1

An introduction to classification and clustering

1.1 Introduction

> An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.
>
> Steven Pinker, *How the Mind Works*, 1997.

One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. The idea of sorting similar things into categories is clearly a primitive one since early man, for example, must have been able to realize that many individual objects shared certain properties such as being edible, or poisonous, or ferocious and so on.

Classification, in its widest sense, is needed for the development of language, which consists of words which help us to recognize and discuss the different types of events, objects and people we encounter. Each noun in a language, for example, is essentially a label used to describe a class of things which have striking features in common; thus animals are named as cats, dogs, horses, etc., and such a name collects individuals into groups. Naming and classifying are essentially synonymous.

As well as being a basic human conceptual activity, classification is also fundamental to most branches of science. In biology for example, classification of organisms has been a preoccupation since the very first biological investigations. Aristotle built up an elaborate system for classifying the species of the animal

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

kingdom, which began by dividing animals into two main groups, those having red blood (corresponding roughly to our own vertebrates), and those lacking it (the invertebrates). He further subdivided these two groups according to the way in which the young are produced, whether alive, in eggs, as pupae and so on.

Following Aristotle, Theophrastos wrote the first fundamental accounts of the structure and classification of plants. The resulting books were so fully documented, so profound and so all-embracing in their scope that they provided the groundwork of biological research for many centuries. They were superseded only in the 17th and 18th centuries, when the great European explorers, by opening the rest of the world to inquiring travellers, created the occasion for a second, similar programme of research and collection, under the direction of the Swedish naturalist, Linnaeus. In 1737, Carl von Linné published his work *Genera Plantarum*, from which the following quotation is taken:

> All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar. The greater the number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary.
>
> For we must not join in the same genus the horse and the swine, though both species had been one hoof'd nor separate in different genera the goat, the reindeer and the elk, tho' they differ in the form of their horns. We ought therefore by attentive and diligent observation to determine the limits of the genera, since they cannot be determined *a priori*. This is the great work, the important labour, for should the genera be confused, all would be confusion.

In biology, the theory and practice of classifying organisms is generally known as *taxonomy*. Initially, taxonomy in its widest sense was perhaps more of an art than a scientific method, but eventually less subjective techniques were developed largely by Adanson (1727–1806), who is credited by Sokal and Sneath (1963) with the introduction of the *polythetic* type of system into biology, in which classifications are based on many characteristics of the objects being studied, as opposed to *monothetic* systems, which use a single characteristic to produce a classification.

The classification of animals and plants has clearly played an important role in the fields of biology and zoology, particularly as a basis for Darwin's theory of evolution. But classification has also played a central role in the developments of theories in other fields of science. The classification of the elements in the periodic table for example, produced by Mendeleyev in the 1860s, has had a profound impact on the understanding of the structure of the atom. Again, in astronomy, the classification of stars into *dwarf* stars and *giant* stars using the Hertzsprung–Russell plot of temperature against luminosity (Figure 1.1) has strongly affected theories of stellar evolution.

Classification may involve people, animals, chemical elements, stars, etc., as the entities to be grouped. In this text we shall generally use the term *object* to cover all such possibilities.

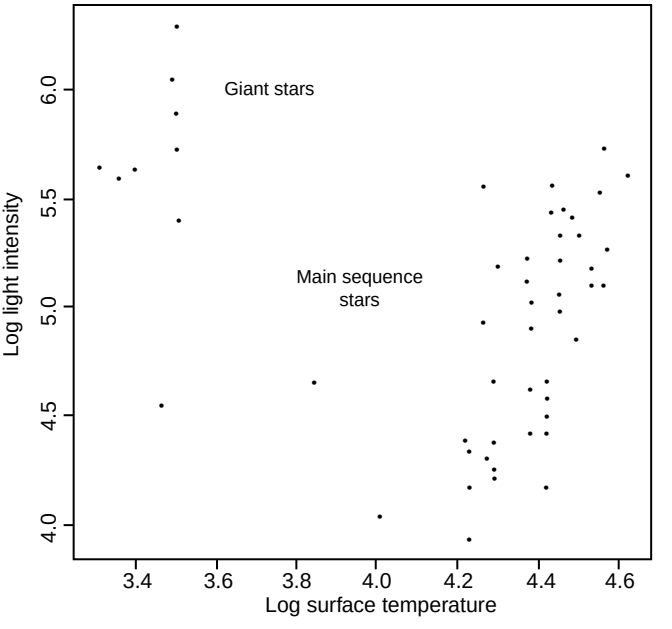

Figure 1.1 *Hertzsprung–Russell plot of temperature against luminosity.*

1.2 Reasons for classifying

At one level, a classification scheme may simply represent a convenient method for organizing a large data set so that it can be understood more easily and information retrieved more efficiently. If the data can validly be summarized by a small number of groups of objects, then the group labels may provide a very concise description of patterns of similarities and differences in the data. In market research, for example, it may be useful to group a large number of respondents according to their preferences for particular products. This may help to identify a 'niche product' for a particular type of consumer. The need to summarize data sets in this way is increasingly important because of the growing number of large databases now available in many areas of science, and the exploration of such databases using cluster analysis and other multivariate analysis techniques is now often called *data mining*. In the 21st century, data mining has become of particular interest for investigating material on the World Wide Web, where the aim is to extract useful information or knowledge from web page contents (see, Liu, 2007 for more details).

In many applications, however, investigators may be looking for a classification which, in addition to providing a useful summary of the data, also serves some more fundamental purpose. Medicine provides a good example. To understand and treat disease it has to be classified, and in general the classification will have two main aims. The first will be *prediction* – separating diseases that require different

treatments. The second will be to provide a basis for research into *aetiology* – the causes of different types of disease. It is these two aims that a clinician has in mind when she makes a diagnosis.

It is almost always the case that a variety of alternative classifications exist for the same set of objects. Human beings, for example, may be classified with respect to *economic status* into groups such as *lower class*, *middle class* and *upper class*; alternatively they might be classified by annual consumption of alcohol into *low*, *medium* and *high*. Clearly such different classifications may not collect the same individuals into groups. Some classifications are, however, more likely to be of general use than others, a point well-made by Needham (1965) in discussing the classification of humans into men and women:

> The usefulness of this classification does not begin and end with all that can, in one sense, be strictly inferred from it – namely a statement about sexual organs. It is a very useful classification because classing a person as a man or woman conveys a great deal more information, about probable relative size, strength, certain types of dexterity and so on. When we say that persons in class *man* are more suitable than persons in class *woman* for certain tasks and conversely, we are only incidentally making a remark about sex, our primary concern being with strength, endurance etc. The point is that we have been able to use a classification of persons which conveys information on many properties. On the contrary a classification of persons into those with hair on their forearms between $\frac{3}{16}$ and $\frac{1}{4}$ inch long and those without, though it may serve some particular use, is certainly of no general use, for imputing membership in the former class to a person conveys information in this property alone. Put another way, there are no known properties which divide up a set of people in a similar manner.

A similar point can be made in respect of the classification of books based on subject matter and their classification based on the colour of the book's binding. The former, with classes such as *dictionaries*, *novels*, *biographies,* etc., will be of far wider use than the latter with classes such as *green*, *blue*, *red,* etc. The reason why the first is more useful than the second is clear; the subject matter classification indicates more of a book's characteristics than the latter.

So it should be remembered that in general a classification of a set of objects is not like a scientific theory and should perhaps be judged largely on its usefulness, rather than in terms of whether it is 'true' or 'false'.

1.3 Numerical methods of classification – cluster analysis

Numerical techniques for deriving classifications originated largely in the natural sciences such as biology and zoology in an effort to rid taxonomy of its traditionally subjective nature. The aim was to provide *objective* and *stable* classifications. Objective in the sense that the analysis of the same set of organisms by the same sequence of numerical methods produces the same classification; stable in that the

classification remains the same under a wide variety of additions of organisms or of new characteristics describing them.

A number of names have been applied to these numerical methods depending largely on the area of application. *Numerical taxonomy* is generally used in biology. In psychology the term *Q analysis* is sometimes employed. In the artificial intelligence literature *unsupervised pattern recognition* is the favoured label, and market researchers often talk about *segmentation*. But nowadays *cluster analysis* is probably the preferred generic term for procedures which seek to uncover groups in data.

In most applications of cluster analysis a *partition* of the data is sought, in which each individual or object belongs to a single cluster, and the complete set of clusters contains all individuals. In some circumstances, however, overlapping clusters may provide a more acceptable solution. It must also be remembered that one acceptable answer from a cluster analysis is that no grouping of the data is justified.

The basic data for most applications of cluster analysis is the usual $n \times p$ multivariate data matrix, $\mathbf{X}$, containing the variable values describing each object to be clustered; that is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}.$$

The entry x_{ij} in $\mathbf{X}$ gives the value of the jth variable on object i. Such a matrix is often termed 'two-mode', indicating that the rows and columns correspond to different things.

The variables in $\mathbf{X}$ may often be a mixture of continuous, ordinal and/or categorical, and often some entries will be missing. Mixed variables and missing values may complicate the clustering of data, as we shall see in later chapters. And in some applications, the rows of the matrix $\mathbf{X}$ may contain *repeated measures* of the *same* variable but under, for example, different conditions, or at different times, or at a number of spatial positions, etc. A simple example in the time domain is provided by measurements of, say, the heights of children each month for several years. Such *structured data* are of a special nature in that all variables are measured on the same scale, and the cluster analysis of structured data may require different approaches from the clustering of unstructured data, as we will see in Chapter 3 and in Chapter 7.

Some cluster analysis techniques begin by converting the matrix $\mathbf{X}$ into an $n \times n$ matrix of inter-object *similarities*, *dissimilarities* or *distances* (a general term is *proximity*), a procedure to be discussed in detail in Chapter 3. (Such matrices may be designated 'one-mode', indicating that their rows and columns index the same thing.) But in some applications the inter-object similarity or dissimilarity matrix may arise directly, particularly in experiments where people are asked to judge the perceived similarity or dissimilarity of a set of stimuli or objects of interest. As an

Table 1.1 Dissimilarity data for all pairs of 10 colas for 2 subjects.

Subject 1

	Cola number									
	1	2	3	4	5	6	7	8	9	10
1	0									
2	16	0								
3	81	47	0							
4	56	32	71	0						
5	87	68	44	71	0					
6	60	35	21	98	34	0				
7	84	94	98	57	99	99	0			
8	50	87	79	73	19	92	45	0		
9	99	25	53	98	52	17	99	84	0	
10	16	92	90	83	79	44	24	18	98	0

Subject 2

	Cola number									
	1	2	3	4	5	6	7	8	9	10
1	0									
2	20	0								
3	75	35	0							
4	60	31	80	0						
5	80	70	37	70	0					
6	55	40	20	89	30	0				
7	80	90	90	55	87	88	0			
8	45	80	77	75	25	86	40	0		
9	87	35	50	88	60	10	98	83	0	
10	12	90	96	89	75	40	27	14	90	0

example, Table 1.1 shows judgements about various brands of cola made by two subjects, using a visual analogue scale with anchor points 'some' (having a score of 0) and 'different' (having a score of 100). In this example the resulting rating for a pair of colas is a dissimilarity – low values indicate that the two colas are regarded as alike and vice versa. A similarity measure would have been obtained had the anchor points been reversed, although similarities are usually scaled to lie in the interval [0,1], as we shall see in Chapter 3.

In this text our main interest will centre on clustering the objects which define the rows of the data matrix **X**. There is, however, no fundamental reason why some clustering techniques could not be applied to the columns of **X** to cluster the variables, perhaps as an alternative to some form of *factor analysis* (see Everitt and Dunn, 2001). This issue of clustering variables will be taken up briefly in Chapter 8.

Cluster analysis is essentially about *discovering* groups in data, and clustering methods should not be confused with *discrimination* and *assignment* methods (in the artificial intelligence world the term *supervised learning* is used), where the groups are known *a priori* and the aim of the analysis is to construct rules for classifying new individuals into one or other of the known groups. A readable account of such methods is given in Hand (1981). More details of recently developed techniques are available in McLachlan (2004).

1.4 What is a cluster?

Up to this point the terms cluster, group and class have been used in an entirely intuitive manner without any attempt at formal definition. In fact it turns out that formal definition is not only difficult but may even be misplaced. Bonner (1964), for example, has suggested that the ultimate criterion for evaluating the meaning of such terms is the value judgement of the user. If using a term such as 'cluster' produces an answer of value to the investigator, that is all that is required.

Bonner has a point, but his argument is not entirely convincing, and many authors, for example Cormack (1971) and Gordon (1999), attempt to define just what a cluster is in terms of internal cohesion – *homogeneity* – and external isolation – *separation*. Such properties can be illustrated, informally at least, with a diagram such as Figure 1.2. The 'clusters' present in this figure will be clear to most observers without attempting an explicit formal definition of the term. Indeed, the example indicates that no single definition is likely to be sufficient for all situations. This may explain why attempts to make the concepts of homogeneity and separation mathematically precise in terms of explicit numerical indices have led to numerous and diverse criteria.

It is not entirely clear how a 'cluster' is recognized when displayed in the plane, but one feature of the recognition process would appear to involve assessment of the relative distances between points. How human observers draw perceptually coherent clusters out of fields of 'dots' will be considered briefly in Chapter 2.

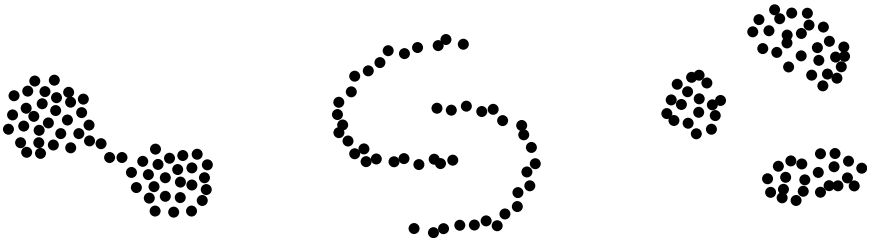

Figure 1.2 *Clusters with internal cohesion and/or external solution. (Reproduced with permission of CRC Press from Gordon, 1980.)*

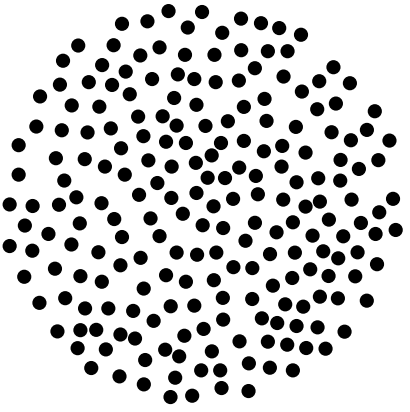

Figure 1.3 *Data containing no 'natural' clusters. (Reproduced with permission of CRC Press from Gordon, 1980.)*

A further set of two-dimensional data is plotted in Figure 1.3. Here most observers would conclude that there is no 'natural' cluster structure, simply a single homogeneous collection of points. Ideally, then, one might expect a method of cluster analysis applied to such data to come to a similar conclusion. As will be seen later, this may not be the case, and many (most) methods of cluster analysis *will* divide the type of data seen in Figure 1.3 into 'groups'. Often the process of dividing a homogeneous data set into different parts is referred to as *dissection*, and such a procedure may be useful in specific circumstances. If, for example, the points in Figure 1.3 represented the geographical locations of houses in a town, dissection might be a useful way of dividing the town up into compact postal districts which contain comparable numbers of houses – see Figure 1.4. (This example was suggested by Gordon, 1980.) The problem is, of course, that since in most cases

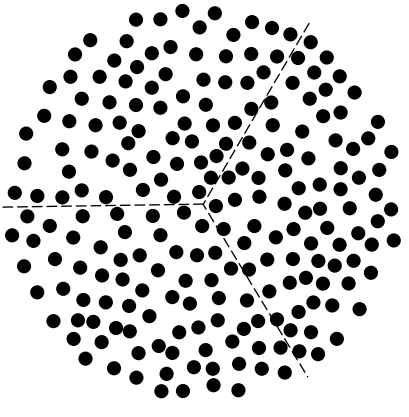

Figure 1.4 *Dissection of data in Figure 1.3 (Reproduced with permission of CRC Press from Gordon, 1980.)*

the investigator does not know *a priori* the structure of the data (cluster analysis is, after all, intended to help to uncover any structure), there is a danger of interpreting *all* clustering solutions in terms of the existence of distinct (natural) clusters. The investigator may then conveniently 'ignore' the possibility that the classification produced by a cluster analysis is an artefact of the method and that actually she is *imposing* a structure on her data rather than discovering something about the actual structure. This is a very real problem in the application of clustering techniques, and one which will be the subject of further discussion in later chapters.

1.5 Examples of the use of clustering

The general problem which cluster analysis addresses appears in many disciplines: biology, botany, medicine, psychology, geography, marketing, image processing, psychiatry, archaeology, etc. Here we describe briefly a number of applications of cluster analysis reported in some of these disciplines. Several of these applications will be described more fully in later chapters, as will a variety of other applications not mentioned below.

1.5.1 Market research

Dividing customers into homogeneous groups is one of the basic strategies of marketing. A market researcher may, for example, ask how to group consumers who seek similar benefits from a product so he or she can communicate with them better. Or a market analyst may be interested in grouping financial characteristics of companies so as to be able to relate them to their stock market performance.

An early specific example of the use of cluster analysis in market research is given in Green *et al.* (1967). A large number of cities were available that could be used as test markets but, due to economic factors, testing had to be restricted to only a small number of these. Cluster analysis was used to classify the cities into a small number of groups on the basis of 14 variables including city size, newspaper circulation and per capita income. Because cities within a group could be expected to be very similar to each other, choosing one city from each group was used as a means of selecting the test markets.

Another application of cluster analysis in market research is described in Chakrapani (2004). A car manufacturer believes that buying a sports car is not solely based on one's means or on one's age but it is more a lifestyle decision, with sports car buyers having a pattern of lifestyle that is different from those who do not buy sports cars. Consequently, the manufacturer employs cluster analysis to try to identify people with a lifestyle most associated with buying sports cars, to create a focused marketing campaign.

1.5.2 Astronomy

Large multivariate astronomical data bases are frequently suspected of containing relatively distinct groups of objects which must be distinguished from each other.

Astronomers want to know how many distinct classes of, for example, stars there are on the basis of some statistical criterion. The typical scientific questions posed are 'How many statistically distinct classes of objects are in this data set and which objects are to be assigned to which classes? Are previously unknown classes of objects present?' Cluster analysis can be used to classify astronomical objects, and can often help astronomers find unusual objects within a flood of data. Examples include discoveries of high-redshift quasars, type 2 quasars (highly luminous, active galactic nuclei, whose centres are obscured by gas and dust), and brown dwarfs.

One specific example is the study reported by Faúndez-Abans *et al.* (1996), who applied a clustering technique due to Ward (1963) (see Chapter 4) to data on the chemical composition of 192 planetary nebulae. Six groups were identified which were similar in many respects to a previously used classification of such objects, but which also showed interesting differences.

A second astronomical example comes from Celeux and Govaert (1992), who apply normal mixture models (see Chapter 6) to stellar data consisting of a population of 2370 stars described by their velocities towards the galactic centre and towards the galactic rotation. Using a three-cluster model, they find a large-size, small-volume cluster, and two small-size, large-volume clusters.

For a fuller account of the use of cluster analysis in astronomy see Babu and Feigelson (1996).

1.5.3 Psychiatry

Diseases of the mind are more elusive than diseases of the body, and there has been much interest in psychiatry in using cluster analysis techniques to refine or even redefine current diagnostic categories. Much of this work has involved depressed patients, where interest primarily centres on the question of the existence of *endogenous* and *neurotic* subtypes. Pilowsky *et al.* (1969), for example, using a method described in Wallace and Boulton (1968), clustered 200 patients on the basis of their responses to a depression questionnaire, together with information about their mental state, sex, age and length of illness. (Notice once again the different types of variable involved.) One of the clusters produced was identified with endogenous depression. A similar study by Paykel (1971), using 165 patients and a clustering method due to Friedman and Rubin (1967) (see Chapter 5), indicated four groups, one of which was clearly psychotic depression. A general review of the classification of depression is given in Farmer *et al.* (1983).

Cluster analysis has also been used to find a classification of individuals who attempt suicide, which might form the basis for studies into the causes and treatment of the problem. Paykel and Rassaby (1978), for example, studied 236 suicide attempters presenting at the main emergency service of a city in the USA. From the pool of available variables, 14 were selected as particularly relevant to classification and used in the analysis. These included age, number of previous suicide attempts, severity of depression and hostility, plus a number of demographic characteristics. A number of cluster methods, for example Ward's method, were applied to the data,

and a classification with three groups was considered the most useful. The general characteristics of the groups found were as follows:

- Group 1: Patients take overdoses, on the whole showing less risk to life, less psychiatric disturbance, and more evidence of interpersonal rather than self-destructive motivation.
- Group 2: Patients in this group made more severe attempts, with more self-destructive motivation, by more violent methods than overdoses.
- Group 3: Patients in this group had a previous history of many attempts and gestures, their recent attempt was relatively mild, and they were overly hostile, engendering reciprocal hostility in the psychiatrist treating them.

A further application of cluster analysis to parasuicide is described in Kurtz *et al.* (1987), and Ellis *et al.* (1996) also investigated the use of cluster analysis on suicidal psychotic outpatients, using *average linkage clustering* (see Chapter 4). They identified four groups which were labelled as follows:

- negativistic/avoidant/schizoid
- avoidant/dependent/negativistic
- antisocial
- histrionic/narcissistic.

And yet another psychiatric example is provided by the controversy over how best to classify eating disorders in which there is recurrent binge eating. Hay *et al.* (1996) investigated the problem by applying Ward's method of cluster analysis to 250 young women each described by five sub-scales derived from the 12th edition of the Eating Disorder Examination (Fairburn and Cooper, 1993). Four subgroups were found:

- objective or subjective bulimic episodes and vomiting or laxative misuse;
- objective bulimic episodes and low levels of vomiting or laxative misuse;
- subjective bulimic episodes and low levels of vomiting or laxative misuse;
- heterogeneous in nature.

1.5.4 Weather classification

Vast amounts of data are collected on the weather worldwide. Exploring such data using cluster analysis may provide new insights into climatological and environmental trends that have both scientific and practical significance. Littmann (2000), for example, applies cluster analysis to the daily occurrences of several surface pressures for weather in the Mediterranean basin, and finds 20 groups that explain rainfall variance in the core Mediterranean regions. And Liu and George (2005) use fuzzy k-means clustering (see Chapter 8) to account for the spatiotemporal nature of weather data in the South Central USA. One further example is provided by Huth

et al. (1993), who analyse daily weather data in winter months (December–February) at Prague Clementinum. Daily weather was characterized by eight variables such as daily mean temperature, relative humidity and wind speed. Average linkage (see Chapter 4) was used to group the data into days with similar weather conditions.

1.5.5 Archaeology

In archaeology, the classification of artefacts can help in uncovering their different uses, the periods they were in use and which populations they were used by. Similarly, the study of fossilized material can help to reveal how prehistoric societies lived. An early example of the cluster analysis of artefacts is given in Hodson *et al*. (1966), who applied single linkage and average linkage clustering (see Chapter 4) to brooches from the Iron Age and found classifications of demonstrable archaeological significance. Another example is given in Hodson (1971), who used a *k-means* clustering technique (see Chapter 5) to construct a taxonomy of hand axes found in the British Isles. Variables used to describe each of the axes included length, breadth and pointedness at the tip. The analysis resulted in two clusters, one of which contained thin, small axes and the other thick, large axes, with axes in the two groups probably being used for different purposes. A third example of clustering artefacts is that given in Mallory-Greenough and Greenough (1998), who again use single linkage and average linkage clustering on trace-element concentrations determined by inductively coupled plasma mass spectrometry in Ancient Egyptian pottery. They find that three groups of Nile pottery from Mendes and Karnak (Akhenatan Temple Project excavations) can be distinguished using lead, lithium, ytterbium and hafnium data.

An example of the clustering of fossilized material is given in Sutton and Reinhard (1995), who report a cluster analysis of 155 coprolites from Antelope House, a prehistoric Anasazi site in Canyon de Chelly, Arizona. The analysis revealed three primary clusters: whole kernel maize, milled maize, and nonmaize, which the authors interpreted as representing seasonal- and preference-related cuisine.

1.5.6 Bioinformatics and genetics

The past decade has been witness to a tremendous growth in *Bioinformatics,* which is the coming together of molecular biology, computer science, mathematics and statistics. Such growth has been accelerated by the ever-expanding genomic and proteomic databases, which are themselves the result of rapid technological advances in DNA sequencing, gene expression measurement and macromolecular structure determination. Statistics and statisticians have played their most important role in this scientific revolution in the study of gene expression. Genes within each cell's DNA provide the templates for building the proteins necessary for many of the structural and biochemical process that take place in each and every one of us. But although most cells in human beings contain the full complement of genes that

make up the entire human genome, genes are selectively expressed in each cell depending on the type of cell and tissue and general conditions both within and outside the cell. Molecular biology techniques have made it clear that major events in the life of a cell are regulated by factors that alter the expression of the gene. Attempting to understand how expression of genes is selectively controlled is now a major activity in modern biological research. DNA microarrays (Cortese, 2000) are a revolutionary breakthrough in experimental molecular biology that have the ability to simultaneously study thousands of genes under a multitude of conditions and provide a mass of data for the researcher. These new types of data share a common characteristic, namely that the number of variables (p) greatly exceeds the number of observations (n); such data is generally labelled *high dimensional*. Many classical statistical methods cannot be applied to high-dimensional data without substantial modifications. But cluster analysis can be used to identify groups of genes with similar patterns of expression, and this can help provide answers to questions of how gene expression is affected by various diseases and which genes are responsible for specific hereditary diseases. For example, Selinski and Ickstadt (2008) use cluster analysis of single-nucleotide polymorphisms to detect differences between diseased and control individuals in case-control studies, and Eisen *et al*. (1998) use clustering of genome-wide expression data to identify cancer subtypes associated with survival; Witten and Tibshirani (2010) describe a similar application of clustering to renal cell carcinoma data. And Kerr and Churchill (2001) investigate the problem of making statistical inferences from clustering tools applied to gene expression data.

1.6 Summary

Cluster analysis techniques are concerned with exploring data sets to assess whether or not they can be summarized meaningfully in terms of a relatively small number of groups or clusters of objects or individuals which resemble each other and which are different in some respects from individuals in other clusters. A vast variety of clustering methods have been developed over the last four decades or so, and to make discussion of them simpler we have devoted later chapters to describing particular classes of techniques – cluster analysis clustered, so-to-speak! But before looking at these formal methods of cluster analysis, we will, in Chapter 2, examine some graphical approaches which may help in uncovering cluster structure, and then in Chapter 3 consider the measurement of similarity, dissimilarity and distance, which is central to many clustering techniques. Finally, in Chapter 9 we will confront the difficult problem of cluster validation, and try to give potential users of cluster analysis some useful hints as to how to avoid being misled by artefactual solutions.