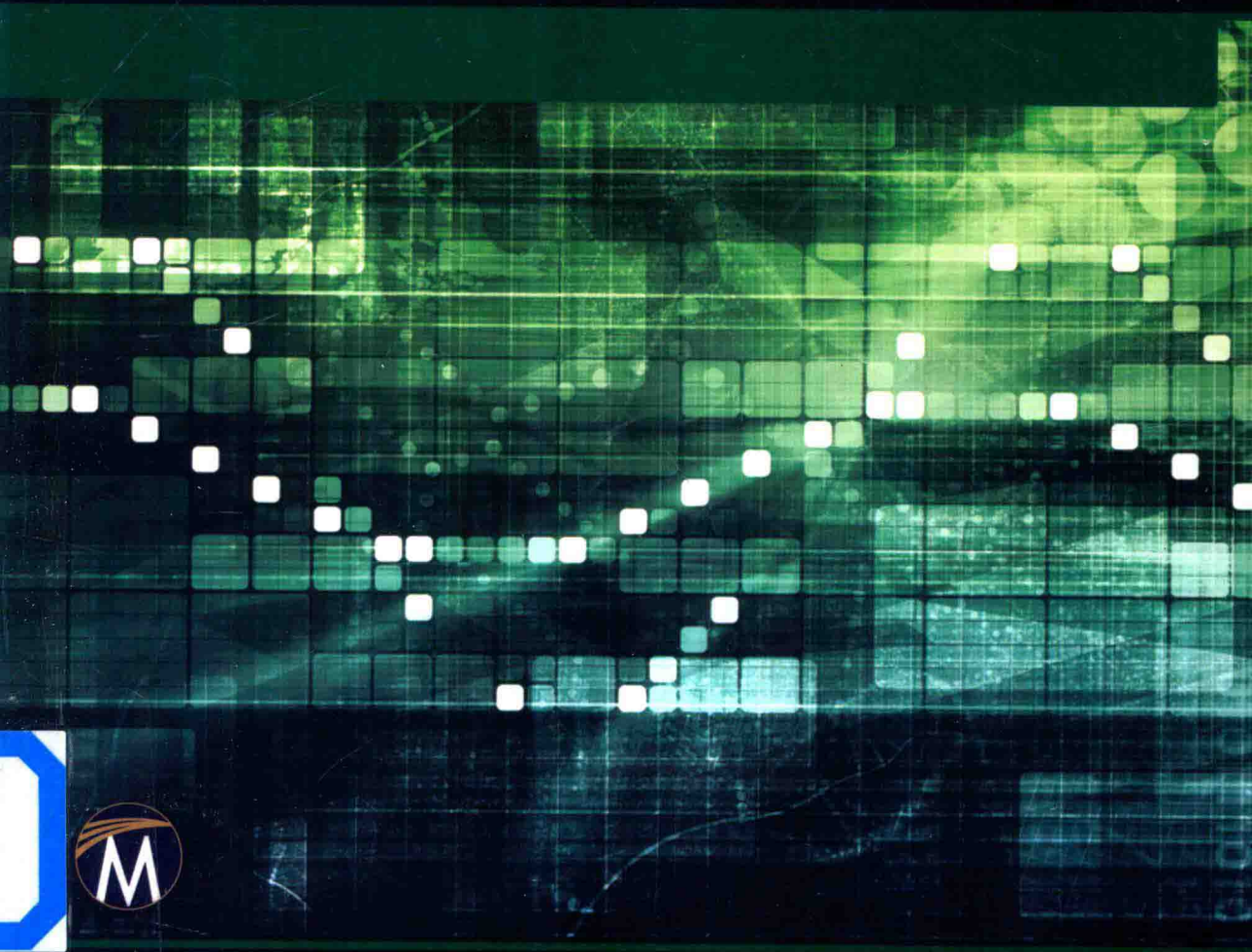
Cluster Analysis and Data Mining

An Introduction



Ronald S. King

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).

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.

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

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.

CONTENTS

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.