Contents

Preface

Clustering is a discipline devoted to finding and describing cohesive or homogeneous chunks in data, the clusters.

Some exemplary clustering problems are:

- Finding common surf patterns in the set of web users;
- Automatically revealing meaningful parts in a digitalized image;
- Partition of a set of documents in groups by similarity of their contents;
- Visual display of the environmental similarity between regions on a country map;
- Monitoring socio-economic development of a system of settlements via a small number of representative settlements;
- Finding protein sequences in a database that are homologous to a query protein sequence;
- Finding anomalous patterns of gene expression data for diagnostic purposes;
- Producing a decision rule for separating potentially bad-debt credit applicants;
- Given a set of preferred vacation places, finding out what features of the places and vacationers attract each other;
- Classifying households according to their furniture purchasing patterns and finding groups' key characteristics to optimize furniture marketing and production.

Clustering is a key area in data mining and knowledge discovery, which are activities oriented towards finding non-trivial or hidden patterns in data collected in databases.

Earlier developments of clustering techniques have been associated, primarily, with three areas of research: factor analysis in psychology [55], numerical taxonomy in biology [122], and unsupervised learning in pattern recognition [21].

Technically speaking, the idea behind clustering is rather simple: introduce a measure of similarity between entities under consideration and combine similar entities into the same clusters while keeping dissimilar entities in different clusters. However, implementing this idea is less than straightforward.

First, too many similarity measures and clustering techniques have been

invented with virtually no support to a non-specialist user in selecting among them. The trouble with this is that different similarity measures and/or clustering techniques may, and frequently do, lead to different results. Moreover, the same technique may also lead to different cluster solutions depending on the choice of parameters such as the initial setting or the number of clusters specified. On the other hand, some common data types, such as questionnaires with both quantitative and categorical features, have been left virtually without any substantiated similarity measure.

Second, use and interpretation of cluster structures may become an issue, especially when available data features are not straightforwardly related to the phenomenon under consideration. For instance, certain data on customers available at a bank, such as age and gender, typically are not very helpful in deciding whether to grant a customer a loan or not.

Specialists acknowledge peculiarities of the discipline of clustering. They understand that the clusters to be found in data may very well depend not on only the data but also on the user's goals and degree of granulation. They frequently consider clustering as art rather than science. Indeed, clustering has been dominated by learning from examples rather than theory based instructions. This is especially visible in texts written for inexperienced readers, such as [4], [28] and [115].

The general opinion among specialists is that clustering is a tool to be applied at the very beginning of investigation into the nature of a phenomenon under consideration, to view the data structure and then decide upon applying better suited methodologies. Another opinion of specialists is that methods for finding clusters as such should constitute the core of the discipline; related questions of data pre-processing, such as feature quantization and standardization, definition and computation of similarity, and post-processing, such as interpretation and association with other aspects of the phenomenon, should be left beyond the scope of the discipline because they are motivated by external considerations related to the substance of the phenomenon under investigation. I share the former opinion and argue the latter because it is at odds with the former: in the very first steps of knowledge discovery, substantive considerations are quite shaky, and it is unrealistic to expect that they alone could lead to properly solving the issues of pre- and post-processing.

Such a dissimilar opinion has led me to believe that the discovered clusters must be treated as an "ideal" representation of the data that could be used for recovering the original data back from the ideal format. This is the idea of the data recovery approach: not only use data for finding clusters but also use clusters for recovering the data. In a general situation, the data recovered from aggregate clusters cannot fit the original data exactly, which can be used for evaluation of the quality of clusters: the better the fit, the better the clusters. This perspective would also lead to the addressing of issues in pre- and post-

processing, which now becomes possible because parts of the data that are explained by clusters can be separated from those that are not.

The data recovery approach is common in more traditional data mining and statistics areas such as regression, analysis of variance and factor analysis, where it works, to a great extent, due to the Pythagorean decomposition of the data scatter into "explained" and "unexplained" parts. Why not try the same approach in clustering?

In this book, two of the most popular clustering techniques, K-Means for partitioning and Ward's method for hierarchical clustering, are presented in the framework of the data recovery approach. The selection is by no means random: these two methods are well suited because they are based on statistical thinking related to and inspired by the data recovery approach, they minimize the overall within cluster variance of data. This seems to be the reason of the popularity of these methods. However, the traditional focus of research on computational and experimental aspects rather than theoretical ones has contributed to the lack of understanding of clustering methods in general and these two in particular. For instance, no firm relation between these two methods has been established so far, in spite of the fact that they share the same square error criterion.

I have found such a relation, in the format of a Pythagorean decomposition of the data scatter into parts explained and unexplained by the found cluster structure. It follows from the decomposition, quite unexpectedly, that it is the divisive clustering format, rather than the traditional agglomerative format, that better suits the Ward clustering criterion. The decomposition has led to a number of other observations that amount to a theoretical framework for the two methods. Moreover, the framework appears to be well suited for extensions of the methods to different data types such as mixed scale data including continuous, nominal and binary features. In addition, a bunch of both conventional and original interpretation aids have been derived for both partitioning and hierarchical clustering based on contributions of features and categories to clusters and splits. One more strain of clustering techniques, one-by-one clustering which is becoming increasingly popular, naturally emerges within the framework giving rise to intelligent versions of K-Means, mitigating the need for user-defined setting of the number of clusters and their hypothetical prototypes. Most importantly, the framework leads to a set of mathematically proven properties relating classical clustering with other clustering techniques such as conceptual clustering and graph theoretic clustering as well as with other data mining concepts such as decision trees and association in contingency data tables.

These are all presented in this book, which is oriented towards a reader interested in the technical aspects of data mining, be they a theoretician or a practitioner. The book is especially well suited for those who want to learn WHAT clustering is by learning not only HOW the techniques are applied

but also WHY. In this way the reader receives knowledge which should allow him not only to apply the methods but also adapt, extend and modify them according to the reader's own ends.

This material is organized in five chapters presenting a unified theory along with computational, interpretational and practical issues of real-world data mining with clustering:
- What is clustering (Chapter 1);
- What is data (Chapter 2);
- What is K-Means (Chapter 3);
- What is Ward clustering (Chapter 4);
- What is the data recovery approach (Chapter 5).

But this is not the end of the story. Two more chapters follow. Chapter 6 presents some other clustering goals and methods such as SOM (self-organizing maps) and EM (expectation-maximization), as well as those for conceptual description of clusters. Chapter 7 takes on "big issues" of data mining: validity and reliability of clusters, missing data, options for data pre-processing and standardization, etc. When convenient, we indicate solutions to the issues following from the theory of the previous chapters. The Conclusion reviews the main points brought up by the data recovery approach to clustering and indicates potential for further developments.

This structure is intended, first, to introduce classical clustering methods and their extensions to modern tasks, according to the data recovery approach, without learning the theory (Chapters 1 through 4), then to describe the theory leading to these and related methods (Chapter 5) and, in addition, see a wider picture in which the theory is but a small part (Chapters 6 and 7).

In fact, my prime intention was to write a text on classical clustering, updated to issues of current interest in data mining such as processing mixed feature scales, incomplete clustering and conceptual interpretation. But then I realized that no such text can appear before the theory is described. When I started describing the theory, I found that there are holes in it, such as a lack of understanding of the relation between K-Means and the Ward method and in fact a lack of a theory for the Ward method at all, misconceptions in quantization of qualitative categories, and a lack of model based interpretation aids. This is how the current version has become a threefold creature oriented toward:

1. Giving an account of the data recovery approach to encompass partitioning, hierarchical and one-by-one clustering methods;

2. Presenting a coherent theory in clustering that addresses such issues as (a) relation between normalizing scales for categorical data and measuring association between categories and clustering, (b) contributions of various elements of cluster structures to data scatter and their use in interpreta-

tion, (c) relevant criteria and methods for clustering differently expressed data, etc.;

3. Providing a text in data mining for teaching and self-learning popular data mining techniques, especially K-Means partitioning and Ward agglomerative and divisive clustering, with emphases on mixed data pre-processing and interpretation aids in practical applications.

At present, there are two types of literature on clustering, one leaning towards providing general knowledge and the other giving more instruction. Books of the former type are Gordon [39] targeting readers with a degree of mathematical background and Everitt et al. [28] that does not require mathematical background. These include a great deal of methods and specific examples but leave rigorous data mining instruction beyond the prime contents. Publications of the latter type are Kaufman and Rousseeuw [62] and chapters in data mining books such as Dunham [23]. They contain selections of some techniques reported in an ad hoc manner, without any concern on relations between them, and provide detailed instruction on algorithms and their parameters.

This book combines features of both approaches. However, it does so in a rather distinct way. The book does contain a number of algorithms with detailed instructions and examples for their settings. But selection of methods is based on their fitting to the data recovery theory rather than just popularity. This leads to the covering of issues in pre- and post-processing matters that are usually left beyond instruction. The book does contain a general knowledge review, but it concerns more of issues rather than specific methods. In doing so, I had to clearly distinguish between four different perspectives: (a) statistics, (b) machine learning, (c) data mining, and (d) knowledge discovery, as those leading to different answers to the same questions. This text obviously pertains to the data mining and knowledge discovery perspectives, though the other two are also referred to, especially with regard to cluster validation.

The book assumes that the reader may have no mathematical background beyond high school: all necessary concepts are defined within the text. However, it does contain some technical stuff needed for shaping and explaining a technical theory. Thus it might be of help if the reader is acquainted with basic notions of calculus, statistics, matrix algebra, graph theory and logics.

To help the reader, the book conventionally includes a list of denotations, in the beginning, and a bibliography and index, in the end. Each individual chapter is preceded by a boxed set of goals and a dictionary of base words. Summarizing overviews are supplied to Chapters 3 through 7. Described methods are accompanied with numbered computational examples showing the working of the methods on relevant data sets from those presented in Chapter 1; there are 58 examples altogether. Computations have been carried out with

self-made programs for MATLAB®, the technical computing tool developed by The MathWorks (see its Internet web site www.mathworks.com).

The material has been used in the teaching of data clustering and visualization to MSc CS students in several colleges across Europe. Based on these experiences, different teaching options can be suggested depending on the course objectives, time resources, and students' background.

If the main objective is teaching clustering methods and there are very few hours available, then it would be advisable to first pick up the material on generic K-Means in sections 3.1.1 and 3.1.2, and then review a couple of related methods such as PAM in section 6.1.2, iK-Means in 3.3.1, Ward agglomeration in 4.1 and division in 4.2.1, single linkage in 6.2.1 and SOM in 6.1.6. Given a little more time, a review of cluster validation techniques from 7.6 including examples in 3.3.2 should follow the methods. In a more relaxed regime, issues of interpretation should be brought forward as described in 3.4, 4.2.3, 6.3 and 7.2.

If the main objective is teaching data visualization, then the starting point should be the system of categories described in 1.1.5, followed by material related to these categories: bivariate analysis in section 2.2, regression in 5.1.2, principal component analysis (SVM decomposition) in 5.1.3, K-Means and iK-Means in Chapter 3, Self-organizing maps SOM in 6.1.6 and graph-theoretic structures in 6.2.

Acknowledgments

Too many people contributed to the approach and this book to list them all. However, I would like to mention those researchers whose support was important for channeling my research efforts: Dr. E. Braverman, Dr. V. Vapnik, Prof. Y. Gavrilets, and Prof. S. Aivazian, in Russia; Prof. F. Roberts, Prof. F. McMorris, Prof. P. Arabie, Prof. T. Krauze, and Prof. D. Fisher, in the USA; Prof. E. Diday, Prof. L. Lebart and Prof. B. Burtschy, in France; Prof. H.-H. Bock, Dr. M. Vingron, and Dr. S. Suhai, in Germany. The structure and contents of this book have been influenced by comments of Dr. I. Muchnik (Rutgers University, NJ, USA), Prof. M. Levin (Higher School of Economics, Moscow, Russia), Dr. S. Nascimento (University Nova, Lisbon, Portugal), and Prof. F. Murtagh (Royal Holloway, University of London, UK).

Author

Boris Mirkin is a Professor of Computer Science at the University of London UK. He develops methods for data mining in such areas as social surveys, bioinformatics and text analysis, and teaches computational intelligence and data visualization.

Dr. Mirkin first became known for his work on combinatorial models and methods for data analysis and their application in biological and social sciences. He has published monographs such as 'Group Choice' (John Wiley & Sons, 1979) and 'Graphs and Genes' (Springer-Verlag, 1984, with S. Rodin).

Subsequently, Dr. Mirkin spent almost ten years doing research in scientific centers such as Ecole Nationale Supérieure des Télécommunications (Paris, France), Deutsches KrebsForschnung Zentrum (Heidelberg, Germany), and National Center for Discrete Mathematics and Theoretical Computer Science DIMACS, Rutgers University (Piscataway, NJ USA). Building on these experiences, he developed a unified framework for clustering as a data recovery discipline.

List of Denotations

I	Entity set
N	Number of entities
V	Feature set
V_l	Set of categories of a categorical feature l
M	Number of column features
$X = (x_{iv})$	Raw entity-to-feature data table
$Y = (y_{iv})$	Standardized entity-to-feature data table; $y_{iv} = (x_{iv} - a_v)/b_v$ where a_v and b_v denote the shift and scale coefficients, respectively
$y_i = (y_{iv})$	M-dimensional vector corresponding to entity $i \in I$ according to data table Y
$y_i = (y_{iv})$	M-dimensional vector corresponding to entity $i \in I$ according to data table Y
(x, y)	Inner product of two vector points $x = (x_j)$ and $y = (y_j)$, $(x, y) = \sum_j x_j y_j$
$d(x, y)$	Distance (Euclidean squared) between two vector points $x = (x_j)$ and $y = (y_j)$, $d(x, y) = \sum_j (x_j - y_j)^2$
$\{S_1, ..., S_K\}$	Partition of set I in K disjoint classes $S_k \subset I$, $k = 1, ..., K$
K	Number of classes/clusters in a partition $S = \{S_1, ..., S_K\}$ of set I
N_k	Number of entities in class S_k of partition S $(k = 1, ..., K)$
$c_k = (c_{kv})$	Centroid of cluster S_k, $c_{kv} = \sum_{i \in S_k} y_{iv}/N_k$, $v \in V$
S_w, S_{w1}, S_{w2}	Parent-children triple in a cluster hierarchy, $S_w = S_{w1} \cup S_{w2}$
$dw(S_{w1}, S_{w2})$	Ward distance between clusters S_{w1}, with centroid c_1, and S_{w2}, with centroid c_2, $dw(S_{w1}, S_{w2}) = \frac{N_{w1} N_{w2}}{N_{w1} + N_{w2}} d(c_{w1}, c_{w2})$
N_{kv}	Number of entities in class S_k of partition S $(k = 1, ..., K)$ that fall in category $v \in V$; an entry in the contingency table between partition S and categorical feature l with set of categories V_l

N_{k+}	Marginal distribution: Number of entities in class S_k of partition S $(k = 1, ..., K)$ as related to a contingency table between partition S and another categorical feature
N_{+v}	Marginal distribution: Number of entities falling in category $v \in V_l$ of categorical feature l as related to a contingency table between partition S and categorical feature l
p_{kv}	Frequency N_{kv}/N
p_{k+}	N_{k+}/N
p_{+v}	N_{+v}/N
q_{kv}	Relative Quetelet coefficient, $q_{kv} = \frac{p_{kv}}{p_{k+}p_{+v}} - 1$
$T(Y)$	Data scatter, $T(Y) = \sum_{i \in I} \sum_{v \in V} y_{iv}^2$
$W(S_k, c_k)$	Cluster's square error, $W(S_k, c_k) = \sum_{i \in S_k} d(y_i, c_k)$
$W(S, c)$	K-Means square error criterion equal to the sum of $W(S_k, c_k)$, $k = 1, ...K$
$\beta(i, S_k)$	Attraction of $i \in I$ to cluster S_k

Introduction: Historical Remarks

Clustering is a discipline aimed at revealing groups, or clusters, of similar entities in data. The existence of clustering activities can be traced a hundred years back, in different disciplines in different countries.

One of the first was the discipline of ecology. A question the scientists were trying to address was of the territorial structure of the settlement of bird species and its determinants. They did field sampling to count numbers of various species at observation spots; similarity measures between spots were defined, and a method of analysis of the structure of similarity dubbed Wrozlaw taxonomy was developed in Poland between WWI and WWII (see publication of a later time [32]). This method survives, in an altered form, in diverse computational schemes such as single-linkage clustering and minimum spanning tree (see section 6.2.1).

Simultaneously, phenomenal activities in differential psychology initiated in the United Kingdom by the thrust of F. Galton (1822-1911) and supported by the mathematical genius of K. Pearson (1855-1936) in trying to prove that human talent is not a random gift but inherited, led to developing a body of multivariate statistics including the discipline of factor analysis (primarily, for measuring talent) and, as its offshoot, cluster analysis. Take, for example, a list of high school students and their marks at various disciplines such as maths, English, history, etc. If one believes that the marks are exterior manifestations of an inner quality, or factor, of talent, then one can assign a student i with a hidden factor score of his talent, z_i. Then marks x_{il} of student i at different disciplines l can be modeled, up to an error, by the product $c_l z_i$ so that $x_{il} \approx c_l z_i$ where factor c_l reflects the impact of the discipline l over students. The problem is to find the unknown z_i and c_l, given a set of students' marks over a set of disciplines. This was the idea behind a method proposed by K. Pearson in 1901 [106] that became the ground for later developments in Principal Component Analysis (PCA), see further explanation in section 5.1.3. To do the job of measuring hidden factors, F. Galton hired C. Spearman who devel-

oped a rather distinct method for factor analysis based on the assumption that no unique talent can explain various human abilities, but there are different, and independent, dimensions of talent such as linguistic or spatial ones. Each of these hidden dimensions must be presented by a corresponding independent factor so that the mark can be thought of as the total of factor scores weighted by their loadings. This idea proved fruitful in developing various personality theories and related psychological tests. However, methods for factor analysis developed between WWI and WWII were computationally intensive since they used the operation of inversion of a matrix of discipline-to-discipline similarity coefficients (covariances, to be exact). The operation of matrix inversion still can be a challenging task when the matrix size grows into thousands, and it was a nightmare before the electronic computer era even with a matrix size of a dozen. It was noted then that variables (in this case, disciplines) related to the same factor are highly correlated among themselves, which led to the idea of catching "clusters" of highly correlated variables as proxies for factors, without computing the inverse matrix, an activity which was referred to once as "factor analysis for the poor." The very first book on cluster analysis, within this framework, was published in 1939 [131], see also [55].

In the 50s and 60s of the 20th century, with computer powers made available at universities, cluster analysis research grew fast in many disciplines simultaneously. Three of these seem especially important for the development of cluster analysis as a scientific discipline.

First, machine learning of groups of entities (pattern recognition) sprang up to involve both supervised and unsupervised learning, the latter being synonymous to cluster analysis [21].

Second, the discipline of numerical taxonomy emerged in biology claiming that a biological taxon, as a rule, could not be defined in the Aristotelian way, with a conjunction of features: a taxon thus was supposed to be such a set of organisms in which a majority shared a majority of attributes with each other [122]. Hierarchical agglomerative and divisive clustering algorithms were supposed to formalize this. They were being "polythetic" by the very mechanism of their action in contrast to classical "monothetic" approaches in which every divergence of taxa was to be explained by a single character. (It should be noted that the appeal of numerical taxonomists left some biologists unimpressed; there even exists the so-called "cladistics" discipline that claims that a single feature ought always to be responsible for any evolutionary divergence.)

Third, in the social sciences, an opposite stance of building a divisive decision tree at which every split is made over a single feature emerged in the work of Sonquist and Morgan (see a later reference [124]). This work led to the development of decision tree techniques that became a highly popular part of machine learning and data mining. Decision trees actually cover three methods, conceptual clustering, classification trees and regression trees, that are usually

considered different because they employ different criteria of homogeneity [58]. In a conceptual clustering tree, split parts must be as homogeneous as possible with regard to all participating features. In contrast, a classification tree or regression tree achieves homogeneity with regard to only one, so-called target, feature. Still, we consider that all these techniques belong in cluster analysis because they all produce split parts consisting of similar entities; however, this does not prevent them also being part of other disciplines such as machine learning or pattern recognition.

A number of books reflecting these developments were published in the 70s describing the great opportunities opened in many areas of human activity by algorithms for finding "coherent" clusters in a data "cloud" placed in 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). In the next decade, some of these developments have been further advanced and presented in such books as Breiman et al. [11], Jain and Dubes [58] and McLachlan and Basford [82]. Still the common view is that clustering is an art rather than a science because determining clusters may depend more on the user's goals than on a theory. Accordingly, clustering is viewed as a set of diverse and ad hoc procedures rather than a consistent theory.

The last decade saw the emergence of data mining, the discipline combining issues of handling and maintaining data with approaches from statistics and machine learning for discovering patterns in data. In contrast to the statistical approach, which tries to find and fit objective regularities in data, data mining is oriented towards the end user. That means that data mining considers the problem of useful knowledge discovery in its entire range, starting from database acquisition to data preprocessing to finding patterns to drawing conclusions. In particular, the concept of an interesting pattern as something which is unusual or far from normal or anomalous has been introduced into data mining [29]. Obviously, an anomalous cluster is one that is further away from the grand mean or any other point of reference – an approach which is adapted in this text.

A number of computer programs for carrying out data mining tasks, clustering included, have been successfully exploited, both in science and industry; a review of them can be found in [23]. There are a number of general purpose statistical packages which have made it through from earlier times: those with some cluster analysis applications such as SAS [119] and SPSS[42] or those entirely devoted to clustering such as CLUSTAN [140]. There are data mining tools which include clustering, such as Clementine [14]. Still, these programs are far from sufficient in advising a user on what method to select, how to pre-process data and, especially, what sense to make of the clusters.

Another feature of this more recent period is that a number of application

areas have emerged in which clustering is a key issue. In many application areas that began much earlier – such as image analysis, machine vision or robot planning – clustering is a rather small part of a very complex task such that the quality of clustering does not much matter to the overall performance; as any reasonable heuristic would do, these areas do not require the discipline of clustering to theoretically develop and mature.

This is not so in Bio-informatics, the discipline which tries to make sense of interrelation between structure, function and evolution of biomolecular objects. Its primary entities, DNA and protein sequences, are complex enough to have their similarity modeled as homology, that is, inheritance from a common ancestor. More advanced structural data such as protein folds and their contact maps are being constantly added to existing depositories. Gene expression technologies add to this an invaluable next step - a wealth of data on biomolecular function. Clustering is one of the major tools in the analysis of bioinformatics data. The very nature of the problem here makes researchers see clustering as a tool not only for finding cohesive groupings in data but also for relating the aspects of structure, function and evolution to each other. In this way, clustering is more and more becoming part of an emerging area of computer classification. It models the major functions of classification in the sciences: the structuring of a phenomenon and associating its different aspects. (Though, in data mining, the term 'classification' is almost exclusively used in its partial meaning as merely a diagnostic tool.) Theoretical and practical research in clustering is thriving in this area.

Another area of booming clustering research is information retrieval and text document mining. With the growth of the Internet and the World Wide Web, text has become one of the most important mediums of mass communication. The terabytes of text that exist must be summarized effectively, which involves a great deal of clustering in such key stages as natural language processing, feature extraction, categorization, annotation and summarization. In author's view, clustering will become even more important as the systems for acquiring and understanding knowledge from texts evolve, which is likely to occur soon. There are already web sites providing web search results with clustering them according to automatically found key phrases (see, for instance, [134]).

This book is mostly devoted to explaining and extending two clustering techniques, K-Means for partitioning and Ward for hierarchical clustering. The choice is far from random. First, they present the most popular clustering formats, hierarchies and partitions, and can be extended to other interesting formats such as single clusters. Second, many other clustering and statistical techniques, such as conceptual clustering, self-organizing maps (SOM), and contingency association measures, appear to be closely related to these. Third, both methods involve the same criterion, the minimum within cluster variance, which can be treated within the same theoretical framework. Fourth, many data

mining issues of current interest, such as analysis of mixed data, incomplete clustering, and conceptual description of clusters, can be treated with extended versions of these methods. In fact, the book contents go far beyond these methods: the two last chapters, accounting for one third of the material, are devoted to the "big issues" in clustering and data mining that are not limited to specific methods.

The present account of the methods is based on a specific approach to cluster analysis, which can be referred to as the *data recovery clustering*. In this approach, clusters are not only found in data but they also feed back into the data: a cluster structure is used to generate data in the format of the data table which has been analyzed with clustering. The data generated by a cluster structure are, in a sense, "ideal" as they reproduce only the cluster structure lying behind their generation. The observed data can then be considered a noisy version of the ideal cluster-generated data; the extent of noise can be measured by the difference between the ideal and observed data. The smaller the difference the better the fit. This idea is not particularly new; it is, in fact, the backbone of many quantitative methods of multivariate statistics, such as regression and factor analysis. Moreover, it has been applied in clustering from the very beginning; in particular, Ward [135] developed his method of agglomerative clustering with implicitly this view of data analysis. Some methods were consciously constructed along the data recovery approach: see, for instance, work of Hartigan [46] at which the single linkage method was developed to approximate the data with an ultrametric matrix, an ideal data type corresponding to a cluster hierarchy. Even more appealing in this capacity is a later work by Hartigan [47].

However, this approach has never been applied in full. The sheer idea, following from models presented in this book, that classical clustering is but a constrained analogue to the principal component model has not achieved any popularity so far, though it has been around for quite a while [89], [90]. The unifying capability of the data recovery clustering is grounded on convenient relations which exist between data approximation problems and geometrically explicit classical clustering. Firm mathematical relations found between different parts of cluster solutions and data lead not only to explanation of the classical algorithms but also to development of a number of other algorithms for both finding and describing clusters. Among the former, principal-component-like algorithms for finding anomalous clusters and divisive clustering should be pointed out. Among the latter, a set of simple but efficient interpretation tools, that are absent from the multiple programs implementing classical clustering methods, should be mentioned.