Automake . . . . .	97
5.3.3	Libtool . . . . .	97

5.3.4	Using GNU Autotools . . . . .	98
5.4	Cygwin . . . . .	98
5.5	Summary . . . . .	99
<b>II</b>	<b>A C++ Data Clustering Framework</b>	<b>101</b>
<b>6</b>	<b>The Clustering Library</b>	<b>103</b>
6.1	Directory Structure and Filenames . . . . .	103
6.2	Specification Files . . . . .	105
6.2.1	configure.ac . . . . .	105
6.2.2	Makefile.am . . . . .	106
6.3	Macros and typedef Declarations . . . . .	109
6.4	Error Handling . . . . .	111
6.5	Unit Testing . . . . .	112
6.6	Compilation and Installation . . . . .	113
6.7	Summary . . . . .	114
<b>7</b>	<b>Datasets</b>	<b>115</b>
7.1	Attributes . . . . .	115
7.1.1	The Attribute Value Class . . . . .	115
7.1.2	The Base Attribute Information Class . . . . .	117
7.1.3	The Continuous Attribute Information Class . . . . .	119
7.1.4	The Discrete Attribute Information Class . . . . .	120
7.2	Records . . . . .	122
7.2.1	The Record Class . . . . .	122
7.2.2	The Schema Class . . . . .	124
7.3	Datasets . . . . .	125
7.4	A Dataset Example . . . . .	127
7.5	Summary . . . . .	130
<b>8</b>	<b>Clusters</b>	<b>131</b>
8.1	Clusters . . . . .	131
8.2	Partitional Clustering . . . . .	133
8.3	Hierarchical Clustering . . . . .	135
8.4	Summary . . . . .	138
<b>9</b>	<b>Dissimilarity Measures</b>	<b>139</b>
9.1	The Distance Base Class . . . . .	139
9.2	Minkowski Distance . . . . .	140
9.3	Euclidean Distance . . . . .	141
9.4	Simple Matching Distance . . . . .	142
9.5	Mixed Distance . . . . .	143
9.6	Mahalanobis Distance . . . . .	144
9.7	Summary . . . . .	147

<b>10 Clustering Algorithms</b>	<b>149</b>
10.1 Arguments	149
10.2 Results	150
10.3 Algorithms	151
10.4 A Dummy Clustering Algorithm	154
10.5 Summary	158
<b>11 Utility Classes</b>	<b>161</b>
11.1 The Container Class	161
11.2 The Double-Key Map Class	164
11.3 The Dataset Adapters	167
11.3.1 A CSV Dataset Reader	167
11.3.2 A Dataset Generator	170
11.3.3 A Dataset Normalizer	173
11.4 The Node Visitors	175
11.4.1 The Join Value Visitor	175
11.4.2 The Partition Creation Visitor	176
11.5 The Dendrogram Class	177
11.6 The Dendrogram Visitor	179
11.7 Summary	180
<b>III Data Clustering Algorithms</b>	<b>183</b>
<b>12 Agglomerative Hierarchical Algorithms</b>	<b>185</b>
12.1 Description of the Algorithm	185
12.2 Implementation	187
12.2.1 The Single Linkage Algorithm	192
12.2.2 The Complete Linkage Algorithm	192
12.2.3 The Group Average Algorithm	193
12.2.4 The Weighted Group Average Algorithm	194
12.2.5 The Centroid Algorithm	194
12.2.6 The Median Algorithm	195
12.2.7 Ward's Algorithm	196
12.3 Examples	197
12.3.1 The Single Linkage Algorithm	198
12.3.2 The Complete Linkage Algorithm	200
12.3.3 The Group Average Algorithm	202
12.3.4 The Weighted Group Average Algorithm	204
12.3.5 The Centroid Algorithm	207
12.3.6 The Median Algorithm	210
12.3.7 Ward's Algorithm	212
12.4 Summary	214

<b>13 DIANA</b>	<b>217</b>
13.1 Description of the Algorithm . . . . .	217
13.2 Implementation . . . . .	218
13.3 Examples . . . . .	223
13.4 Summary . . . . .	227
<b>14 The <math>k</math>-means Algorithm</b>	<b>229</b>
14.1 Description of the Algorithm . . . . .	229
14.2 Implementation . . . . .	230
14.3 Examples . . . . .	235
14.4 Summary . . . . .	240
<b>15 The c-means Algorithm</b>	<b>241</b>
15.1 Description of the Algorithm . . . . .	241
15.2 Implementaion . . . . .	242
15.3 Examples . . . . .	246
15.4 Summary . . . . .	253
<b>16 The <math>k</math>-prototypes Algorithm</b>	<b>255</b>
16.1 Description of the Algorithm . . . . .	255
16.2 Implementation . . . . .	256
16.3 Examples . . . . .	258
16.4 Summary . . . . .	263
<b>17 The Genetic <math>k</math>-modes Algorithm</b>	<b>265</b>
17.1 Description of the Algorithm . . . . .	265
17.2 Implementation . . . . .	267
17.3 Examples . . . . .	274
17.4 Summary . . . . .	277
<b>18 The FSC Algorithm</b>	<b>279</b>
18.1 Description of the Algorithm . . . . .	279
18.2 Implementation . . . . .	281
18.3 Examples . . . . .	284
18.4 Summary . . . . .	290
<b>19 The Gaussian Mixture Algorithm</b>	<b>291</b>
19.1 Description of the Algorithm . . . . .	291
19.2 Implementation . . . . .	293
19.3 Examples . . . . .	300
19.4 Summary . . . . .	306

<b>20 A Parallel <math>k</math>-means Algorithm</b>	<b>307</b>
20.1 Message Passing Interface . . . . .	307
20.2 Description of the Algorithm . . . . .	310
20.3 Implementation . . . . .	311
20.4 Examples . . . . .	316
20.5 Summary . . . . .	320
<b>A Exercises and Projects</b>	<b>323</b>
<b>B Listings</b>	<b>325</b>
B.1 Files in Folder <code>ClusLib</code> . . . . .	325
B.1.1 Configuration File <code>configure.ac</code> . . . . .	325
B.1.2 m4 Macro File <code>acinclude.m4</code> . . . . .	326
B.1.3 Makefile . . . . .	327
B.2 Files in Folder <code>cl</code> . . . . .	328
B.2.1 Makefile . . . . .	328
B.2.2 Macros and <code>typedef</code> Declarations . . . . .	328
B.2.3 Class <code>Error</code> . . . . .	329
B.3 Files in Folder <code>cl/algorithms</code> . . . . .	331
B.3.1 Makefile . . . . .	331
B.3.2 Class <code>Algorithm</code> . . . . .	332
B.3.3 Class <code>Average</code> . . . . .	334
B.3.4 Class <code>Centroid</code> . . . . .	334
B.3.5 Class <code>Cmean</code> . . . . .	335
B.3.6 Class <code>Complete</code> . . . . .	339
B.3.7 Class <code>Diana</code> . . . . .	339
B.3.8 Class <code>FSC</code> . . . . .	343
B.3.9 Class <code>GKmode</code> . . . . .	347
B.3.10 Class <code>GMC</code> . . . . .	353
B.3.11 Class <code>Kmean</code> . . . . .	358
B.3.12 Class <code>Kprototype</code> . . . . .	361
B.3.13 Class <code>LW</code> . . . . .	362
B.3.14 Class <code>Median</code> . . . . .	364
B.3.15 Class <code>Single</code> . . . . .	365
B.3.16 Class <code>Ward</code> . . . . .	366
B.3.17 Class <code>Weighted</code> . . . . .	367
B.4 Files in Folder <code>cl/clusters</code> . . . . .	368
B.4.1 Makefile . . . . .	368
B.4.2 Class <code>CenterCluster</code> . . . . .	368
B.4.3 Class <code>Cluster</code> . . . . .	369
B.4.4 Class <code>HClustering</code> . . . . .	370
B.4.5 Class <code>PClustering</code> . . . . .	372
B.4.6 Class <code>SubspaceCluster</code> . . . . .	375
B.5 Files in Folder <code>cl/datasets</code> . . . . .	376
B.5.1 Makefile . . . . .	376



B.5.2	Class AttrValue . . . . .	376
B.5.3	Class AttrInfo . . . . .	377
B.5.4	Class CAttrInfo . . . . .	379
B.5.5	Class DAttrInfo . . . . .	381
B.5.6	Class Record . . . . .	384
B.5.7	Class Schema . . . . .	386
B.5.8	Class Dataset . . . . .	388
B.6	Files in Folder cl/distances . . . . .	392
B.6.1	Makefile . . . . .	392
B.6.2	Class Distance . . . . .	392
B.6.3	Class EuclideanDistance . . . . .	393
B.6.4	Class MahalanobisDistance . . . . .	394
B.6.5	Class MinkowskiDistance . . . . .	395
B.6.6	Class MixedDistance . . . . .	396
B.6.7	Class SimpleMatchingDistance . . . . .	397
B.7	Files in Folder cl/patterns . . . . .	398
B.7.1	Makefile . . . . .	398
B.7.2	Class DendrogramVisitor . . . . .	399
B.7.3	Class InternalNode . . . . .	401
B.7.4	Class LeafNode . . . . .	403
B.7.5	Class Node . . . . .	404
B.7.6	Class NodeVisitor . . . . .	405
B.7.7	Class JoinValueVisitor . . . . .	405
B.7.8	Class PCVisitor . . . . .	407
B.8	Files in Folder cl/utilities . . . . .	408
B.8.1	Makefile . . . . .	408
B.8.2	Class Container . . . . .	409
B.8.3	Class DataAdapter . . . . .	411
B.8.4	Class DatasetGenerator . . . . .	411
B.8.5	Class DatasetNormalizer . . . . .	413
B.8.6	Class DatasetReader . . . . .	415
B.8.7	Class Dendrogram . . . . .	418
B.8.8	Class nnMap . . . . .	421
B.8.9	Matrix Functions . . . . .	423
B.8.10	Null Types . . . . .	425
B.9	Files in Folder examples . . . . .	426
B.9.1	Makefile . . . . .	426
B.9.2	Agglomerative Hierarchical Algorithms . . . . .	426
B.9.3	A Divisive Hierarchical Algorithm . . . . .	429
B.9.4	The $k$ -means Algorithm . . . . .	430
B.9.5	The $c$ -means Algorithm . . . . .	433
B.9.6	The $k$ -prototypes Algorithm . . . . .	435
B.9.7	The Genetic $k$ -modes Algorithm . . . . .	437
B.9.8	The FSC Algorithm . . . . .	439
B.9.9	The Gaussian Mixture Clustering Algorithm . . . . .	441