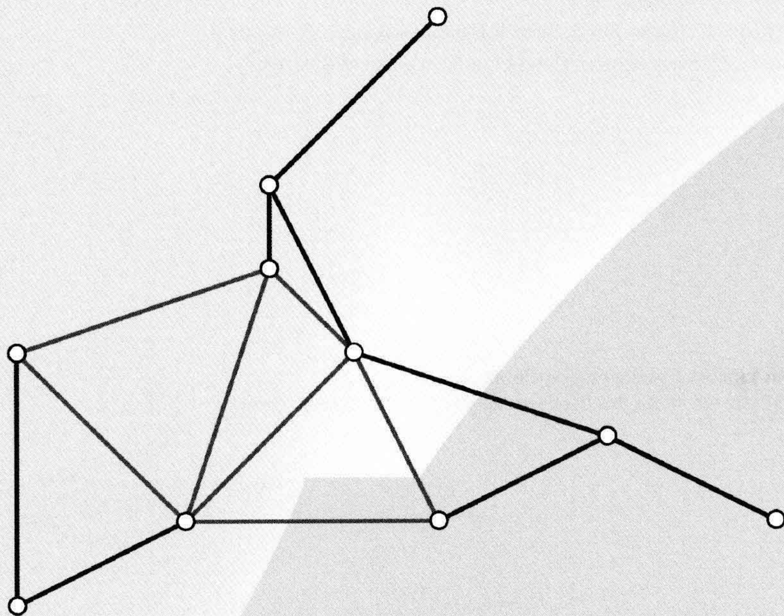
Interdisciplinary Mathematical Sciences – Vol. 10

Ordinal and Relational Clustering

Melvin F Janowitz

Interdisciplinary Mathematical Sciences – Vol. 10



Ordinal and Relational Clustering

Melvin F Janowitz
Rutgers University, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Interdisciplinary Mathematical Sciences — Vol. 10
ORDINAL AND RELATIONAL CLUSTERING
(With CD-ROM)

ISBN-13 978-981-4287-20-3
ISBN-10 981-4287-20-2

Printed in Singapore by Mainland Press Pte Ltd.

Ordinal and Relational Clustering

INTERDISCIPLINARY MATHEMATICAL SCIENCES

Published

Vol. 1: Global Attractors of Nonautonomous Dissipative Dynamical Systems
David N. Cheban

Vol. 2: Stochastic Differential Equations: Theory and Applications
A Volume in Honor of Professor Boris L. Rozovskii
eds. Peter H. Baxendale & Sergey V. Lototsky

Vol. 3: Amplitude Equations for Stochastic Partial Differential Equations
Dirk Blömker

Vol. 4: Mathematical Theory of Adaptive Control
Vladimir G. Sragovich

Vol. 5: The Hilbert–Huang Transform and Its Applications
Norden E. Huang & Samuel S. P. Shen

Vol. 6: Meshfree Approximation Methods with MATLAB
Gregory E. Fasshauer

Vol. 7: Variational Methods for Strongly Indefinite Problems
Yanheng Ding

Vol. 8: Recent Development in Stochastic Dynamics and Stochastic Analysis
eds. Jinqiao Duan, Shunlong Luo & Caishi Wang

Vol. 9: Perspectives in Mathematical Sciences
eds. Yisong Yang, Xinchu Fu & Jinqiao Duan

Vol. 10: Ordinal and Relational Clustering (with CD-ROM)
Melvin F. Janowitz

For Trudy

Who makes all things possible!

Preface

Underlying motivation for the book: *If data has only ordinal significance, then taking averages is not a meaningful basis for comparison.* This is not a new observation. There is a deep theory of measurement, and a substantial literature involving this and related issues. The interested reader might consult sources like Refs. [46, 55, 56].

This book is rather unusual in its organization as well as in its purpose. It is neither a text book nor is it a well documented exhaustive research volume. A middle ground is sought. Its subject is a type of exploratory data analysis that has come to be known as *cluster analysis*. Despite claims to the contrary in the literature, arguments will be presented to show that the input data to a clustering algorithm might have nothing more than some version of ordinal significance. For that reason, a careful approach will be presented for the analysis of ordinal data. Though mathematicians often try to develop general properties of whatever system they are studying, it seems clear that there is no axiom system that will provide a "best" method of discovering any internal structure of an arbitrary data set. For that reason, we will present properties that cluster algorithms may or may not enjoy, rather than axioms that should necessarily hold for every algorithm and all data.

Where appropriate, we shall illustrate abstract ideas with concrete numerical examples, as well as with careful mathematical arguments. Much of our treatment is based on ideas originally appearing in Ref. [44]. For readers wishing a more comprehensive introduction to cluster analysis, possible references include Refs. [28, 30, 32, 51, 63]. Though the treatment is mathematical in nature, it will be written so as to make it accessible to researchers interested in concrete data analysis. The first five chapters deal with standard cluster analysis, while the remaining two chapters present a new and more general theory that subsumes both cluster analysis and also other related techniques of data analysis. In particular, it includes some aspects of formal concept analysis (Refs. [18, 26]), and makes contact with symbolic data analysis [13].

A *clustering problem* may take as its input a finite data set P having at least three elements, together with a list of finitely many attributes that the data might have in varying degrees. Alternately, it may consist of a numerical (perhaps ordinal) measure of similarity or dissimilarity between pairs of objects of P. Though the ultimate output is to discover some sort of internal classification structure that P may have, we argue that an appropriate intermediate step is to form a dissimilarity coefficient that somehow summarizes this internal structure. The key idea is that the sought after output structure must be determined entirely by the input, and is in no way directly dependent on any external criteria. We do not argue that external criteria should be ignored, but only take the view that when such considerations are made, we are no longer in the realm of cluster analysis. We shall largely ignore the passage from attribute data to similarity or dissimilarity measures, and concentrate instead on the transformation of a dissimilarity measure into some sort of classification or nested sequence of classifications by means of first creating a "summary" measure of dissimilarity from which the ultimate output may be constructed. Thus, we usually take our input for a cluster algorithm to be a dissimilarity coefficient. The reader should note that the terms "dissimilarity coefficient" and "dissimilarity measure" are being used interchangeably. The abbreviation DC provides a convenient way to refer to them.

Research in the social and behavioral sciences is replete with examples where perceptions or preferences are first quantified, and then on the basis of this quantification, a classification or a decision must be reached. For example, judges will give numerical ratings to paintings or to athletic performances. Though these ratings are subjective with criteria that vary from judge to judge, one hopes that if the rating for A is less than the rating for B, then in some interpretable sense, B is superior to A. In other words, even though the numbers themselves may not have meaning, the fact that the rating for A is less than the rating for B may still have significance. Much the same situation applies in general to cluster analysis. We will discuss algorithms whose input is a numerical measure of dissimilarity. If one knew the nature of the data and just how the measure d of dissimilarity was constructed, the actual values of $d(x, y)$ might be interpretable. But we cannot assume such knowledge. For a measure d of unspecified origin, the most one can hope for, and all that is claimed, is the fact that $d(x, y) < d(s, t)$ should imply that x is more similar to y, than s is to t. We will briefly consider the behavior of cluster algorithms when faced with noisy data. Here versions of continuity are a natural consideration, as are the nature of various properties of the input data implied by the choice of cluster algorithm. Naturally, there will be other considerations.

Here is an intuitive road map that will guide your journey through the text. We begin by introducing the underlying ideas by means of examples and informal discussions. This is followed by a more mathematical treatment of the same material, again illustrated by examples. The general theme we follow is that of looking at the action of various mappings on the reals with a numerical dissimilarity measure d. Thus we carefully study the effects of the transformation of the dissimilarity coefficient d into the dissimilarity coefficient θd, where θ is some mapping on the nonnegative reals. This is the content of the third chapter, while Chapter 4 considers ideas related to continuity. It turns out to be possible to classify cluster methods according to how they interact with various mappings on the reals. The idea is that this classification tells us something about the data assumptions that are implicitly made by the choice of cluster method. Such is the content of the fifth chapter.

Chapter 6 begins with an introduction to formal concept analysis, thereby laying a foundation for making a connection between cluster analysis and this seemingly unrelated discipline. The reader ia shown situations where dissimilarities are measured in settings more general than the real number system. Dissimilarities taking values in a partially ordered set L will be of special interest. It is argued here that it is appropriate to actually view the dissimilarity between x and y as an *order filter* of the poset L in which dissimilarities are measured. This leads to a dramatically different view of a dissimilarity coefficient d. For objects x, y, the idea is that $d(x, y)$ should denote the set of levels of the poset L at which $\{x, y\}$ is a *candidate* for clustering. A cluster method F decides at each level which cluster candidates should actually be merged to form clusters. The output of F is then a dissimilarity coefficient whose image takes values in the order filters of L. Associated with this output is a collection of partitions of subsets of the underlying set, thus making a natural connection with the type of display used in formal concept analysis. This is a display of some or all of the possible clusters that arise. The goal of this type of clustering is to provide possible hypotheses for the internal structure of P. Thus rather than assume that there is a unique "true" hierarchical structure that we are trying to estimate, we make no statistical assumptions, and instead take the view that we are seeking possible hidden internal structures of the given data set. Note that this approach allows us to consider a DC taking values in the attribute space associated with the input data. For that reason, we will call this new general discipline "relational clustering". Naturally, we do not preclude the possibility of an underlying "true" structure that we are estimating, but our goal is to present a model that is also suitable for use with data mining applications. The actual connection with FCA is made by considering dissimilarities taking values in a finite Boolean algebra.

The final chapter presents a more formal and more general view of this material, and illustrates the idea by presenting the results of some clustering algorithm based on these ideas. In particular, this allows one to consider attributes whose values vary within the confines of the individual objects being clustered, This enables contacts with percentile clustering, confidence interval clustering, and symbolic data analysis. It is hoped that this introduction to such a general model of cluster analysis will inspire further development of this new subject.

We close with a word about organization and terminology. The text is organized into chapters and sections. Numbering will be by divisions, so that a reference of the form Theorem 2.3.5 would indicate Theorem 5 of Section 3 of Chapter 2. Though we try to make the treatment as self-contained as possible, we do assume a basic knowledge of set theory, functions, and the real number system. Sections prefaced with a $\star$ should be viewed as optional and may safely be omitted. These sections tend to be technical and require a certain amount of extra sophistication on the part of the reader. They may also contain ideas that are not often of interest to the typical user of clustering techniques. The terminology and notation are (at least to the author) standard. There is included a list of tables, a list of figures, a list of symbols, as well as a detailed index for the text. You will find a CD containing MATLAB software accompanying the book. This software is designed to allow the reader to personally check the illustrative examples in the book; it will be discussed in Sections 3.4.3, 6.5, 7.4, 7.5 and 7.6. Any needed corrections or improvements will be placed in the website

`http://www.worldscibooks.com/mathematics/7449.html`

Acknowledgments

Before proceeding, the author expresses his gratitude to CLUSTAN Ltd. for granting permission to use the ClustanGraphics clustering package in connection with processing some of the illustrative examples. Illustrative software is provided that makes use of version 7.9 of MATLAB® and the Optimization Toolbox. Product information may be obtained from

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098 USA
Tel: 508-647-7000, Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

We especially thank Ed Scheinerman for making available the Matgraph Toolbox `http://www.ams.jhu.edu/~ers/matgraph` that was used as a basis for many of our MATLAB® calculations.

Professor Thomas Riedel, Mathematics Department, University of Louisville taught a graduate course based on an early version of the text during the Summer of 2008. Both he and his class made valuable suggestions that improved the quality of the exposition. We especially thank his students for their input. Among them are Richard Lagani, Amanda Ledford, and Christy Parks. We also thank Professors Robert C. Powers (University of Louisville), Richard Greechie (Louisiana Polytechnic University), and Li Chen (University of the District of Columbia) for their comments and suggestions. Finally, we thank Ralph Freese (University of Hawaii) for modifying and making available for use with the book his software package *LatDraw* for drawing lattices, (`http://www.latdraw.org`). We also thank the Computer Science Laboratory at the University of the District of Columbia for a version of the software that was used to produce Figures 7.1, 7.2, and 7.3 of the text. A big thanks goes to Fred McMorrris (Illinois Institute of Technology) both for his editorial role with World Scientific, and his encouragement over many years as a valued research colleague. We also owe a debt of gratitude to the editor of the series in which this volume appears: Professor Jinqiao (Jeffrey) Duan (Illinois Institute of Technology), as well as to the editorial staff of World Scientific: notably to Rajesh Babu and Rok Ting Tan.

Contents

List of Figures

List of Tables