CLUSTER ANALYSIS AND DATA MINING

LICENSE, DISCLAIMER OF LIABILITY, AND LIMITED WARRANTY

By purchasing or using this book (the "Work"), you agree that this license grants permission to use the contents contained herein, but does not give you the right of ownership to any of the textual content in the book or ownership to any of the information or products contained in it. *This license does not permit uploading of the Work onto the Internet or on a network (of any kind) without the written consent of the Publisher.* Duplication or dissemination of any text, code, simulations, images, etc. contained herein is limited to and subject to licensing terms for the respective products, and permission must be obtained from the Publisher or the owner of the content, etc., in order to reproduce or network any portion of the textual material (in any media) that is contained in the Work.

Mercury Learning and Information ("MLI" or "the Publisher") and anyone involved in the creation, writing, or production of the companion disc, accompanying algorithms, code, or computer programs ("the software"), and any accompanying Web site or software of the Work, cannot and do not warrant the performance or results that might be obtained by using the contents of the Work. The author, developers, and the Publisher have used their best efforts to insure the accuracy and functionality of the textual material and/or programs contained in this package; we, however, make no warranty of any kind, express or implied, regarding the performance of these contents or programs. The Work is sold "as is" without warranty (except for defective materials used in manufacturing the book or due to faulty workmanship).

The author, developers, and the publisher of any accompanying content, and anyone involved in the composition, production, and manufacturing of this work will not be liable for damages of any kind arising out of the use of (or the inability to use) the algorithms, source code, computer programs, or textual material contained in this publication. This includes, but is not limited to, loss of revenue or profit, or other incidental, physical, or consequential damages arising out of the use of this Work.

The sole remedy in the event of a claim of any kind is expressly limited to replacement of the book, and only at the discretion of the Publisher. The use of "implied warranty" and certain "exclusions" vary from state to state, and might not apply to the purchaser of this product.

CLUSTER ANALYSIS AND DATA MINING

An Introduction

R.S. King

MERCURY LEARNING AND INFORMATION
Dulles, Virginia
Boston, Massachusetts
New Delhi

Publisher: David Pallai
Mercury Learning and Information
22841 Quicksilver Drive
Dulles, VA 20166
info@merclearning.com
www.merclearning.com
1-800-758-3756

This book is printed on acid-free paper.

R.S. King. *Cluster Analysis and Data Mining: An Introduction*
ISBN: 978-1-938549-38-0

Library of Congress Control Number: 2014941165

141516321 Printed in the United States of America

Our titles are available for adoption, license, or bulk purchase by institutions, corporations, etc. Digital versions of this title are available at www.authorcloudware.com. Companion disc files may be obtained by contacting info@merclearning.com. For additional information, please contact the Customer Service Dept. at 1-800-758-3756 (toll free).

To the memory of my father, who made it possible and to LaJuan, the shining light in my life who survived the process.

CONTENTS

PREFACE

This book is appropriate for a first course in clustering methods and data mining. Clustering and data mining methods are applicable in many fields of study, for example:

1. in the life sciences for developing complete taxonomies,
2. in the medical sciences for discovering more effective and economical means for making positive diagnosis in the treatment of patients,
3. in the behavioral and social sciences for discerning human judgments and behavior patterns,
4. in the earth sciences for identifying and classifying geographical regions,
5. in the engineering sciences for pattern recognition and artificial intelligence applications, and
6. in decision and information sciences for analysis of markets and documents.

The first five chapters consider early historical clustering methods. Chapters 1 and 2 are an introduction to general concepts in clustering methods, with an emphasis on proximity measures and data mining. Classical numerical clustering methods are presented in Chapters 3 and 4: hierarchical and partitioned clustering. These methods are particularly defined

only on numeric data files. A clustering method implemented via multiple linear regression, judgmental analysis (JAN), is discussed in Chapter 5. JAN allows for numerical and categorical variables to be included in a clustering study.

All of the methods in Chapters 1 through 5 generate partitions on a study's data file, referred to as *crisp clustering* results. *Fuzzy clustering* methods presented in Chapter 6, capture partitions plus modified versions for the partitions. The modified partitions allow for overlapping clusters.

Chapter 7 is an introduction to the data mining topics of classification and association rules, which enable qualitative rather than simply quantitative data mining studies to be conducted.

Cluster analysis is essentially an art, but can be accomplished scientifically if the results of a clustering study can be validated. This is discussed in Chapter 8. Determination of the validity of individual clusters and the validation of a clustering, or collection of clusters, are discussed.

Chapter 9 surveys a variety of algorithms for clustering categorical data: ROCK, STIRR, CACTUS, and CLICK. These methods are dependent on underlying data structures and are applicable to relational databases.

Applications of clustering methods are presented in Chapters 10 through 11. Chapter ten discusses classical statistical methods for identifying outliers. Additionally, crisp and fuzzy clustering methods are applied to the outlier identification problem. Chapter 11 is an overview of model-based clustering. This is often used in physical science research studies for data generation.

A summary of the issues and trends in the cluster analysis field is made in Chapter 12. Besides giving recommendations for further study, an introduction to neural networks is presented. The appendices provide a variety of resources (software, URLs, algorithms, references) for the cluster analysis plus URLs for test data files.

The text is applicable to either a course on clustering, data mining, and classification or as a companion text for a first class in applied statistics. Clustering and data mining are good motivators and applications of the topics commonly included in an introductory applied statistics course.

The scheduling references for each of the chapters, in an applied statistics class, could be as follows:

Chapters 1-4: after study of descriptive statistics.

Chapter 9: immediately following Chapters 1-3.

Chapter 6: after study of descriptive statistics.

Chapter 10: after studying the Empirical Rule and Chebychev's Law.

Chapter 7: after studying probability.

Chapter 8: after study of hypothesis testing.

Chapter 5: after study of correlation, and both linear and multiple linear regression.

Chapter 11: after study of statistical inference.

No previous experience or background in clustering is assumed. Elementary statistics plus a brief exposure to data structures are the prerequisites. Informal algorithms for clustering data and interpreting results are emphasized. In order to evaluate the results of clustering and to explore data, graphical methods and data structures are used for representing data. Throughout the text, examples and references are provided, in order to enable the material to be comprehensible for a diverse audience.

CHAPTER 1

INTRODUCTION TO CLUSTER ANALYSIS

In This Chapter

1.1 WHAT IS A CLUSTER?

Many of the decisions being made today involve more than one person. An important question in the group decision process is: "How does the group arrive at its final decision?" There have been a number of different mathematical and statistical approaches used by researchers attempting to model the decision-making process including game theory, information theory, and linear programming. Due to the large variety of decision-making situations, different types of decision processes, and the kinds of skills required, there is still a great deal of concern about the best way to make decisions. In many cases there is no objective approach. The individuals in the decision-making group each use their own set of criterion in reaching a decision. This approach might work in a situation where a consensus is

not needed. However, in the case where a single group decision is needed, there must be a "meeting of the minds."

One approach used is the *Delphi Technique*. This technique was designed in the early 1950s by the Rand Corporation to predict future outcomes. It is a group information gathering process to develop consensus opinion from a panel of experts on a topic of interest. In the normal Delphi scenario, the panel never meets face to face but interacts through questionnaires and feedback. This noncontact approach alleviates the worry over such issues as individual defensiveness or persuasiveness. However, opinions can be swayed due to a participant observing the responses of the rest of the panel. Another problem with the Delphi Technique is that the noncontact aspect is not feasible when, for example, the panel is the graduate admissions committee at a university.

Cluster analysis is another technique that has been used with success in the decision-making process. First, the investigator must determine the answer to "What is a cluster?" The **premise in cluster analysis** is: given a number of individuals, each of which is described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that the objects within classes are *similar* in some respect and *unlike* those from other classes. These deduced classes are the clusters. The number of classes and the characteristics of each class must be determined from the data as discussed by Everett.[1]

The key difference between cluster analysis and the Delphi Technique is that cluster analysis is strictly an objective technique. Whereas individual decisions can be swayed in an attempt to reach consensus in the Delphi process, or a "happy medium" is reached which does not really portray the feelings of the group as a whole. This is not the case in cluster analysis. Clusters of individuals are reached using an objective mathematical function. One particular type of cluster analysis called Judgmental ANalysis (JAN) takes the process one step further. Not only does it classify the panel into similar groups based on a related regression equation, but it also allows for these equations to be combined into a single policy equation. The JAN technique has been in use since the 1960s. It has proven to be an effective first step for methods of capturing and clustering the policies of judges.

Attempts at classification, that is sorting similar things into categories, can be traced back to primitive humans. The ability to classify is a necessary prerequisite for the development of language. Nouns, for example, are labels

[1] Everitt, B. S. (1980). *Cluster analysis (2nd ed.)*. New York: Halsted Press.

used to classify a particular group of objects. Saying that a particular four-legged animal is a "dog" allows us to put that animal into a category separate from cats, sheep, and horses. In other words, it allows us to communicate.

The classification of people and animals is almost as old as language. The early Hindus categorized humans into six types based on sex, physical, and behavioral characteristics. The early Greeks and Romans used classification to get a better understanding of the world around them. Galen, A.D. 129-199, defined nine temperamental types that were assumed to be related to a person's susceptibility to various diseases and to individual differences in behavior as discussed by Everitt.[1] Development of a method to categorize animals into species was initiated by Aristotle. He started by dividing them into red blooded (vertebrates) and those not having red blood (invertebrates). He then subdivided the two groups again based on how their young were born. Theophrastus continued Aristotle's work, providing the groundwork for biological research for centuries. Eventually, new taxonomic systems were developed by such people as Linnaeus, Lindley, and Darwin. Classification was not restricted to the biological sciences. In chemistry, Mendeleyev used classification to develop the periodic tables, discussion by Everitt.[1]

In the 1960s, two events led to an explosion of interest in cluster analysis. The availability and spread of large, high-speed computers opened up new possibilities for researchers. Additionally, the publication of *Principles of Numerical Taxonomy* by Sokal and Sneath[2] covered the following three important areas:

1. a number of different cluster analysis techniques
2. the use of computers in classification research
3. a radically empirical approach to biological taxonomy presented by Blashfield and Aldenderfer[3]

The need for cluster analysis arises in many fields of study. For example, Anderberg[4] lists six areas where cluster analysis has been used successfully:

1. In the life sciences (biology, botany, zoology, etc.), the objects of analysis are life forms such as plants, animals, and insects. The

---

[2] Sokal, R. R., and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman.
[3] Blashfield and Aldenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research*, 13, 271-295.
[4] Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

operational purpose of the analysis may range from developing complete taxonomies to delimiting the subspecies of a distinct but varied species.

2. In the medical sciences (psychiatry, pathology, etc.), the objects of a cluster analysis may be diseases, patients, symptoms, and laboratory tests. The operational emphasis here is on discovering more effective and economical means for making positive diagnosis in the treatment of patients.

3. In the behavioral and social sciences (psychology, sociology, education, etc.), some of the wide variety of objects of analysis are training methods, behavior patterns, organizations, human judgments, families, and teaching techniques.

4. Applications of cluster analysis in the earth sciences (geology, geography, etc.) have included the study of land and rock formations, soils, river systems, cities, and regions of the world.

5. Examples of entities that have been clustered in the engineering sciences (pattern recognition, artificial intelligence, cybernetics, electrical engineering, etc.) include handwritten characters, speech, fingerprints, electrocardiograms, radar signals, and circuit designs.

6. In the area of information and decision sciences (information retrieval, political science, economics, marketing research, operational research, etc.), cluster analysis has been applied to the analysis of documents, markets, investments, and credit risks.

As can be seen, the areas in which cluster analysis has been used with success are large and varied. It is also interesting to note some of the other names for cluster analysis in these different fields. Some of the aliases mentioned by Anderberg[3] are numerical taxonomy (biology, botany, ecology), typology (social sciences), learning without a leader (pattern recognition, cybernetics, electrical engineering), clumping (information retrieval, linguistics), regionalization (geography), partition (graph theory, circuit designers), and serration (anthropology). The reasons for clustering are as many and as varied as the fields and names. Everitt[1] mentions seven possible uses of clustering techniques including data reduction, data exploration, hypothesis generating, hypothesis testing, model fitting, and prediction based on groups, and finding a true typology.

1.2 CAPTURING THE CLUSTERS

Cluster analysis employs a measure of similarity or dissimilarity for assigning points in space to a cluster. In general terms, points exist in a space (which could be a plane, the surface of a sphere, three-dimensional space, etc.) that relate to the concept of distance that matches geometrical intuition. Formally, the operational definition of this type of distance is: a **proximity measure** in space $M = \{A, d\}$ consists of a nonempty set A together with a distance function d: $A \times A \rightarrow \mathbf{R}^2$ which satisfies:

1. $d(x, y) \geq 0$; $d(x, y) = 0$ if and only if $x = y$

That is, the distance between two distinct points is strictly positive.

2. $d(x, y) = d(y, x)$ for all x, y in A

The distance from x to y is equal to the distance from y to x.

3. (a) for a dissimilarity $d(i,i) = 0$, for all i

The distance between a point and itself is zero, or

points aren't different from themselves.

(b) for a similarity $d(i,i) \geq max_k\, d(i, k)$ for all i

The points are most similar to themselves.

So, what does this actually mean? First, there must exist a nonempty set A, basically a collection of one or more points. Given a distance function, d, which can be used to determine the distance between any two points of A, d, must also follow certain rules.

The first rule states that one cannot have a negative distance and the distance between two points can only be zero if the two points are, in fact, in exactly the same place. The second rule states that the distance between two points must be the same for whichever direction is measured, going from x to y covers the same distance as going from y to x. Finally, measurement is either based upon similarity or dissimilarity between points.

Let $M = \{A, d\}$ be the space being studied, let a be in A, and let $\varepsilon > 0$. The **ε-neighborhood** of a in M is defined to be:

$$\mathbf{N}_\varepsilon(a) = \{x \text{ in } A \mid d(x, a) < \varepsilon\}$$

That is, the collection of points x in A within distance ε of a.

It is worth emphasizing that $\mathbf{N}_\varepsilon(a)$ does *not* include the boundary. It consists *only* of the interior of the "neighborhood." If the boundary is included, the neighborhood is called a *closed* neighborhood.

In $\mathbf{R}^2$ (the plane), $\mathbf{N}_\varepsilon(a)$ is the interior of a disc of radius ε centered on a.

In $\mathbf{R}^3$ (three-dimensional space), $\mathbf{N}_\varepsilon(a)$ is the interior of a solid ball of radius ε centered on a.

The previous two examples have all used the Euclidean metric, that is, our intuitive notion of distance:

$d_1 = d(i,k) = \left(\sum_{j=1}^{d} \left|x_{ij} - x_{kj}\right|^2\right)^{1\backslash 2}$, in two-dimensional space and d is the number of features.

In $\mathbf{R}^2$, using the **Minkowski measure**,

$d_2 = d(i,k) = \left(\sum_{j=1}^{d} \left|x_{ij} - x_{kj}\right|^r\right)^{1\backslash r}$, where d is the number of features, n is the number of patterns, and $r = 2$ is the dimension of the space, $\mathrm{N}_\varepsilon(a)$ is the interior of a square centered on a, with sides of length 2ε parallel to the co-ordinate axes.

The **"sup" distance,**

$d_3 = d(i,k) = max_{i \le j \le d} \left|x_{ij} - x_{kj}\right|$, where d is the number of patterns, generates diamond shaped ε-neighborhoods.

Clusters are captured by attempting to find nonoverlapping ε-neighborhoods using the given proximity. The goal is to group objects in a group (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the finer granularity is present in the clustering.

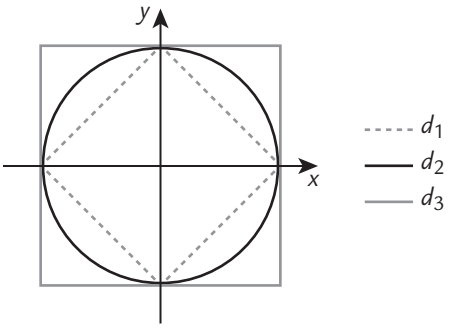

FIGURE 1.1 Example ε-Neighborhoods.

Other commonly used proximity measures include:

City-block (Manhattan) distance. This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. The city-block distance is: distance $(x,y) = \sum_i \left|y_i - x_i\right|$

Chebychev distance. This distance measure may be appropriate in cases when defining two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as: distance (x,y) = Maximum $|x_i - y_i|$

Power distance. Sometimes the emphasis is to increase or decrease the progressive weight that is placed on dimensions for different objects. This can be accomplished via the *power distance*. The power distance is computed as: distance $(x,y) = \sum_i \left(|x_i - y_i|^p\right)^{1/r}$, where r and p are user-defined parameters. Parameter p controls the progressive weight that is placed on differences on individual dimensions, parameter r controls the progressive weight that is placed on larger differences between objects. If r and p are equal to two, then this distance is equal to the Euclidean distance.

Percent disagreement. This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature. This distance is computed as: distance (x,y) = (Number of $x_i \neq y_i$)/ i

Consider the following set of points in two-dimensional Euclidean space:

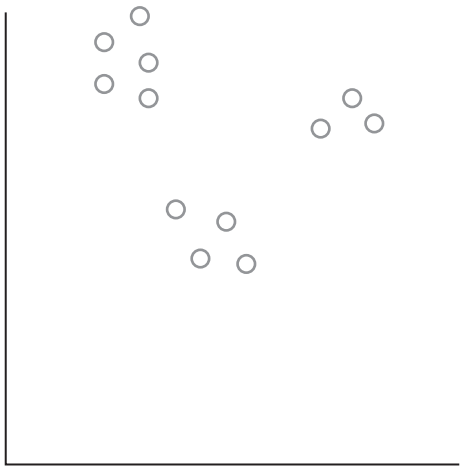

FIGURE 1.2 Points in Two-dimensional Euclidian Space.

Three groups would be identified with Euclidian neighborhoods.

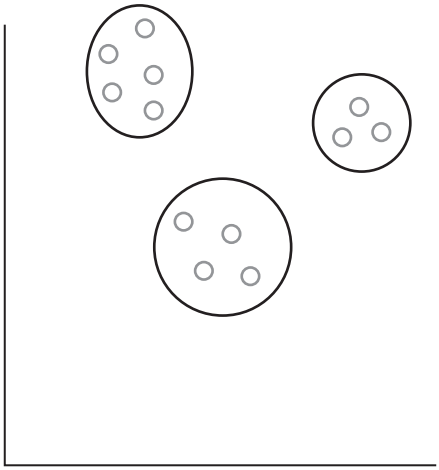

FIGURE 1.3 Three Groups Captured by Euclidean Neighborhoods.

Several primary questions need to be investigated when capturing the clusters. These questions include:

1. "How many clusters are present?" Consider the following situation:

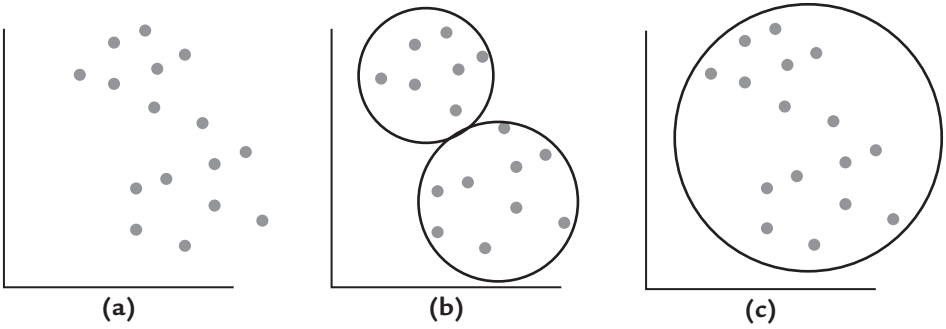

FIGURE 1.4 (a) Original Points, (b) Two Groups, and (c) One Group.

2. "Does the current ε-neighborhood and proximity measure correctly identify the clusters?"

For instance, only one cluster can be captured for the following set of points in two-dimensional Euclidean space:

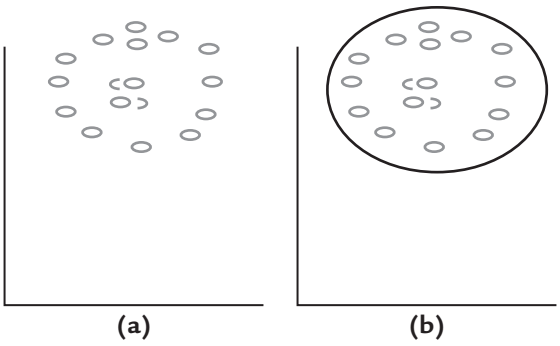

FIGURE 1.5 (a) Original Points and (b) One Cluster Interpretation.

In reality, the actual number of clusters should be two. In this case, the Euclidean neighborhoods are incapable of obtaining the correct number of clusters.

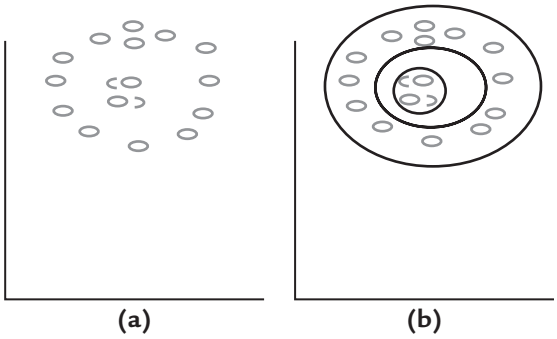

FIGURE 1.6 (a) Original Points and (b) Actual Groups.

This example illustrates that cluster analysis is sensitive to both the proximity measure selected and related ε-neighborhood shapes. Different approaches may yield different results. Consequently, the distance metric should be chosen carefully. The results should also be compared to analyses based on different proximity measures to enable determination of the robustness of the results.

3. Do the captured clusters have realistic interpretations?

4. Do any of the clusters overlap? If so, to what degree?

1.3 NEED FOR VISUALIZING DATA

The discussion on proximity measures for capturing clusters demonstrates that clustering software needs features that make clustering practical for a wide variety of applications. Such a package should at least provide highly optimized implementations of agglomerative, k-means,

and graph clustering, especially in the context of sparse high-dimensional data. Additionally, the package should help the user sort through the algorithm options and resulting data files by providing an intuitive graphical interface. Clustering software should provide both standard statistics and unique visualizations for interpreting clustering results. Given the wide range of options and factors that are involved in clustering, the user should carefully analyze his results and compare them with results generated with different options. Visualizations enhance the analysis and comparisons.

1.4 THE PROXIMITY MATRIX

According to *Oxford Dictionary of Statistics*,a square matrix in which the entry in cell (j, k) is some measure of the similarity (or distance) between the items to which row j and column k correspond. A simple example would be a standard mileage chart—the smaller the entry, the closer together are the two items. Proximity matrices form the data for multidimensional scaling. Asymmetric matrices can occur (for example, if the measurement is time taken, then the journey from top to bottom of a hill will be shorter than the journey from bottom to top).

Suppose we are given the following **ordinal proximity matrix**:

	x_1	x_2	x_3	x_4	x_5
x_1	0	6	8	2	7
x_2	6	0	1	5	3
x_3	8	1	0	10	8
x_4	2	5	10	0	4
x_5	7	3	9	4	0

TABLE 1.1 An Ordinal Proximity Matrix.

The objects x_i, for the cell the value represents the Euclidean distance between the objects. Next we construct a **proximity ratio matrix**; a matrix where a proximity measure has been derived from the proximity matrix. The Euclidean proximity measure would generate the following proximity ratio matrix for the matrix in Table 1.2:

	x_2	x_3	x_4	x_5
x_1	12.08	14.77	4.69	10.58
x_2		7.48	12.17	9.17
x_3			16.37	13.64
x_4				7.87

TABLE 1.2 A Proximity Ratio Matrix.

Using the proximity ratio matrix we can perform a hierarchical clustering $\gamma_0, \gamma_2, \ldots, \gamma_{n-1}$ where the mth clustering contains $n - m$ clusters. A level function, records the proximity for each clustering formed. For the start of this process, $L(k) = k$, because the levels are evenly spread apart.

$$L(m) = \min\{d(x_i, x_j) \text{ I } \gamma_m \text{ is defined}\}$$

where the mth clustering contains $n - m$ clusters:

$$\gamma_{Cm} = \left\{\gamma_{m1}, \gamma_{m2}, \ldots, \gamma_{m(n-m)}\right\}$$

The **cophenetic proximity measure** d_c on the n objects is the level at which objects x_i and x_j are first in the cluster.

$$d_c(i,j) = L(k_{ij})$$

where

For single-link, use $k_{ij} = min\left\{m: \left(x_i, x_j\right) \varepsilon \gamma_{Cmq} \text{ for some } q\right\}$

For complete-link, use $k_{ij} = max\left\{m: \left(x_i, x_j\right) \varepsilon \gamma_{Cmq} \text{ for some } q\right\}$

This process generates the following results for the complete-link solution:

	x_2	x_3	x_4	x_5
x_1	12.08	14.77	4.69	10.58
x_2		7.48	12.17	9.17
x_3			16.37	13.64
x_4				7.87

	x_2	$x_{3,4}$	x_5
x_1	12.08	14.77	10.58
x_2		12.17	9.17
$x_{3,4}$			13.64

	x_2	x_5
$x_{1,3,4}$	12.17	13.64
x_2		9.17

	x_2
$x_{1,3,4,5}$	12.17

FIGURE 1.7 Complete-Link Clustering Process.

1.5 DENDROGRAMS

A special type of tree structure, called a **dendrogram**, provides a graphical presentation for the clustering. A dendrogram consists of layers of nodes, where each node represents a cluster. Lines connect nodes representing clusters which are nested together. Horizontal slices of a dendrogram indicate a clustering. For the latter complete-link clustering, we have:

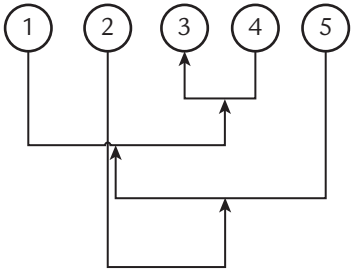

FIGURE 1.8 Dendrogram for a Complete-Link Clustering.

At the start $L(1) = 16.37$ for the clustering $\{(x_1), (x_2), (x_3), (x_4), (x_5)\}$.

On the first iteration, $L(2) = 14.77$ for the clustering $\{(x_1), (x_2), (x_3, x_4), (x_5)\}$.

After completion of the second iteration, $L(3) = 13.64$ for the clustering $\{(x_2), (x_1, x_3, x_4), (x_5)\}$.

Completion of the third iteration generates: $L(4) = 12.17$ for the clustering $\{(x_2), (x_1, x_3, x_4, x_5)\}$.

The last iteration generates: $L(5) = 12.17$ for the clustering $\{(x_1, x_2, x_3, x_4, x_5)\}$.

Clearly several obvious questions arise at this point in the cluster analysis:

- How many groups are present in the data?
- What is the group membership interpretation?
- Will different grouping algorithms have a common clustering result? Not necessarily!
- CLUSTERING IS AN ART AS WELL AS A SCIENCE!

1.6 SUMMARY

Cluster analysis is the formal study of algorithms and methods for grouping, or classifying objects.

Questions to resolve include:

What defines similarity between objects or between clusters?

What defines a distance between two clusters?

How can you capture clusters? What are the different methods for identifying these clusters?

When is it "best" to partition or identify clusters?

When is it "best" to stop joining clusters?

What are the right data elements to utilize in clustering for a problem in a specific application domain where we may have hundreds of variables?

What are the limitations of cluster analysis?

Basically the steps completed in clustering include:

Step One: Form similarities between all pairs of the objects based on a given attribute.

Step Two: Groups are constructed where within-group similarities are larger than the between-group similarities.

Let $d(i,k)$ be a proximity between the ith and kth patterns. Then:

- For a dissimilarity $d(i,i) = 0$, for all i.
- For a similarity, $d(i,i) = \max d(i,k)$, for all (i,k).
- $d(i,k) = d(k,i)$, for all (i,k).
- $d(i,k) \geq 0$, for all (i,k).

The common proximity measures are:

- Minkowski metric: $d(i,k) = \left(\sum_{j=1}^{d} \left| x_{ij} - x_{kj} \right| \right)$
- Euclidean distance: $d(i,k) = \left(\sum_{j=1}^{d} \left(x_{ij} - x_{kj} \right)^2 \right)^{1/2}$
- Manhattan distance: $d(i,k) = \sum_{j=1}^{d} \left| x_{ij} - x_{kj} \right|$
- Sup distance: $d(i,k) = \max_{1 \leq j \leq d} \left| x_{ij} - x_{kj} \right|$

1.7 EXERCISES

1. Suppose you are given two decks of playing cards, one with a blue backing and the other with a red backing. Discuss ways in which the 104 cards without jokers or 108 cards with jokers can be clustered. Is it possible to form a clustering which is not a partition?
2. What is the distinction between the following terms: similarity measure, dissimilarity measure, metric, and distance measure?
3. Complete the computations for the single-link clustering presented in the chapter.

4. Complete a single-link cluster on the ordinal proximity matrix given in Table 1.2.

5. Perform a complete-link cluster for each proximity measure given the following example considering the following data {(16,19), (20,23), (8,20), (1,23), (18,6), (5,28)} with associated labels {1, 2, 3, 4, 5, 6}, which are points within two-dimensional Euclidean space:

   (a) Minkowski metric (b) Manhattan distance

   (c) Sup distance (d) Percent disagreement

6. The cluster analysis presented in this chapter is a bottom-to-top (agglomerative) process. Perform a top-to-bottom (divisive) process. In other words, start with one cluster containing all the objects and finish with a clustering containing all the singleton clusters.

7. For problems 3 through 5, discuss how to determine when the clustering stops.

8. How could clustering methods be used for identifying outlier(s)? Note that outlier(s) by itself (themselves) will be a cluster. Think of an example of a tree diagram which will point out few outliers; and how the grouping pattern and the stem will be represented by a cluster of outliers.

9. What is the relationship between the linkage distance measure and the number of clusters?

10. Why would the number of clusters not be a simple continuously increasing number? Is it possible that there may not be a one-to-one relationship between the linkage distance and the number of clusters?

11. How does variable selection play a role in cluster analysis; what method is best to use?

12. Why is linkage distance inversely related to the number of clusters in general?

13. What happens if a similarity measure is used instead of distance measure?

14. What is meant by similarity or distance measures when we have qualitative data?

15. What is the major problem with the nonhierarchical method? (Hint: start point of the seed or center of the cluster)

16. Why should you standardize the data when doing cluster analysis?

17. Discuss how to use the dendrogram. (Tree structure for identifying the distance "between clusters" and which observations belong to which cluster—a graphical representation issue.)

18. Various factors affect the stability of a clustering solution, including: selection, distance/similarity measure used, different significance levels, and type of method (divisive vs. agglomerative) among others. Do some background research and present a report on a method that converges to the right solution in the midst of the above mentioned parameters.

19. You are given the following contingency table for binary data:

Object *i* \ Object *j*	1	0	*sum*
1	*a*	*b*	*a* + *b*
0	*c*	*d*	*c* + *d*
sum	*a* + *c*	*b* + *d*	*p*

TABLE 1.3 Contingency Table.

(a) Define a symmetry measure based upon the contingency table.

(b) Define an asymmetry measure based upon the contingency table.

20. Describe the type of variables in the following table and define a dissimilarity measure for the binary variables.

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

21. Define a similarity measure for nominal variables.

22. What is an ordinal variable? Discuss how a dissimilarity measure can be defined for an ordinal variable.

23. What is a ratio-scaled variable? Discuss how nonlinear scale, such as an exponential scale, should be represented in a cluster analysis.

24. Databases consist of variables of mixed types. Discuss how to operationally define similarity/dissimilarity measures for a typical database.

25. What is a vector object? How can one operationally define a dissimilarity measure for vector objects?

CHAPTER 2

OVERVIEW OF DATA MINING

In This Chapter

2.1 WHAT IS DATA MINING?

A primary goal for many twenty-first century companies is to simultaneously maximize their rate of return and customer satisfaction. Supply chain management coupled with the associated entity relationship program enable firms to be competitive in the workforce. These firms are able to deliver a high-quality product that is highly useful to the customer in a timely fashion. Success is based upon understanding their customers, vendors, and supply chain.

Often this type of understanding is partially obtained by drilling-into-the-database. For example, consider a database for a chain of grocery stores. The top ten customers could be found simply by using filters. Pivot table