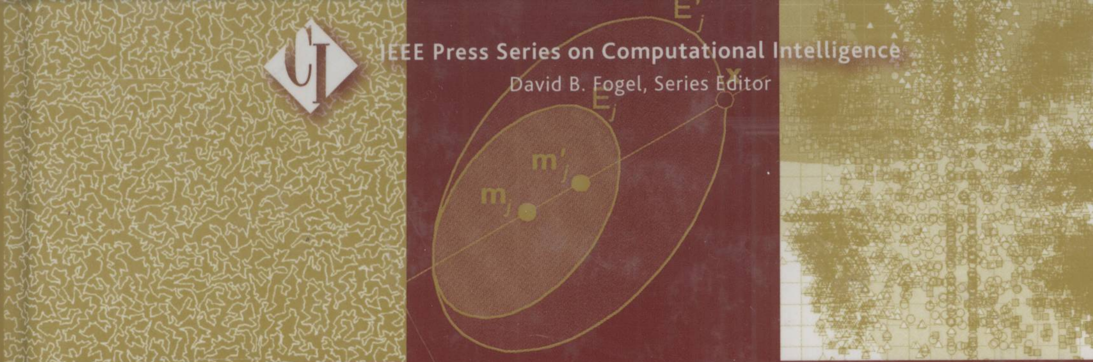
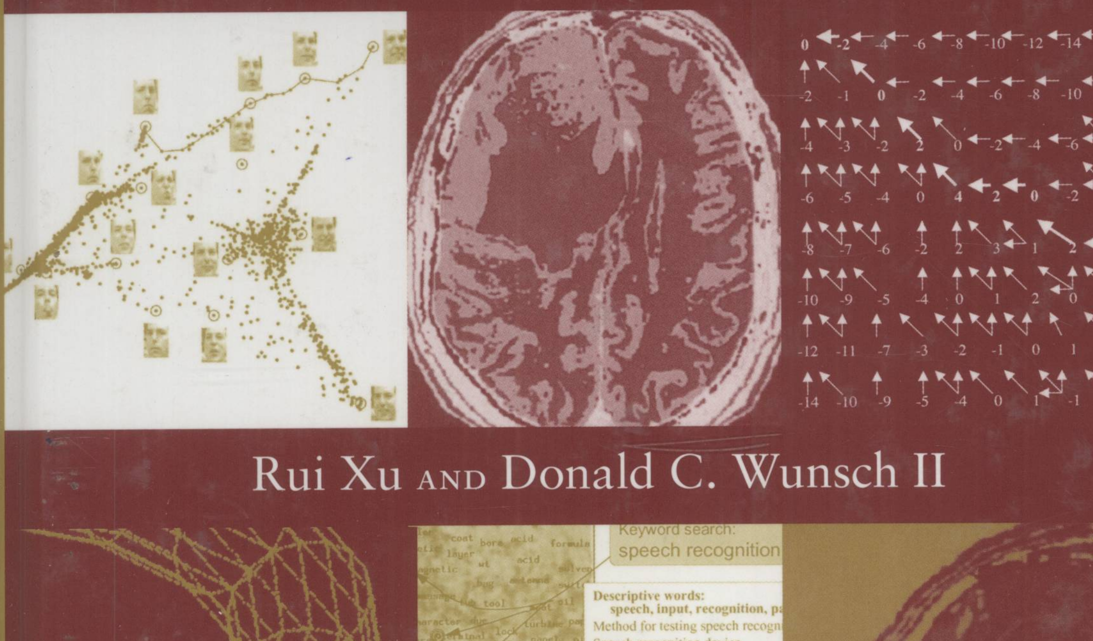
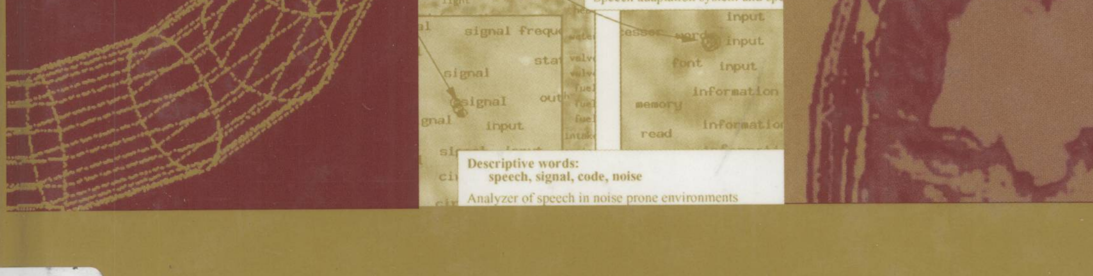
IEEE Press Series on Computational Intelligence
David B. Fogel, Series Editor



CLUSTERING



Rui Xu AND Donald C. Wunsch II



WILEY

IEEE
IEEE PRESS

TP182
X8

CLUSTERING

RUI XU

DONALD C. WUNSCH, II

华南理工大学
图书馆藏书

IEEE Computational Intelligence Society, *Sponsor*

IEEE Press Series on Computational Intelligence
David B. Fogel, *Series Editor*

IEEE

IEEE PRESS

WILEY

E2009002482

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2009 by Institute of Electrical and Electronics Engineers. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: White the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-470-27680-8

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CLUSTERING

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Lajos Hanzo, *Editor in Chief*

R. Abari
J. Anderson
S. Basu
A. Chatterjee
T. Chen
T. G. Croda
S. Farshchi
B. M. Hammerli
O. Malik
S. Nahavandi
M. S. Newman
W. Reeve

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*
Steve Welch, *IEEE Press Manager*
Jeanne Audino, *Project Editor*

IEEE Computational Intelligence Society, *Sponsor*
IEEE-CIS Liaison to IEEE Press, Gary B. Fogel

Technical Reviewers
Alexander Gorban, University of Leicester, UK
Rod Taber, Northrop Grumman Information Technology

Books in the IEEE Press Series on Computational Intelligence

Introduction to Evolvable Hardware: A Practical Guide for Designing Self-Adaptive Systems
Garrison W. Greenwood and Andrew M. Tyrrell
2007 978-0471-71977-9

Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, Third Edition
David B. Fogel
2006 978-0471-66951-7

Emergent Information Technologies and Enabling Policies for Counter-Terrorism
Edited by Robert L. Popp and John Yen
2006 978-0471-77615-4

Computationally Intelligent Hybrid Systems
Edited by Seppo J. Ovaska
2005 0-471-47668-4

Handbook of Learning and Appropriate Dynamic Programming
Edited by Jennie Si, Andrew G. Barto, Warren B. Powell, Donald Wunsch II
2004 0-471-66054-X

Computational Intelligence: The Experts Speak
Edited by David B. Fogel and Charles J. Robinson
2003 0-471-27454-2

Computational Intelligence in Bioinformatics
Edited by Gary B. Fogel, David W. Corne, Yi Pan
2008 978-0470-10526-9

Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches
Richard Jensen and Qiang Shen
2008 978-0470-22975-0

此为试读，需要完整PDF请访问：www.ertongbook.com

PREFACE

Clustering has become an increasingly important topic in recent years, caused by the glut of data from a wide variety of disciplines. However, due to the lack of good communication among these communities, similar theories or algorithms are redeveloped many times, causing unnecessary waste of time and resources. Furthermore, different terminologies confuse practitioners, especially those new to cluster analysis. Clear and comprehensive information in this field is needed. This need, among others, has encouraged us to produce this book, seeking to provide a comprehensive and systematic description of the important clustering algorithms rooted in statistics, computer science, computational intelligence, and machine learning, with an emphasis on the new advances in recent years. The book consists of 11 chapters, ranging from the basic concept of cluster analysis, proximity measures, and cluster validation, to a wide variety of clustering algorithms, including hierarchical clustering, partitional clustering, neural network-based clustering, kernel-based clustering, sequential data clustering, large-scale data clustering, and high dimensional data clustering. It also includes rich references and illustrates examples in recent applications, such as bioinformatics and web document clustering. Exercises are provided at the end of the chapters to help readers understand the corresponding topics.

The book is intended as a professional reference and also as a course textbook for graduate students in math, science, or engineering. We expect it to be particularly interesting to computer scientists and applied mathematicians applying it to data-intensive applications like bioinformatics, data mining, sensor networks, and computer security, among many other fields. It is a natural fit for computational intelligence researchers, who often must use clustering for feature selection or data reduction. The book will not have extensive assumptions of prerequisite background but will provide enough detail to allow the reader to select the method that best fits his or her application.

We have been working on cluster analysis for many years. Support from the National Science Foundation, Sandia Laboratories, and the M.K. Finley Missouri endowment is gratefully acknowledged.

We are grateful to the thousands of researchers who have contributed to this field, many of whom are our current and past collaborators, mentors, role models, and friends. It is not possible to reference all of the countless publications in this area, but we are always interested in finding outstanding ones we may have overlooked, perhaps to cover in a future edition. We thank the anonymous Associate Editor of our 2005 paper in *IEEE Transactions on Neural Networks** for the part on classification and clustering. We wish to thank the reviewers for their helpful comments. We are grateful to Bart Kosko for encouraging us to write this book after the success of the journal article. The manuscript of the book has been used in a course at Missouri University of Science and Technology. Many thanks to the graduate students Soumya De, Tae-hyung Kim, Ryan Meuth, Paul Robinette, John Seiffertt IV, and Hanzheng Wang for their valuable feedback and help in solving the homework problems. We also wish to thank Ms. Barbie Kuntemeier for her proofreading assistance.

Finally, Rui Xu would like to thank his family: Xiaomin, Benlin, Shuifeng, Wei, Jie, and Qiong; and Don Wunsch would like to thank Hong and Donnie. Without their encouragement, understanding, and patience, this book would not exist.

* Reference (Xu and Wunsch, 2005), which remained on the IEEE Explore top 100 list for over a year.

CONTENTS

CHAPTER 1

CLUSTER ANALYSIS

1.1. CLASSIFICATION AND CLUSTERING

We are living in a world full of data. Every day, people deal with different types of data coming from all types of measurements and observations. Data describe the characteristics of a living species, depict the properties of a natural phenomenon, summarize the results of a scientific experiment, and record the dynamics of a running machinery system. More importantly, data provide a basis for further analysis, reasoning, decisions, and ultimately, for the understanding of all kinds of objects and phenomena. One of the most important of the myriad of data analysis activities is to classify or group data into a set of categories or clusters. Data objects that are classified in the same group should display similar properties based on some criteria. Actually, as one of the most primitive activities of human beings (Anderberg, 1973; Everitt et al., 2001), classification plays an important and indispensable role in the long history of human development. In order to learn a new object or understand a new phenomenon, people always try to identify descriptive features and further compare these features with those of known objects or phenomena, based on their similarity or dissimilarity, generalized as proximity, according to some certain standards or rules. As an example, all natural objects are basically classified into three groups: animal, plant, and mineral. According to the biological taxonomy, all animals are further classified into categories of kingdom, phylum, class, order, family, genus, and species, from general to specific. Thus, we have animals named tigers, lions, wolves, dogs, horses,

Clustering, by Rui Xu and Donald C. Wunsch, II
Copyright © 2009 Institute of Electrical and Electronics Engineers

此为试读，需要完整PDF请访问：www.ertongbook.com

sheep, cats, mice, and so on. Actually, naming and classifying are essentially synonymous, according to Everitt et al. (2001). With such classification information at hand, we can infer the properties of a specific object based on the category to which it belongs. For instance, when we see a seal lying easily on the ground, we know immediately that it is a good swimmer without really seeing it swim.

Basically, classification systems are either supervised or unsupervised, depending on whether they assign new data objects to one of a finite number of discrete supervised classes or unsupervised categories, respectively (Bishop, 1995; Cherkassky and Mulier, 1998; Duda et al., 2001). In supervised classification, the mapping from a set of input data vectors, denoted as $\mathbf{x} \in \Re^d$, where d is the input space dimensionality, to a finite set of discrete class labels, represented as $y \in 1, \ldots, C$, where C is the total number of class types, is modeled in terms of some mathematical function $y = y(\mathbf{x},\mathbf{w})$, where $\mathbf{w}$ is a vector of adjustable parameters. The values of these parameters are determined (optimized) by an inductive learning algorithm (also termed inducer), whose aim is to minimize an empirical risk functional (related to an inductive principle) on a finite data set of input-output examples, $(\mathbf{x}_i,y_i)$, $i = 1, \ldots, N$, where N is the finite cardinality of the available representative data set (Bishop, 1995; Cherkassky and Mulier, 1998; Kohavi, 1995). When the inducer reaches convergence or terminates, an induced classifier is generated (Kohavi, 1995).

In unsupervised classification, also called clustering or exploratory data analysis, no labeled data are available (Everitt et al., 2001; Jain and Dubes, 1988). The goal of clustering is to separate a finite, unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than to provide an accurate characterization of unobserved samples generated from the same probability distribution (Baraldi and Alpaydin, 2002; Cherkassky and Mulier, 1998). This can make the task of clustering fall outside of the framework of unsupervised predictive learning problems, such as vector quantization (Cherkassky and Mulier, 1998) (see Chapter 4), probability density function estimation (Bishop, 1995) (see Chapter 4), and entropy maximization (Fritzke, 1997). It is noteworthy that clustering differs from multidimensional scaling (perceptual maps), whose goal is to depict all the evaluated objects in a way that minimizes topographical distortion while using as few dimensions as possible. Also note that, in practice, many (predictive) vector quantizers are also used for (non-predictive) clustering analysis (Cherkassky and Mulier, 1998).

It is clear from the above discussion that a direct reason for unsupervised clustering comes from the requirement of exploring the unknown natures of the data that are integrated with little or no prior information. Consider, for example, disease diagnosis and treatment in clinics. For a particular type of disease, there may exist several unknown subtypes that exhibit similar morphological appearances while responding differently to the same therapy. In this context, cluster analysis with gene expression data that measure the activities of genes provides a promising method to uncover the subtypes and thereby determine the corresponding therapies. Sometimes, the process of labeling data samples may become extremely expensive and time consuming, which also makes clustering a good choice considering the great savings in both cost and time. In addition, cluster analysis provides a compressed representation of the data and is useful in large-scale data analysis. Aldenderfer and Blashfield (1984) summarized the goals of cluster analysis in the following four major aspects:

- Development of a classification;
- Investigation of useful conceptual schemes for grouping entities;
- Hypothesis generation through data exploration;
- Hypothesis testing or the attempt to determine if types defined through other procedures are in fact present in a data set.

Nonpredictive clustering is a subjective process in nature that precludes an absolute judgment as to the relative efficacy of all clustering techniques (Baraldi and Alpaydin, 2002; Jain et al., 1999). As pointed out by Backer and Jain (1981), "in cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups." Moreover, a different clustering criterion or clustering algorithm, even for the same algorithm but with different selection of parameters, may cause completely different clustering results. For instance, human beings may be classified based on their ethnicity, region, age, socioeconomic status, education, career, hobby, weight and height, favorite food, dressing style, and so on. Apparently, different clustering criteria may assign a specific individual to very different groups and therefore produce different partitions. However, there is absolutely no way to determine which criterion is the best in general. As a matter of fact, each criterion has its own appropriate use corresponding to particular occasions, although some of them may be applied to wider situations than others. Figure 1.1 illustrates another example of the effect of subjectivity on the resulting clusters. A coarse partition divides the regions into four major clusters, while a finer one suggests that the data consist of nine clusters. Whether we adopt a coarse or fine scheme depends on the requirement of the specific problem, and in this sense, we would not say which clustering results are better, in general.

1.2. DEFINITION OF CLUSTERS

Clustering algorithms partition data objects (patterns, entities, instances, observances, units) into a certain number of clusters (groups, subsets, or

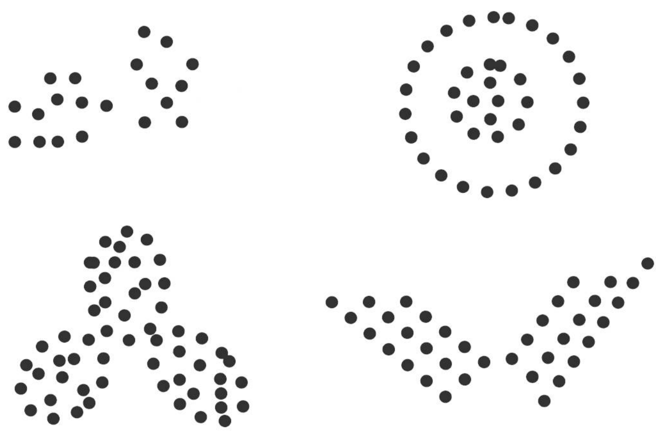


Fig. 1.1. Illustration of subjectivity of cluster analysis. Clustering at a coarse level produces four major clusters, while a finer clustering leads to nine clusters.

categories). However, there is no universally agreed upon and precise definition of the term cluster. Everitt et al. (2001) indicate that "formal definition (of cluster) is not only difficult but may even be misplaced." In spite of this difficulty, several operational definitions are still available, as summarized by Everitt (1980) and illustrated as follows:

"A cluster is a set of entities which are alike, and entities from different clusters are not alike."

A cluster is "an aggregate of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it."

"Clusters may be described as continuous regions of this space (d-dimensional feature space) containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points."

Clearly, a cluster in these definitions is described in terms of internal homogeneity and external separation (Gordon, 1999; Hansen and Jaumard, 1997; Jain and R. Dubes, 1988), i.e., data objects in the same cluster should be similar to each other, while data objects in different clusters should be dissimilar from one another. Both the similarity and the dissimilarity should be elucidated in a clear and meaningful way. Here, we give some simple mathematical descriptions of two types of clustering, known as partitional and hierarchical clustering, based on the discussion in Hansen and Jaumard (1997).

Given a set of input patterns $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \in \Re^d$, with each measure x_{ji} called a feature (attribute, dimension, or variable):

1. Hard partitional clustering attempts to seek a K-partition of $\mathbf{X}$, $C = \{C_1, \ldots, C_K\}$ $(K \leq N)$, such that

- $C_i \neq \phi, i = 1, \cdots, K;$ (1.1)
- $\bigcup_{i=1}^{K} C_i = \mathbf{X};$ (1.2)
- $C_i \cap C_j = \phi, i, j = 1, \cdots, K$ and $i \neq j$. (1.3)

2. Hierarchical clustering attempts to construct a tree-like, nested structure partition of $\mathbf{X}$, $H = \{H_1, \ldots, H_Q\}$ $(Q \leq N)$, such that $C_i \in H_m$, $C_j \in H_l$, and $m > l$ imply $C_i \subset C_j$ or $C_i \cap C_j = \phi$ for all $i, j \neq i, m, l = 1, \ldots, Q$.

For hard partitional clustering, each data object is exclusively associated with a single cluster. It may also be possible that an object is allowed to belong to all K clusters with a degree of membership, $u_{i,j} \in [0,1]$, which represents the membership coefficient of the j^{th} object in the i^{th} cluster and satisfies the following two constraints:

$$\sum_{i=1}^{K} u_{i,j} = 1, \forall j, \tag{1.4}$$

and

$$\sum_{j=1}^{N} u_{i,j} < N, \forall i, \tag{1.5}$$

as introduced in fuzzy set theory (Zadeh, 1965). This is known as fuzzy clustering and will be discussed in Chapter 4.

Figure 1.2 depicts the procedure of cluster analysis with the following four basic steps:

1. *Feature selection or extraction*. As pointed out by Jain et al. (1999, 2000) and Bishop (1995), feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones. Clearly, feature extraction is potentially capable of producing features that could be of better use in uncovering the data structure. However, feature extraction may generate features that are not physically interpretable, while feature selection assures the retention of the original physical meaning of the selected features. In the literature, these two terms sometimes are used interchangeably without further identifying the difference. Both feature selection and feature extraction are very important to the effectiveness of clustering applications. Elegant selection or generation of salient features can greatly decrease the storage requirement and measurement cost, simplify the subsequent design process, and facilitate the understanding of the data. Generally, ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, and easy to obtain and interpret. We elaborate

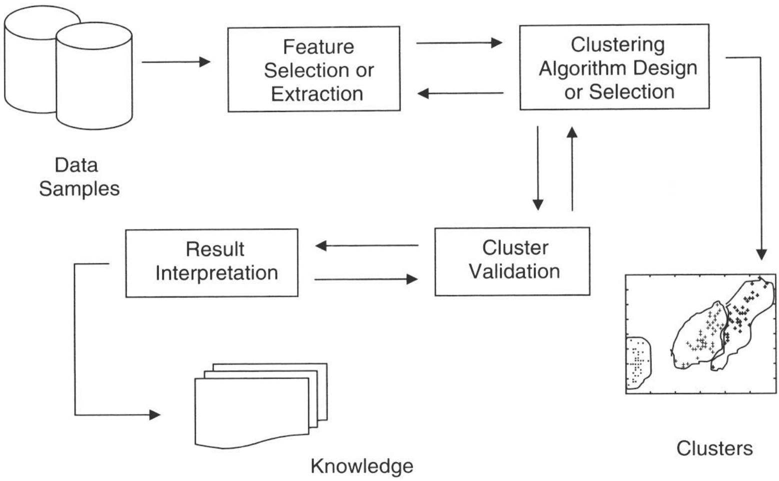


Fig. 1.2. Clustering procedure. The basic process of cluster analysis consists of four steps with a feedback pathway. These steps are closely related to each other and determine the derived clusters.

on the discussion of feature extraction in Chapter 9 in the context of data visualization and dimensionality reduction. Feature selection is more often used in the context of supervised classification with class labels available (Jain et al., 2000; Sklansky and Siedlecki, 1993). Jain et al. (2000), Liu and Yu (2005), and Theodoridis and Koutroumbas (2006) all provided good reviews of the feature selection techniques for supervised learning. A method of simultaneous feature selection and clustering, under the framework of finite mixture models, was proposed in Law et al. (2004). Kim et al. (2000) employed the genetic algorithm for feature selection in a K-means algorithm. Mitra et al. (2002) introduced a maximum information compression index to measure feature similarity and examine feature redundancy. More discussions on feature selection in clustering were given in Dy and Brodley (2000), Roth and Lange (2004), and Talavera (2000).

2. *Clustering algorithm design or selection.* This step usually consists of determining an appropriate proximity measure and constructing a criterion function. Intuitively, data objects are grouped into different clusters according to whether they resemble one another or not. Almost all clustering algorithms are explicitly or implicitly connected to some particular definition of proximity measure. Some algorithms even work directly on the proximity matrix, as defined in Chapter 2. Once a proximity measure is determined, clustering could be construed as an optimization problem with a specific criterion function. Again, the obtained clusters are dependent on the selection of the criterion function. The subjectivity of cluster analysis is thus inescapable.

Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems from a wide variety of fields. However, there is no universal clustering algorithm to solve all problems. "It has been very difficult to develop a unified framework for reasoning about it (clustering) at a technical level, and profoundly diverse approaches to clustering" (Kleinberg, 2002). Therefore, it is important to carefully investigate the characteristics of a problem in order to select or design an appropriate clustering strategy. Clustering algorithms that are developed to solve a particular problem in a specialized field usually make assumptions in favor of the application of interest. For example, the K-means algorithm is based on the Euclidean measure and hence tends to generate hyperspherical clusters. However, if the real clusters are in other geometric forms, K-means may no longer be effective, and we need to resort to other schemes. Similar considerations must be kept in mind for mixture-model clustering, in which data are assumed to come from some specific models that are already known in advance.

3. *Cluster validation*. Given a data set, each clustering algorithm can always produce a partition whether or not there really exists a particular structure in the data. Moreover, different clustering approaches usually lead to different clusters of data, and even for the same algorithm, the selection of a parameter or the presentation order of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are critically important to provide users with a degree of confidence for the clustering results. These assessments should be objective and have no preferences to any algorithm. Also, they should be able to provide meaningful insights in answering questions like how many clusters are hidden in the data, whether the clusters obtained are meaningful from a practical point of view or just artifacts of the algorithms, or why we choose one algorithm instead of another. Generally, there are three categories of testing criteria: external indices, internal indices, and relative indices. They are defined on three types of clustering structures, known as partitional clustering, hierarchical clustering, and individual clusters (Gordon, 1998; Halkidi et al., 2002; Jain and Dubes, 1988). Tests for situations in which no clustering structure exists in the data are also considered (Gordon, 1998) but seldom used because users are usually confident of the presence of clusters in the data of interest. External indices are based on some prespecified structure, which is the reflection of prior information on the data and is used as a standard to validate the clustering solutions. Internal tests are not dependent on external information (prior knowledge). Instead, they examine the clustering structure directly from the original data. Relative criteria emphasize the comparison of different clustering structures in order to provide a reference to decide which one may best reveal the characteristics of the objects. Cluster validation will be discussed in Chapter 10, with a focus on the methods for estimating the number of clusters.

此为试读，需要完整PDF请访问：www.ertongbook.com