

Part I

Cluster Analysis and Graph Clustering

Cluster analysis

Cluster analysis came into being as a bundling of exploratory data techniques which were scattered over many sciences. This embodied an effort to unify and distinguish different frameworks, to separate method from data, and to separate implementation from method. Methods still exist in multitudes, but the reasons for this are well understood. Contrasting this numerosity is the prevalence of a single data model in the cluster analysis monographs, namely that where entities are represented by vectors. Special attention is paid to this issue throughout this chapter and the next, as this thesis is concerned with a cluster algorithm in the setting of graphs.

In Section 2.1 the position of cluster analysis as an exploratory data analysis science is summarized. This section is concise, because it is of limited interest for the subject of graph clustering. A short history of cluster analysis, tied to the perspective of exploratory data analysis, is found in Section 4 of the appendix *A cluster miscellany* (page 154). In Section 2.2 problems and methods from the pattern recognition sciences are introduced. These are interesting because clustering methods have found employment there, and because (intermediate) problems and results in pattern recognition are often phrased in terms of graphs. Two graph based stochastic methods, namely Hidden Markov Models and Markov Random Fields, are discussed in order to illustrate the versatility of stochastic graph concepts and to exemplify the significant conceptual and mathematical differences between these methods and the *MCL* process. Section 2.3 is concerned with the history and characteristics of the research that is collectively labelled as cluster analysis, and the extent to which this label corresponds with a coherent discipline from a mathematical point of view.

2.1 Exploratory data analysis

Cluster analysis is usually seen as the result of cross-fertilization between mathematics and sciences such as biology, chemistry, medicine, and psychology. The latter provide the practical applications that yield the problem formulation, while the study of those problems in an abstract setting belongs to mathematics proper. In this classic setting (more than a century old now) the nature of the problem formulation is that of *exploring* data to see if cohesive structure is present; cluster analysis is then ranged under the flag of *exploratory data analysis*. Other mathematical disciplines of this type are *discriminant analysis*: assigning objects to (a priori known) classes given a number of possibly incomplete observations, *factor analysis*: uncovering correlations between variables by their observed behaviour, *mixture resolving*: estimating parameters for mixtures of distributions (e.g. via the Maximum-Likelihood method), and *dispersion analysis*: methods

for detecting the effect of individual factors on the results of an experiment. Among this list cluster analysis is the most vague in the sense that neither problem nor aim allow a satisfactory description. The least is known — observations only — and the most is wanted: a classification of these, resulting in a bootstrapping problem in optima forma. Usually, the more constrained a problem is, the more this suggests a particular way of solving it. The reverse is true as well; Cluster analysis is pervaded by a host of different mathematical techniques (Section 2.3). The most prominent data model in exploratory data analysis is that where entities are represented by vectors (representing scores on sets of attributes). I refer to this setting as *the vector model*. It is also discussed in greater detail in Section 2.3.

2.2 Pattern recognition sciences

New employment for exploratory techniques has been found in the young but vibrant field of *pattern recognition*, along with a host of other techniques from for example statistical decision theory, signal processing, and linear algebra. In the pattern recognition sciences methods are studied for emulation of cognitive skills, i.e. for detection or recognition of structure in data. Very often these data correspond with auditory or visual signals, or with variables which are measured along the coordinates of a two or three-dimensional space. Examples of this kind are speech recognition, food screening, matching of fingerprints, machine vision, satellite image processing, and more generally the processing of geographic, atmospheric, oceanic, or astronomic data. The hardness of the problems varies widely: the more is known a priori, the easier the problem. The difficulty of matching problems, where it has to be decided whether a new object is the same as one out of a collection of known objects, depends highly on the variability of the object appearances. Iris or fingerprint matching is thus relatively easy, whereas speech recognition or recognition of 3-D objects is much more difficult. Utterances may vary in tone, pronunciation, emphasis, colour, and duration, even for a single person. The appearance of objects may vary according to the angle of view, the orientation of the object, the viewing distance, the background, the location and intensity of the light source(s), and the overlap by other objects.

2.2.1 Cluster analysis applications. Some of the applications of cluster analysis in pattern recognition (usually corresponding with an intermediate processing stage) as listed by Jain and Dubes in [94] are: grammatical inference, speech and speaker recognition, image segmentation, and image matching. In general, the role of cluster analysis is the joining of primitive data elements in regions or time-frames, which do not yet need to correspond with high-level objects. Clustering is thus a base method for diminishing dimensionality. Other applications of clustering in pattern recognition are reducing feature dimensionality and reducing the dimensionality of a search space (e.g. the set of all Chinese characters).

2.2.2 The data model in spatial pattern recognition. At first sight image segmentation and spatial data segmentation seem to fit well within the vector model, as they concern measurements in Euclidean spaces. However, the data model differs considerably from

that in exploratory data analysis. In the latter setting, the distribution of the vectors themselves over the geometry is of interest. In spatial data segmentation the vectors are just the (x, y) coordinates of pixels or the (x, y, z) coordinates of voxels (the 3-D equivalent of a pixel). The vectors sample some area, which is usually box-shaped. The proximity between two vectors is not related to their distance, but defined in terms of the similarity between the measurements on the corresponding units of the sample space. Typically, only neighbouring pixels or voxels are considered for the proximity relationship, as one is interested in finding regions of contiguous units, which are homogeneous with respect to the measurements. This localization of the neighbour relationship induces a lattice or grid-like graph. Consequently, graph-theoretical concepts and techniques play an important role in spatial data segmentation.

This is in fact true for the field of pattern recognition at large. Data may be split up in collections of primitive patterns. The primitive patterns are to be matched with a catalogue of generic primitives, while at the same time the collection has to be split into higher-level patterns representing objects. Examples of primitive patterns are phonemes in speech recognition and stroke primitives in optical character recognition. In the latter case, characters can also be viewed as primitives for the word-level, and words can be viewed as primitives for the sentence level. The important thing is that the interaction or succession of such primitives is governed by constraints. This induces graphs where the primitives are nodes and the constraints induce and exclude neighbour relations. The fertile combination of graphs and stochastic models in pattern recognition is the subject of the following section.

2.2.3 Graph-based stochastic methods. Two stochastic methods in pattern recognition deserve special mention, namely Hidden Markov Models (HMM) and Markov Random Fields (MRF). These techniques draw upon the same mathematical apparatus as the *MCL* process — which is where the resemblance more or less stops. The common denominator of HMM, MRF, and MCL is that they all use stochastic concepts in the setting of graphs.

A hidden Markov model is a probabilistic model for data observed in a sequential fashion, based on two primary assumptions. The first assumption is that the observed data arise from a mixture of K probability distributions corresponding with K states. The second assumption is that there is a discrete Markov chain generating the observed data by visiting the K states according to the Markov model. The hidden aspect of the model arises from the fact that the state sequence is not directly observed. Instead, the state sequence must be inferred from a sequence of observed data using the probability model (adapted from [170], page 373). The most noteworthy application of the hidden Markov model is found in speech recognition, where it is used both for the recognition of words in terms of phonemes (which are the states), and the construction of sentences from words. In both cases the sequence of primitives (phonemes or words) is governed by constraints, in the sense that the true state of the previously observed primitive induces a probability distribution on the most likely state of the currently observed primitive. This is modelled by a Markov chain of transition probabilities. In [170] computational

biology is listed as another area where HMMs have found employment, and [157] names the applications *lip reading* and *face tracking*.

A Markov Random Field is the counterpart of a one-dimensional Markov chain, where the bidirectionality of past–present–future is superseded by the spatial concept of neighbouring sites. The origin of this concept is found in the Ising model for ferromagnetism, and in physics it is applied to statistical mechanical systems exhibiting phase transitions ([100], page 120). In pattern recognition, MRFs are applied to fundamental tasks such as segmentation and restoration of images, edge detection, and texture analysis and synthesis.

As an example, consider a rectangular array of pixels, where each pixel may assume one of G grey levels. Typically an MRF is used to model the fact that the grey levels of neighbouring pixels are correlated. This correlation reflects higher-level properties of an image, such as the presence of regions, boundaries, and texture. An MRF yields the tools to infer such high-level properties from the low-level pixel information by stochastic techniques. It is assumed that the overall shading of the array is the result of a stochastic process indexed by the pixels. An MRF requires a symmetric neighbourhood structure defined for the pixels; the simplest case is that where only adjacent pixels are neighbours. The property that defines an MRF is that *The conditional probability that a pixel assumes a grey level, given the grey levels of all other pixels, is equal to the conditional probability that it assumes the grey level given only the grey levels of its neighbouring pixels.*

The state space of the MRF is thus enormous; it amounts to all possible grey colourings of the array. An MRF can also be used to model the presence of edges or boundaries in a picture, by creating an edge array similar to the pixel array¹. The edge array can be observed only indirectly by looking at the pixel array. The equivalence of MRFs with Gibbs distributions allows the modelling of expectations regarding boundaries — smoothness, continuity — in the MRF via so called clique potentials. Combining the intensity and boundary processes yields a compound MRF which can be used to restore noisy images, detect edges, detect regions, or a combination of these [33]. This requires specification of a measurement model for the observed image (i.e. choice of the stochastic model, parameter estimation). A typical approach for image restoration is: *A maximum a posteriori probability estimate of the image based on the noisy observations then [after specification of the measurement model] is found by minimizing the posterior Gibbs energy via simulated annealing* ([63], page 499). The MRF model is very powerful in its flexibility — diverse types of interaction (corresponding e.g. with texture, regions, and edges) between pixels can be modelled, and it is backed up by a sound mathematical apparatus. However, its computational requirements are considerable, and the issues of supervised and unsupervised parameter learning are no less difficult than they are in general.

The Hidden Markov Model and the theory of Markov Random Fields illustrate the versatility of the graph model in pattern recognition, and its use in inferencing higher-level structure by stochastic techniques. The stochastic concepts used in the HMM and MRF

¹A horizontal respectively vertical edge is thought to separate two vertically respectively horizontally adjacent pixels.

and their applications are closely tied to a stochastic view on the realization and structure of patterns in real life. This situation is somewhat different for the *MCL* algorithm: its formulation uses the notion of *flow*, utilizing the presence of cohesive structure, rather than drawing upon a model for the realization of this structure and retrieving it via a posteriori optimization.

2.2.4 The position of cluster analysis. Publications on cluster analysis in the setting of pattern recognition are usually rather theoretical, and blend in with publications in the more traditional line of exploratory data research. Most publications which are solely concerned with clustering contain either a new cluster algorithm or suggestions for improvement of existing algorithms. A few publications focusing on the graph model do exist, but these are mostly concerned with neighbourhood graphs (see Section 3.3 in the next chapter).

Clustering methods are occasionally mentioned as a tool for segmentation or as a necessary intermediate processing step in various applications, but I have not found any systematic comparison of different methods plugged into the same hole. The most likely reason for this is that the role of clustering is not sufficiently essential, as it is not a dedicated solution for a specific problem. Benchmarking is in principle very well possible and Chapter 12 is devoted to this issue.

2.3 Methods and concepts in multitudes

Cluster analysis is a field that has always been driven by a demand from various disciplines engaged in exploratory data analysis, like for example taxonomy, chemistry, medicine, psychiatry, market research, et cetera. The monographs on cluster analysis give long listings of applications, see e.g. Anderberg [10], Everitt [54], and Mirkin [132]. This wide range of interest groups, most of which do not have common channels of communication, has resulted in an enormous variety of clustering methods. Adding to this is the elusive nature of the problem. There is no obvious step immediately transforming the loose formulation of the problem into a strategy for solving it. Rather, ten different people will come up with ten different methods, and they may well come up with twenty, because the problem is attractive in that it seems intuitively so simple, yet in practice so hard. On a related note, Blashfield et al report in [22] on a questionnaire regarding the use of cluster software, with fifty-three respondents yielding fifty different programs and packages. It is sometimes said that there are as many cluster methods as there are cluster analysis users. This thesis indeed offers Yet Another Cluster Method². The method operates in a setting that is relatively new to cluster analysis though (see the next chapter), it depends on an algebraic process that has not been studied before, and this process has properties that make it particularly suitable as a means for detecting cluster structure (Parts II and III of this thesis).

²This is yet another Yet Another qualification; see page 160.

2.3.1 Conceptual confusion. Different disciplines use different concepts and different wordings, so cluster-related research has yielded a plethora of methods and concepts. Kaufman and Rousseeuw put it this way in [103], page vii: *Rather than giving an extensive survey of clustering methods, leaving the user with a bewildering multitude of methods to choose from (...)*. They have a positive attitude, as they write on page 3:

(...) automatic classification is a very young scientific discipline in vigorous development, as can be seen from the thousands of articles scattered over many periodicals (mostly journals of statistics, biology, psychometrics, computer science, and marketing). Nowadays, automatic classification is establishing itself as an independent scientific discipline (...)

The qualification ‘very young’ is debatable though, as the first survey articles and monographs in cluster analysis began to appear in the sixties and early seventies. The purpose of Kaufman and Rousseeuw was to write an applied book for the general user. In the field of classification, the book by Jardine and Sibson [95] is a long standing reference which aims to give a mathematical account of the methods employed in taxonomy (the theory of classification of living entities). The significance of this book is enlarged by its emphasis on mathematics rather than limited by its focus on taxonomy. The tone of their introduction is somewhat more dejected ([95], page ix):

Terminological confusion and the conceptual confusions which they conceal have loomed large in the development of methods of automatic classification. Partly this has arisen from the parallel development of related methods under different names in biology, pattern recognition, psychology, linguistics, archaeology, and sociology. Partly it has arisen from failure by mathematicians working in the various fields to realize how diverse are the problems included under the headings ‘classification’, ‘taxonomy’, and ‘data analysis’.

In mathematics, the elusive nature of the vector cluster problem is reflected in the fact that there is no particular piece of mathematical machinery just right for the job. The result is that many mathematical tools and disciplines are used in modelling of the problem domain and formulation of cluster strategies. The list includes matrix algebra, geometry, metric spaces, statistics, set theory, graph theory, information theory, and several combinations out of these. Witness Mirkin in [132], page xiv: (...) *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 words *introductory* and *elementary* should be noted though. Cluster analysis draws upon the different disciplines mainly for formulation of both problems and strategies sought to apply to them. With few exceptions, not much new material is being developed nor is there much need to apply existing theorems.

Clustering has also been presented in the setting of computing paradigms such as simulated annealing [97, 106], genetic algorithms, and neural networks. In these instances however, the computing paradigms embraced the clustering problem rather than vice versa, and the approaches have not yet swept the area.

2.3.2 The prevalent data model in cluster analysis. *Cluster Analysis is the mathematical study of methods for recognizing natural groups within a class of entities.* Consider this definition once more. It speaks of 'natural groups', indicating that entities are somehow related to each other. Apparently, there is a notion of greater or lesser distance or similarity between entities, or it would be impossible to discriminate between different pairs and constellations of entities. It is actually difficult to imagine a class of entities where such a notion is lacking; the art of discrimination is that essential to sensory systems, intelligence, and awareness. Whereas this digression may seem too aspiring, it appears that attempts to formulate the essence of cluster analysis readily lead to such cognitive and philosophical issues. For example, Sokal writes in [156], page 1:

But since classification is the ordering of objects by their similarities (...) and objects can be conceived of in the widest sense including processes and activities — anything to which a vector of descriptors can be attached, we recognize that classification transcends human intellectual endeavor and is indeed a fundamental property of living organisms. Unless they are able to group stimuli into like kinds so as to establish classes to which favorable or avoidance reactions can be made, organisms would be ill-adapted for survival.

This kind of comment is frequently found in the literature, and it spells out the burdensome idea that methods in cluster analysis have to compete with nothing less than a fundamental property of living organisms. The citation also demonstrates that at the time of writing (1976) cluster analysis was more or less tied to the framework where entities are represented by vectors, an observation still valid today. In the framework each entity is examined with respect to a fixed number of characteristics; this gives a vector of numbers, and this vector represents the entity. For example, if the entities are medical records, then the characteristics might be *length, weight, age, alcohol consumption, blood pressure, liver thickening*, if the entities are meteorite rocks then the characteristics can be the respective fractions of different types of chemical compounds. The distance between two entities is then defined to be a function of the difference of the two corresponding vectors. Usually a scaling of the vectors is involved in order to weigh different characteristics, and some norm is taken of the difference vector.

This setting is prevalent throughout the cluster analysis literature, and it is the de facto standard in monographs on the subject — see the list of monographs compiled in the next section. On the one hand, this persistence of the same data model is hardly surprising, as it is a highly generic model that apparently suits the existing needs. On the other hand, there is an interest in clustering algorithms in the graph partitioning community and occasional other areas, feeding a small but steady stream of publications which is isolated from the cluster analysis literature at large in terms of references and surveys. This is discussed in the next chapter.

2.3.3 The coherence of cluster analysis as a discipline. As evidenced in the beginning of this section, the field of cluster analysis is of fragmented heritage. This situation began to change somewhat with the advent of dedicated conferences and publication of several monographs on the subject, beginning with the books by Sokal and Sneath [155] (1963), Jardine and Sibson [95] (1971), Anderberg [10] (1973), Duda and

Hart [50] (1973), Everitt [54] (1974), and Hartigan [79] (1975). Other often cited references are Ryzin [166] (conference proceedings, 1977), Kaufman and Rousseeuw [103] (1983), Jain and Dubes [94] (1988), Massart and Kaufman [119] (1990), and Mirkin [132] (1996). These monographs often contain huge collections of references to articles from different disciplines, and they embody considerable efforts to create a unified approach to cluster analysis. The citation of Good on page 157 reflects a certain longing for the field to become whole. The same feeling is occasionally expressed by other researchers. According to Kaufman and Rousseeuw the era of cluster analysis as an independent scientific discipline has already begun (see the citation on page 22). However, much can be brought to stand against this point of view, without disputing the viability of the label 'cluster analysis' as a flag under which a motley crew sails.

The most prominent reason why cluster analysis is not the home ground of a coherent scientific community is simply the wide variety of mathematical concepts that can be employed. No single branch of mathematics is particularly apt for the job, and problems in cluster analysis defy a clinching formal description. Such a description may emerge, but the general opinion is that the conditions are unfavourable, as the generic clustering problem formulation assumes little and aspires much.

A second reason is found in the lack of an application area with a very definite and urgent need for good clustering algorithms, with common cost functions and benchmarks, and preferably with problem instances that obey some variant of Moore's law — a doubling of the typical problem size every three years or so. Graph partitioning, with its current emphasis on multilevel approaches, may turn out to provide such pull from the application side for cluster methods in the setting of graphs. Otherwise, this is a phenomenon unfamiliar to cluster analysis, simply because the large majority of applications either has an exploratory nature, or concerns the embedding of clustering methods in a compound (e.g. recognition) method that is not widely recognized as a fundamental step in the solution of a problem of general importance and urgency.

All this can be summed up in the observation that researchers in cluster analysis share neither a common mathematical framework nor a common application area generating sufficient pull. This does not imply that there is no use for the gathering forces of monographs and conferences dedicated to the subject. The opposite is true, as there is without doubt a general interest in clustering methods, and a real danger of wasted efforts and reinvention of the wheel over and over again. However, the state of affairs in cluster analysis is that the gathering of efforts does not have a strong synergetic effect.