Mathematical Classification and Clustering

Nonconvex Optimization and Its Applications

Volume 11

The titles published in this series are listed at the end of this volume.

Mathematical Classification and Clustering

by

Boris Mirkin
DIMACS, Rutgers University

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress

ISBN-13: 978-1-4613-8057-3 e-ISBN-13: 978-1-4613-0457-9
DOI: 10.1007/978-1-4613-0457-9

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates
the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Table of Contents

Foreword

I am very happy to have this opportunity to present the work of Boris Mirkin, a distinguished Russian scholar in the areas of data analysis and decision making methodologies.

The monograph is devoted entirely to clustering, a discipline dispersed through many theoretical and application areas, from mathematical statistics and combinatorial optimization to biology, sociology and organizational structures. It compiles an immense amount of research done to date, including many original Russian developments never presented to the international community before (for instance, cluster-by-cluster versions of the K-Means method in Chapter 4 or uniform partitioning in Chapter 5). The author's approach, approximation clustering, allows him both to systematize a great part of the discipline and to develop many innovative methods in the framework of optimization problems. The optimization methods considered are proved to be meaningful in the contexts of data analysis and clustering.

The material presented in this book is quite interesting and stimulating in paradigms, clustering and optimization. On the other hand, it has a substantial application appeal. The book will be useful both to specialists and students in the fields of data analysis and clustering as well as in biology, psychology, economics, marketing research, artificial intelligence, and other scientific disciplines.

Panos Pardalos, Series Editor.

Preface

The world is organized via classification: elements in physics, compounds in chemistry, species in biology, enterprises in industries, illnesses in medicine, standards in technology, firms in economics, countries in geography, parties in politics — all these are witnesses to that. The science of classification, which deals with the problems of how classifications emerge, function and interact, is still unborn. What we have in hand currently is clustering, the discipline aimed at revealing classifications in observed real-world data. Though we can trace the existence of clustering activities back a hundred years, the real outburst of the discipline occurred in the sixties, with the computer era coming to handle the real-world data.

Within just a few years, a number of books appeared describing the great opportunities opened in many areas of human activity by algorithms for finding "coherent" clusters in a data "cloud" put in a geometrical space (see, for example, Benzécri 1973, Bock 1974, Clifford and Stephenson 1975, Duda and Hart 1973, Duran and Odell 1974, Everitt 1974, Hartigan 1975, Sneath and Sokal 1973, Sonquist, Baker, and Morgan 1973, Van Ryzin 1977, Zagoruyko 1972).

The strict computer eye was supposed to substitute for imprecise human vision and transform the art of classification into a scientific exercise (for instance, numerical taxonomy was to replace handmade and controversial taxonomy in biology). The good news in that was that the algorithms did find clusters. The bad news was that there was no rigorous theoretical foundation underlying the algorithms. Moreover, for a typical case in which no clear cluster structure prevailed in the data, different algorithms produced different clusters. More bad news was the lack of any rigorous tool for interpreting the clusters found, which yielded eventually to the emergence of the so-called conceptual clustering as a counterpart to the traditional one.

The pessimism generated by these obstacles can be felt in popular sayings like these: "There are more clustering techniques suggested than the number of real-world problems resolved with them", and "Clustering algorithms are worth a dime a dozen." However, the situation is improving, in the long run. More and more

real-world problems, such as early diagnostics in medicine, knowledge discovery and message understanding in artificial intelligence, machine vision and robot planning in engineering, require developing a sound theory for clustering.

In the last two decades, beyond the traditional activity of inventing new clustering concepts and algorithms, we can distinguish two overlapping mainstreams potentially leading to bridging the gaps within the clustering discipline. One is related to modeling cluster structures in terms of observed data, and the other is connected with analyzing particular kinds of phenomena, such as image processing or biomolecular-data-based phylogeny reconstructing – even though in the latter kind of analyses, clustering is only a part, however important, of the entire problem.

Within the former movements, initially, the effort was concentrated on developing probabilistic models in a statistical framework (see, for example, monographs by Breiman et al. 1984, Jain and Dubes 1988, McLachlan and Basford 1988), leaning more to testing rather than to revealing the cluster structures. However, all along, work was being done on modeling of clusters in the data just as it is, without any connection to a possible probabilistic mechanism of data generation. In this paradigm, probabilistic clusters are just a particular clustering structure, and the clustering discipline seems more related to mathematics and artificial intelligence than to statistics. The present book offers an account of clustering in the framework of this wider paradigm.

Actually, the book's goal is threefold. First, it is supposed to be a reference book for the enormous amount of existing clustering concepts and methods; second, it can be utilized as a clustering text-book; and, third, it is a presentation of the author's and his Russian colleagues' results, put in the perspective of the current development.

As a reference book, it features:

(a) a review of classification as a scientific notion;

(b) an updated review of clustering algorithms based on a systematic typology of input-data/output-cluster-structures (the set of cluster structures considered is quite extensive and includes such structures as neural networks);

(c) a detailed description of the approaches in single cluster clustering, partitioning, and hierarchical clustering, including most recent developments made in various countries (Canada, France, Germany, Russia, USA);

(d) development of a unifying approximation approach;

(e) an extensive bibliography, and

(f) an index.

To serve in the text-book capacity, the monograph includes:

(a) a dozen illustrative and small, though real-world, data sets, along with clustering problems quite similar to those for larger real data sets;

(b) detailed description and discussion of the major algorithms and underlying theories;

(c) solutions to the illustrative problems found with the algorithms described (which can be utilized as a stock of exercises).

It should be pointed out that the data sets, mostly, are taken from published sources and have been discussed in the literature extensively, which provides the reader with opportunity to look at them from various perspectives. The examples are printed with a somewhat smaller font, like this.

The present author's results are based on a different approach to cluster analysis, which can be referred to as *approximation clustering*, developed by him and his collaborators starting in the early seventies. Some similar work is being done in the USA and in the other countries. In this approach, clustering is considered to approximate data by a simpler, cluster-wise structure rather than to reveal the geometrically explicit "coherent clusters" in a data point-set. The results found within the approximation approach amount to a mathematical theory for clustering involving the following directions of development: (a) unifying a considerable part of the clustering techniques, (b) developing new techniques, (c) finding relations among various notions and algorithms both within the clustering discipline and outside – especially in statistics, machine learning and combinatorial optimization.

The unifying capability of approximation clustering is grounded on convenient relations which exist between approximation problems and geometrically explicit clustering. Based on this, the major clustering techniques have been reformulated as locally optimal approximation algorithms and extended to many situations untreatable with explicit approaches such as mixed-data clustering. Firm mathematical relations have been found between traditional and conceptual clustering; moreover, unexpectedly, some classical statistical concepts such as contingency measures have been found to have meaning in the approximation framework. These yield a set of simple but efficient interpretation tools. Several new methods have been developed in the framework, such as additive and principal cluster analyses, uniform partitioning, box clustering, and fuzzy additive type clustering. In a few cases, approximation clustering goes into substantive phenomena modeling, as in the case of aggregating mobility tables.

The unifying features of the approximation approach fit quite well into some general issues raised about clustering goals (defined here in the general classification context) and the kinds of data tables treated. Three data types – column-conditional, comparable and aggregable table – defined with regard to extent of

comparability among the data entries, are considered here through all the material in terms of different approximation clustering models.

Though all the mathematical notions used are defined in the book, the reader is assumed to have an introductory background in calculus, linear algebra, graph theory, combinatorial optimization, elementary set theory and logic, and statistics and multivariate statistics.

The contents of the book are as follows. In Chapter 1, the classification forms and functions are discussed, especially as involved in the sciences. Such an analysis is considered a prerequisite to properly defining the scope and goals of clustering; probably, it has never been undertaken before, which explains why the discussion takes more than two dozen pages. The basic data formats are discussed, and a set of illustrative clustering problems is presented based on small real-world data sets. In Chapter 2, the data table notions are put in a geometrical perspective. The major low-rank approximation model is considered as related to data analysis techniques such as the principal component and correspondence analyses, and its extension to arbitrary additive approximation problems is provided. In Chapter 3, a systematic review of the clustering concepts and techniques is given, sometimes accompanied by examples. Chapters 4 through 7, the core of the book, are devoted to a detailed account of the mathematical theories, including the most current ones, on clustering, with three kinds of discrete clustering structures: single cluster (Chapter 4), partition (Chapters 5 and 6), and hierarchy and its extensions (Chapter 7). There are not too many connections between the latter Chapters,

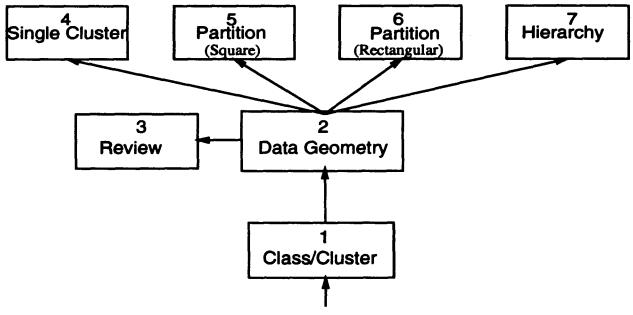

Figure 0.1: Basic dependence structure.

which allows us to present the structure of the book in the following fan-shaped format (see Fig.0.1).

The Sections are accompanied by reviewing discussions while the Chapters' main features are listed as their preambles.

The following are suggested as subjects for a college course/seminar, based on the material presented: a review of clustering (Chapters 1 through 3), clustering algorithms (any subset of algorithms presented in Chapters 3 through 6 along with the illustrative examples from these chapters and corresponding data descriptions from Chapter 1), and combinatorial clustering (Chapters 4 through 7).

Last, but not least, the author would like to acknowledge the role of some researchers and organizations in preparing of this volume: my collaborators in Russia, who participated in developing the approximation approach, especially Dr. V. Kupershtoh, Dr. V. Trofimov and Dr. P. Rostovtsev (Novosibirsk); Dr. S. Aivazian (Moscow), who made possible the development of a program, ClassMaster, implementing (and, thus, testing) many of the approximation clustering algorithms in the late eighties; Ecole Nationale Supérieure des Télécommunications (ENST, Paris), which provided a visiting position for me at 1991-1992, and Dr. L. Lebart and Dr. B. Burtschy from ENST, who helped me in understanding and extending the contingency data analysis techniques developed in France; Dr. F.S. Roberts, Director of the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS, a NSF Science and Technology Center), in the friendly atmosphere of which I did most of my research in 1993-1996; support from the Office of Naval Research (under a grant to Rutgers University) that provided me with opportunities for further developing the approach as reflected in my most recent papers and talks, the contents of which form the core of the monograph presented; discussions with Dr. I. Muchnik (Rutgers University) and Dr. T. Krauze (Hofstra University) have been most influential for my writing; the Editor of the series, Dr. P. Pardalos, has encouraged me to undertake this task; and Mr. R. Settergren, a PhD student, has helped me in language editing. I am grateful to all of them.