

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

Contents

List of Figures	xv
List of Tables	xix
Preface	xxi
I Data Clustering and C++ Preliminaries	1
1 Introduction to Data Clustering	3
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
2	The Unified Modeling Language	29
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
3	Object-Oriented Programming and C++	41
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
4	Design Patterns	57
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
5	C++ Libraries and Tools	77
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
II	A C++ Data Clustering Framework	101
6	The Clustering Library	103
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
7	Datasets	115
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
8	Clusters	131
8.1	Clusters	131
8.2	Partitional Clustering	133
8.3	Hierarchical Clustering	135
8.4	Summary	138
9	Dissimilarity Measures	139
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

10 Clustering Algorithms	149
10.1 Arguments	149
10.2 Results	150
10.3 Algorithms	151
10.4 A Dummy Clustering Algorithm	154
10.5 Summary	158
11 Utility Classes	161
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
III Data Clustering Algorithms	183
12 Agglomerative Hierarchical Algorithms	185
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

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

20 A Parallel k-means Algorithm	307
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
A Exercises and Projects	323
B Listings	325
B.1 Files in Folder ClusLib	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 Error	329
B.3 Files in Folder <code>cl/algorithms</code>	331
B.3.1 Makefile	331
B.3.2 Class Algorithm	332
B.3.3 Class Average	334
B.3.4 Class Centroid	334
B.3.5 Class Cmean	335
B.3.6 Class Complete	339
B.3.7 Class Diana	339
B.3.8 Class FSC	343
B.3.9 Class GKmode	347
B.3.10 Class GMC	353
B.3.11 Class Kmean	358
B.3.12 Class Kprototype	361
B.3.13 Class LW	362
B.3.14 Class Median	364
B.3.15 Class Single	365
B.3.16 Class Ward	366
B.3.17 Class Weighted	367
B.4 Files in Folder <code>cl/clusters</code>	368
B.4.1 Makefile	368
B.4.2 Class CenterCluster	368
B.4.3 Class Cluster	369
B.4.4 Class HClustering	370
B.4.5 Class PClustering	372
B.4.6 Class SubspaceCluster	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

B.9.10 A Parallel k -means Algorithm	444
B.10 Files in Folder <code>test-suite</code>	450
B.10.1 Makefile	450
B.10.2 The Master Test Suite	451
B.10.3 Test of <code>AttrInfo</code>	451
B.10.4 Test of <code>Dataset</code>	453
B.10.5 Test of <code>Distance</code>	454
B.10.6 Test of <code>nnMap</code>	456
B.10.7 Test of <code>Matrices</code>	458
B.10.8 Test of <code>Schema</code>	459
C Software	461
C.1 An Introduction to Makefiles	461
C.1.1 Rules	461
C.1.2 Variables	462
C.2 Installing Boost	463
C.2.1 Boost for Windows	463
C.2.2 Boost for Cygwin or Linux	464
C.3 Installing Cygwin	465
C.4 Installing GMP	465
C.5 Installing MPICH2 and Boost MPI	466
Bibliography	469
Author Index	487
Subject Index	493