2 Theoretical Aspects of Pattern Analysis

Arjen van Ooyen
Netherlands Institute for Brain Research, Meibergdreef 33, 1105 AZ Amsterdam, The Netherlands

CONTENTS

2.1 INTRODUCTION TO PATTERN DETECTION

The purpose of most pattern detection methods is to represent the variation in a data set in a more manageable form by recognising classes or groups. The data typically consist of a set of objects described by a number of characters. An object could be (e.g.) a strain of bacteria, while a character could define how well a strain of bacteria grows on a particular C-source, or whether a strain of bacteria contains a particular protein.

If the objects were always described by only two or three characters, there would not be much need for pattern detection methods. Just plotting the data in two or three dimensions, respectively, would be sufficient to distinguish groups (the number of dimensions is the number of axes that are needed in order to plot the data, with one axis for each character). However, typically, objects are characterised by more than three characters, so that simply plotting the data is not possible. Other ways need to be found to represent the data.

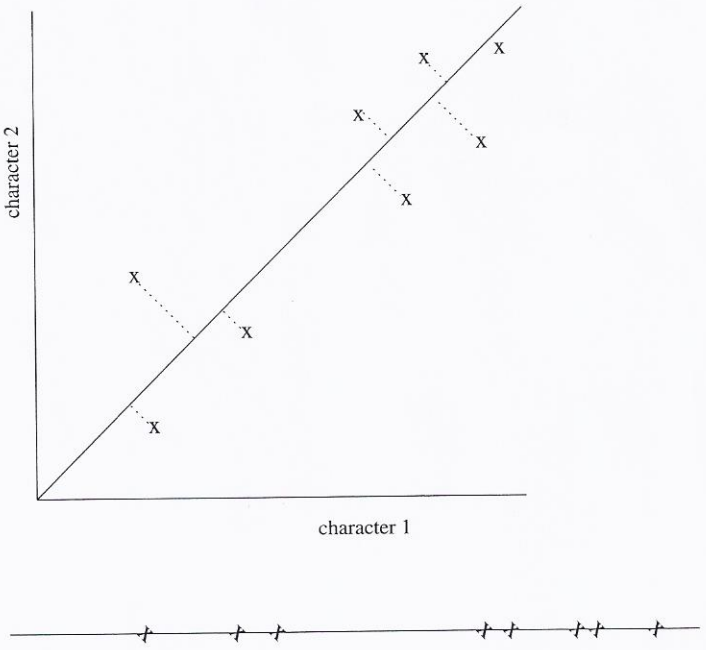

Fig. 2.1. Simple example illustrating principal component analysis (see text).

There are two main approaches that can be taken to manage large data sets. The first involves reducing the number of characters by finding two or three new characters that are combinations of the old characters. Using these new characters, the data can again be plotted in two or three dimensions, and groups can be distinguished by visual inspection. This is the approach taken by *principal component analysis* (see section 2.2). The second approach for managing large data sets does not reduce the number of characters, but involves a stepwise reduction in the number of objects by placing them into groups. This is the approach taken by *cluster analysis* (see section 2.3).

In this chapter, simple examples of both principal component analysis and cluster analysis will be given to explain the ideas behind the methods. Detailed reviews of pattern detection methods and their applications can be found elsewhere (Sokal & Sneath, 1963; Sneath & Sokal, 1973; Bock, 1974; Hogeweg, 1976a; Aldenderfer & Blashfield, 1984; Everitt, 1993; Applied Maths, 1998).

2.2 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis studies large data sets by reducing the number of characters. This is achieved by forming new characters that are combinations of the old ones. A simple example can be used to illustrate the principle behind the method. In the example, the number of characters will be reduced from two to one. In real applications, the method would be used to reduce the number of characters

from "many" to two or three.

In Fig. 2.1, a number of objects characterised by only two characters are plotted. The space spanned by the two axes is called the character space, which in this case is two-dimensional (i.e., has two axes, the x- and y-axis) as there are only two characters. A line then needs to be drawn so that the variance among the points when projected on to this line will be as large as possible (this line is called the first principal component). This ensures that as much information as possible about the original data set will be retained. When this line has been found, all the points are projected on to it. On this line (i.e., the reduced character space), it may be possible to distinguish clusters by visual inspection. This new line, or character, can be interpreted in terms of the contributions that the original characters have made to it.

When principal component analysis is used to reduce the number of characters from "many" to two or three, not only the first but also the second and third principal components are calculated, and the points are projected, not on to a line, but on to a two- or three-dimensional character space.

2.3 CLUSTER ANALYSIS

In contrast to principal component analysis, cluster analysis does not reduce the number of characters, but involves a stepwise reduction in the number of objects by placing them into groups. An agglomerative clustering method starts with as many clusters as there are objects (each cluster thus contains a single object), and then sequentially joins objects (or clusters), on the basis of their similarity, to form new clusters. This process continues until one big cluster is obtained that contains all objects. The result of this process is usually depicted as a dendrogram, in which the sequential union of clusters, together with the similarity value leading to this union, is depicted. A dendrogram, therefore, does not define one partitioning of the data set, but contains many different classifications. A particular classification is obtained by "cutting" the dendrogram at some optimal value (defined relative to the dendrogram). In order to interpret the pattern(s) revealed by the cluster analysis, each pattern is studied to determine its relationship with several characteristics of the objects, including characteristics that were not part of the data set proper, i.e., so-called label information such as epidemic sites of origin of strains, dates of sampling, etc.

To illustrate the clustering process, a simple example will be given in the next section, followed by a general protocol for cluster analysis and a description of different similarity measures and clustering methods.

A. A simple example of cluster analysis

The following example illustrates the whole clustering protocol, from the basic data to the formation of a dendrogram (Fig. 2.2). The data set consists of only four objects, each described by only two characters (Fig. 2.2a). Thus, each object is

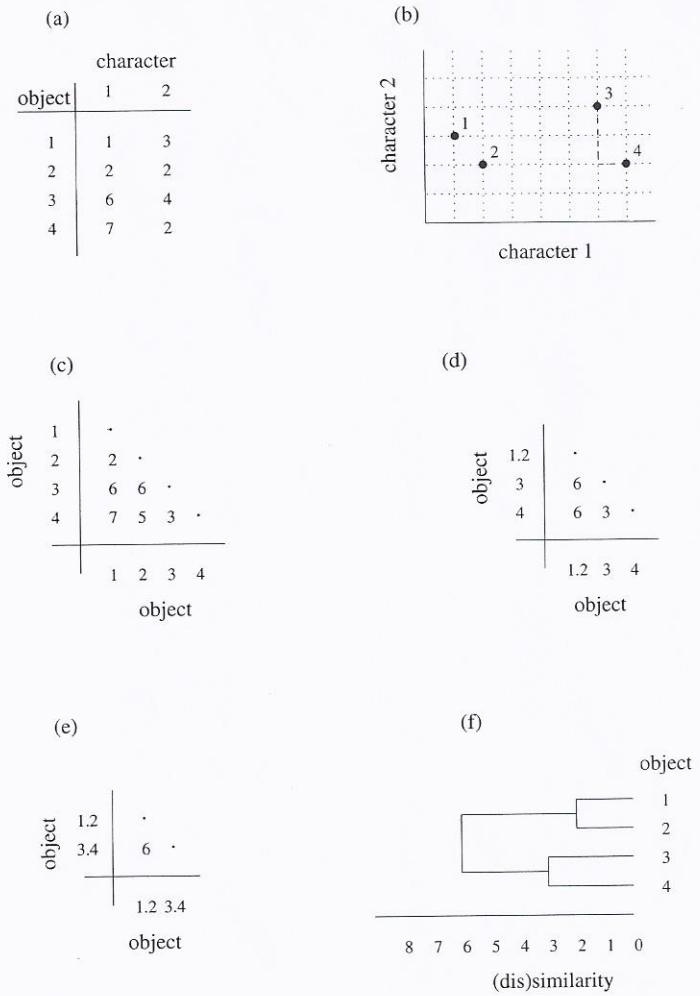

Fig. 2.2. Simple example illustrating the protocol for cluster analysis (see text): (**a**) data set, consisting of four objects, each characterised by two characters; (**b**) objects plotted in character space; (**c**) similarity matrix showing dissimilarity between objects; (**d**) and (**e**) derived similarity matrices used in successive steps of the clustering process; (**f**) dendrogram.

characterised by the values it takes on for these two characters. The objects could be (e.g.) four strains of bacteria, and the characters could (e.g.) describe how well the different strains grow on two different C-sources. Fig. 2.2b shows what the data look like when plotted. The x-coordinate of an object (point) is taken to be the value that the object takes on for character one, and the y-coordinate is the value that the object takes on for character two. As explained earlier, the space spanned by the two axes is called the character space, which in this case is again two-dimensional (i.e., has two axes, the x- and y-axis) as there are only two characters. In general, there are as many dimensions (i.e., axes) as there are different

characters. Plotting objects that are characterised by more than three characters is not possible because it would require more than three axes. Although these data cannot be plotted, they can still be treated mathematically in the same way. The advantage of this simple example is that the data and the clustering process can be easily visualised.

The aim of the clustering procedure is to join the objects (i.e., points in the figure) into clusters, or groups, of similar objects. Two objects will be similar if they are close together in character space. Thus, the first step in any clustering procedure is to determine the similarity between each pair of objects. In order to determine the similarity between two objects, a similarity measure is required. In principle, there are a large number of different measures that can be used. For example, the distance between two objects in character space can be used as a measure of their similarity (or rather dissimilarity). In this example, an even simpler similarity measure will be used. The similarity between, for example, objects 1 and 2, is defined as the difference in the values for the first character plus the difference in the values for the second character. This is what is called city-block distance and can be expressed formally for this example as

$$D_{i,j} = \left|C_{1,i} - C_{1,j}\right| + \left|C_{2,i} - C_{2,j}\right|, \tag{1}$$

where $D_{i,j}$ is the dissimilarity between objects i and j, and $C_{1,i}$ is the value that object i takes on for character 1. The fact that absolute differences are taken is indicated by $|\ldots|$. Using equation (1), the similarity between each pair of objects is determined, which yields a so-called similarity matrix (Fig. 2.2c). This matrix will have a triangular shape because the similarity between, e.g., objects 1 and 2 is the same as the similarity between objects 2 and 1. The clustering of objects starts by joining the objects that are most similar to each other, i.e., that have the lowest value in the similarity matrix. In this case, objects 1 and 2 are most similar to each other, and these will be joined to form the first cluster. The new situation is then a cluster consisting of objects 1 and 2 (which is denoted as cluster {1,2}), and two single objects, 3 and 4. The cluster can then be treated as a new object.

The next step is to calculate a similarity matrix for the new situation. To do this, the similarities between the cluster and the two single objects need to be calculated, i.e., the similarity between object 3 and cluster {1,2}, and the similarity between object 4 and cluster {1,2}. The similarity between objects 3 and 4 is, of course, not changed. In this example, the similarity between object 3 and cluster {1,2} is simply defined as the average of the following two similarities: (a) the similarity between object 3 and object 1, and (b) the similarity between object 3 and object 2. In the same way, the similarity between object 4 and cluster {1,2} can be defined. Thus,

$$D_{3,\{1,2\}} = \frac{D_{3,1} + D_{3,2}}{2}, \tag{2}$$

where $D_{3,\{1,2\}}$ is the similarity between object 3 and cluster {1,2}. Similarly,

$$D_{4,\{1,2\}} = \frac{D_{4,1} + D_{4,2}}{2}, \tag{3}$$

where $D_{4,\{1,2\}}$ is the similarity between object 4 and cluster {1,2}.

There are other ways to define the similarity between single objects and clusters of objects, and the method used to calculate the new similarity is what is called the clustering criterion or clustering method. In the new similarity matrix (Fig. 2.2d), the lowest value is again searched for, which is that between objects 3 and 4, and these objects are subsequently joined. Again a new similarity matrix is calculated, which now consists only of the similarity between cluster {1,2} and cluster {3,4} (Fig. 2.2e). Using the same clustering criterion as before, we obtain

$$D_{\{1,2\},\{3,4\}} = \frac{D_{\{1,2\},3} + D_{\{1,2\},4}}{2}. \tag{4}$$

The similarities $D_{\{1,2\},3}$ and $D_{\{1,2\},4}$ are given by equations (2) and (3), respectively (note that by definition $D_{\{1,2\},3} = D_{3,\{1,2\}}$ and $D_{\{1,2\},4} = D_{4,\{1,2\}}$).

The sequential union of points (groups) is now depicted in a dendrogram (Fig. 2.2f). First, objects 1 and 2 are joined. In the dendrogram, the level at which objects 1 and 2 are connected is the dissimilarity level in the similarity matrix that led to their union. Then, objects 3 and 4 are joined, and finally clusters {1,2} and {3,4}. In the dendrogram, the level at which the clusters are joined is the similarity value as calculated in equation (4); this is a measure for the similarity between *cluster* {1,2} and *cluster* {3,4}. Thus, the similarity between, for example, objects 2 and 4 is not shown in the dendrogram.

B. *General protocol for cluster analysis*

Keeping in mind the previous example, the general procedure for clustering is as follows (Fig. 2.3):

1. *Data set*. The starting point is a data set of objects that are described by the values they take on for a number of characters.
2. *Transformation*. Before calculating a similarity matrix, it may first be necessary to transform the data. This is necessary if the characters are qualitatively different or are expressed in different units. Transformation ensures that equal weight is given to all characters.
3. *Similarity matrix*. The next step is to choose a similarity measure and calculate the similarity between each pair of objects, yielding a triangular similarity matrix. Similarity measures are usually distance measures, but can also be derived from (e.g.) correlation coefficients. For electrophoresis data, the similarity between two objects can be expressed as the correlation between their banding patterns.
4. *Clustering*. Once the clustering method has been chosen – which is basically the formula that defines how to calculate the cluster-to-cluster similarities (and object-to-cluster similarities) from the basic object-to-object similarities – the

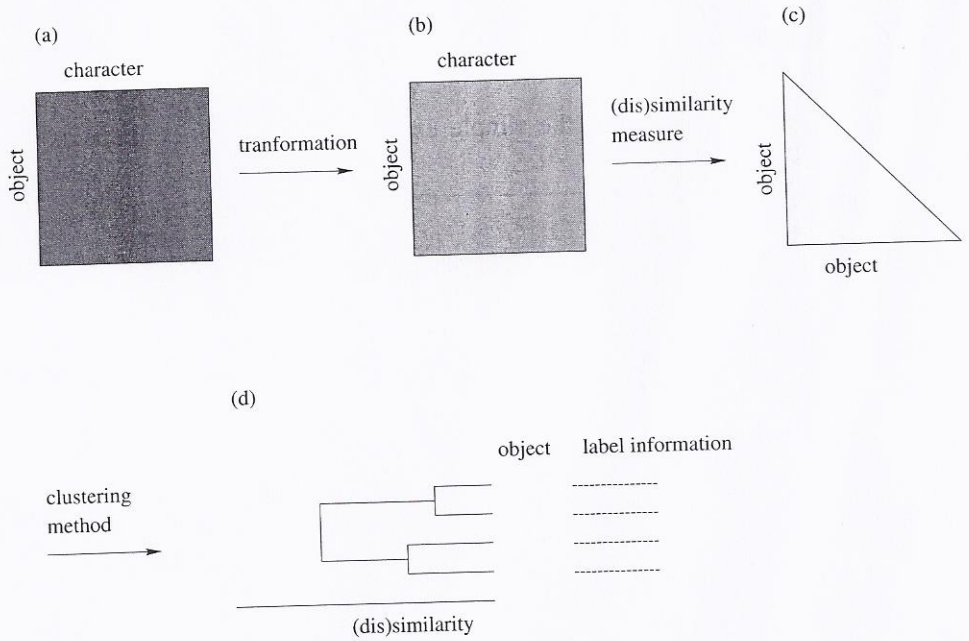

Fig. 2.3. The general protocol for cluster analysis (see text): (**a**) data set; (**b**) data set after transformation; (**c**) similarity matrix; (**d**) dendrogram.

similarity matrix can be used to form clusters.

5. *Dendrogram.* The result of this sequential joining of clusters is depicted in a dendrogram. In a dendrogram, the sequential union of objects and clusters is represented, together with the similarity value leading to this union. A dendrogram, therefore, does not define one partitioning, or grouping, of the set of objects, but contains many different partitionings of the set of objects. A particular partitioning can be obtained by "cutting" the dendrogram at some optimal value, defined relative to the dendrogram. For criteria to determine this cut-off value, see (e.g.) Blanc *et al.* (1994) and Hogeweg (1976b). In interpreting the groupings obtained, so-called label information can play an important role. Label information is basically all the information that is known about the objects which was not actually used in the clustering process itself (i.e., in determining the similarity between objects). Label information includes (e.g.) date of sampling, place of sampling, the date of analysis of the sampling, etc. It may be found – sometimes unexpectedly or unwanted – that the groupings obtained in the cluster analysis correlate with certain label information.

In the next sections, some of the most frequently used similarity measures and clustering methods will be briefly described.

C. *Similarity measures*

(i) *City-block distance*

The similarity measure used in the simple example, the city-block distance (or character difference), is given by

$$D_{i,j} = \sum_{k=1}^{N} \left| C_{k,i} - C_{k,j} \right|, \tag{5}$$

where $D_{i,j}$ is the dissimilarity between objects i and j, N is the total number of characters, and $C_{k,i}$ is the value that object i takes on for character k (index k runs from 1 to N). To calculate the mean city-block distance, the total number of characters is used as the denominator, i.e.,

$$D_{i,j} = \frac{1}{N} \sum_{k=1}^{N} \left| C_{k,i} - C_{k,j} \right|. \tag{6}$$

(ii) Euclidean distance

The distance between two objects in character space is used as a measure of their dissimilarity:

$$D_{i,j} = \sqrt{\sum_{k=1}^{N} (C_{k,i} - C_{k,j})^2}, \tag{7}$$

where $D_{i,j}$ is the distance between objects i and j, and $C_{k,i}$ is the value that object i takes on for character k (that $D_{i,j}$ represents distance can easily be seen for $N = 2$, using the Pythagorean theorem). To avoid the use of the square root, the value of the distance is often squared, and this expression is referred to as "squared Euclidean distance".

In comparing electrophoresis patterns, the matrix of similarities can be based either on the Pearson correlation coefficient or on one of the band-matching coefficients (Applied Maths, 1998).

(iii) Pearson or product-moment correlation coefficient

The similarity between two objects is calculated as the correlation between the two arrays of character values (typically densitometric values) taken on by the two objects:

$$S_{i,j} = \frac{\sum_{k=1}^{N} (C_{k,i} - \overline{C_i})(C_{k,j} - \overline{C_j})}{\sqrt{\sum_{k=1}^{N} (C_{k,i} - \overline{C_i})^2 \sum_{k=1}^{N} (C_{k,j} - \overline{C_j})^2}}, \tag{8}$$

where $S_{i,j}$ is the similarity (i.e., correlation coefficient) between objects i and j, $C_{k,i}$ is the value that object i takes on for character k, and $\overline{C_i}$ is the mean of all the character values of object i. The value of the correlation coefficient ranges from

+1 for perfect association to –1 for negative association; a value of 0 indicates that there is no association. That a correlation of 1 means perfect association can be seen by correlating object i to itself, i.e.,

$$S_{i,i} = \frac{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)(C_{k,i} - \overline{C}_i)}{\sqrt{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)^2 \sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)^2}} = \frac{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)^2}{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)^2} = 1. \quad (9)$$

The correlation coefficient is a shape measure; i.e., it is sensitive to the pattern of dips and rises across the character values. Two profiles can have a correlation of +1 and yet not be truly identical (i.e., take on the same values). This occurs, for example, when the two profiles have the same pattern of dips and rises, but one profile is elevated compared to the other (see also Chapter 3).

(iv) Band-based similarity coefficients

(a) Coefficient of Jaccard. The similarity between two tracks of bands is the number of matching bands divided by the total number of bands in both tracks (i.e., the corresponding bands plus the track-specific bands):

$$S_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}, \quad (10)$$

where $S_{i,j}$ is the similarity between tracks i and j, $n_{i,j}$ is the number of corresponding bands for i and j, n_i is the total number of bands in i, and n_j is the total number of bands in j. So $n_i + n_j - n_{i,j}$ is the total number of bands in both tracks, not double counting the corresponding ones. If all bands in i match those in j, then $S_{i,j} = 1$.

(b) Area-sensitive coefficient. This is a more sophisticated similarity measure, which also takes into account the possible differences in areas of the matching bands:

$$S_{i,j} = \frac{A_{i,j}}{n_i + n_j - n_{i,j}}, \quad (11)$$

where

$$A_{i,j} = \sum_{k=1}^{n_{i,j}} \frac{\alpha}{\alpha + \left| B_{i,k} - B_{j,k} \right|}, \quad (12)$$

where α is a constant, and $\left| B_{i,k} - B_{j,k} \right|$ is the absolute difference between the areas of the k-th corresponding band in i and j, where k runs from 1 to $n_{i,j}$. Thus, differences in band areas of the corresponding bands are penalised. If the areas of all corresponding bands of both tracks are equal, this coefficient is reduced to the coefficient of Jaccard: if $B_{i,k} = B_{j,k}$ for all k, $A_{i,j} = \sum_{k=1}^{n_{i,j}} 1 = n_{i,j}$.

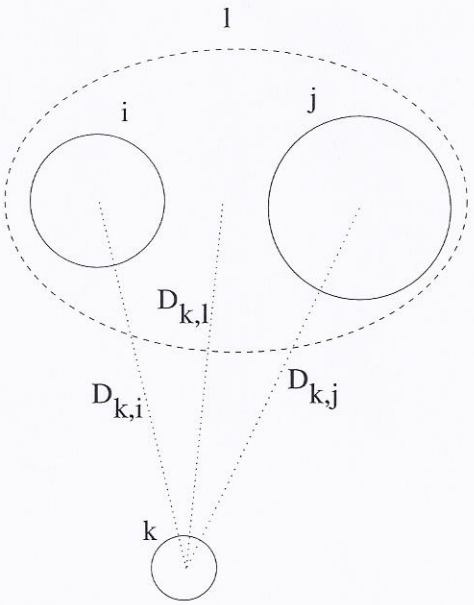

Fig. 2.4. UPGMA or group average (see text). The dissimilarity between an object or cluster *k*, and a cluster *l* formed by joining objects or clusters *i* and *j*, is the average of the dissimilarities between *k* and *i*, and between *k* and *j*, weighted for the number of points in clusters *i* and *j*.

(c) Dice coefficient. The Dice coefficient is very similar to the coefficient of Jaccard, but gives more weight to matching bands:

$$S_{i,j} = \frac{2n_{i,j}}{n_i + n_j}, \tag{13}$$

where $S_{i,j}$ is the similarity between tracks i and j, $n_{i,j}$ is the number of matching bands for i and j, n_i is the total number of bands in i, and n_j is the total number of bands in j.

D. *Clustering methods*

(i) *UPGMA or group average*

This similarity measure, termed the unweighted pair group method using arithmetic averages (UPGMA), was used in the simple example discussed earlier in this chapter. It states that the dissimilarity between an object or cluster *k*, and a cluster *l* formed by joining objects or clusters *i* and *j*, is simply the average of the dissimilarities between *k* and *i*, and between *k* and *j* (taking into account the number of points in clusters *i* and *j*) (Fig. 2.4). This is given by the formula

$$D_{k,l} = \frac{N_i D_{k,i} + N_j D_{k,j}}{N_i + N_j}, \tag{14}$$

where k is the index used for an existing cluster or object, l is the index used for

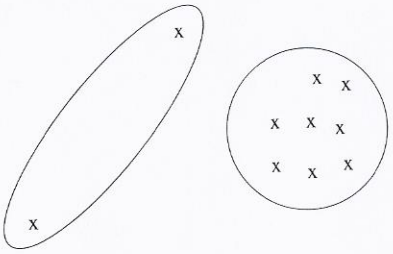

Fig. 2.5. With Ward's clustering method, a cluster of aberrant points (in this example, the cluster with two points) is often found which have nothing in common with each other except that they are dissimilar to the other objects.

the newly formed cluster, $D_{k,l}$ is the dissimilarity between k and l, N_i is the number of objects in cluster i, and N_j is the number of objects in cluster j. This clustering method effectively leads to minimisation of the average dissimilarity between the objects in a cluster. This interpretation holds for all types of similarity measures. The clustering structure is less pronounced and the clusters are more limited in diameter than with Ward's clustering method (see below).

(ii) Ward's averaging

With Ward's averaging, those clusters (objects) are joined which lead to a minimal increase in the total within group variance. This results in the following properties of the method: (a) a cluster of aberrant points is often found which have nothing in common with each other, except that they are dissimilar to the other objects (Fig. 2.5); (b) more groups are distinguished in dense areas of the character space (i.e., where most of the objects are); and (c) every data set shows a clear cluster structure, which does not necessarily imply that there are clear separations.

2.4 EXAMPLES OF APPLICATIONS OF CLUSTER ANALYSIS

Among the many possible areas of applications, pattern detection techniques are now widely used in both taxonomy and epidemiology. In taxonomy, the objective is to classify organisms into genera and species on the basis of their genotypic *or* phenotypic relationships (i.e., taxonomy is not necessarily limited to identifying relationships by ancestry); in epidemiology, the objective is confined to identifying bacterial isolates in terms of their recent ancestry (i.e., their epidemiological origin). Many examples of both applications can be found throughout this book. In this chapter, just three examples are given to illustrate the various goals of cluster analysis.

The first example (Coenye *et al.*, 2000) shows how cluster analysis used on different types of data, in combination with the evaluation of the groups obtained in terms of label and other information, can help to unravel the taxonomy of microorganisms. A polyphasic taxonomic study was performed on a group of isolates identified tentatively as *Burkholderia cepacia*, a bacterial pathogen that causes

life-threatening lung infections in cystic fibrosis patients. Using cluster analysis with the Pearson or product-moment correlation coefficient as the similarity measure, and UPGMA as the clustering method, analysis of SDS-PAGE fingerprints of whole-cell proteins (see Chapter 4) and AFLP fingerprints (see Chapter 8) identified at least five different species, and this was confirmed by DNA-DNA hybridisation experiments. Based on genotypic and phenotypic characteristics, these organisms were then classified in a novel genus, *Pandoraea*.

The second example (Sloos *et al.*, 1998) demonstrates the application of cluster analysis to microbial epidemiology. The diversity of strains of *Staphylococcus epidermidis* in a neonatal care unit of a secondary care hospital in The Netherlands was studied. Samples were taken consecutively from patients, and the isolates obtained were typed by pulsed-field gel electrophoresis (PFGE; see Chapter 7) and quantitative antibiogram analysis. The antibiograms were used to group the organisms (Fig. 2.6), using squared Euclidean distance as the similarity measure and Ward's averaging as the clustering method. The main grouping obtained was evaluated for its correlation with other characteristics of the individual isolates, including PFGE type, length of stay, usage of antibiotics, birth weight and cubicle number. Thus, these characteristics of the isolates were not used in the generation of clusters, but were used as label information to help interpret the grouping. The cluster analysis revealed that 14 isolates from six patients had a common PFGE pattern and were of one multiresistant antibiogram type. The remaining isolates belonged to a variety of PFGE types and were more susceptible to antibiotics. Colonisation with the multiresistant strain correlated with a long period of stay and with the use of specific antibiotics. Cluster analysis on the basis of antibiograms was also performed on a combined collection that included multiresistant strains from another hospital in the same area. This analysis revealed that the multiresistant strains from both hospitals were closely related, and suggested that transfer of the multiresistant strain had occurred between hospitals.

In the third example (Blanc *et al.*, 1996), cluster analysis of quantitative antibiograms was performed to test whether a typology based on antibiograms would correspond to typologies based on other characteristics. It was found that the grouping obtained by cluster analysis of antibiograms was equivalent to the grouping obtained by ribotyping (see Chapter 5) when the ribotyping was used as label information to evaluate the clusters.

2.5 DISCUSSION

Cluster analysis is a procedure that starts with a data set containing information about a set of objects, and then attempts to organise these objects into groups that are in some sense optimal for the data set under consideration. Cluster analysis can be used for a variety of goals (Aldenderfer & Blashfield, 1984), including developing typologies or classifications, generating concepts or hypotheses through data exploration, and testing whether typologies or classifications generated by other procedures, or by using other data, are present in the data set under consideration.

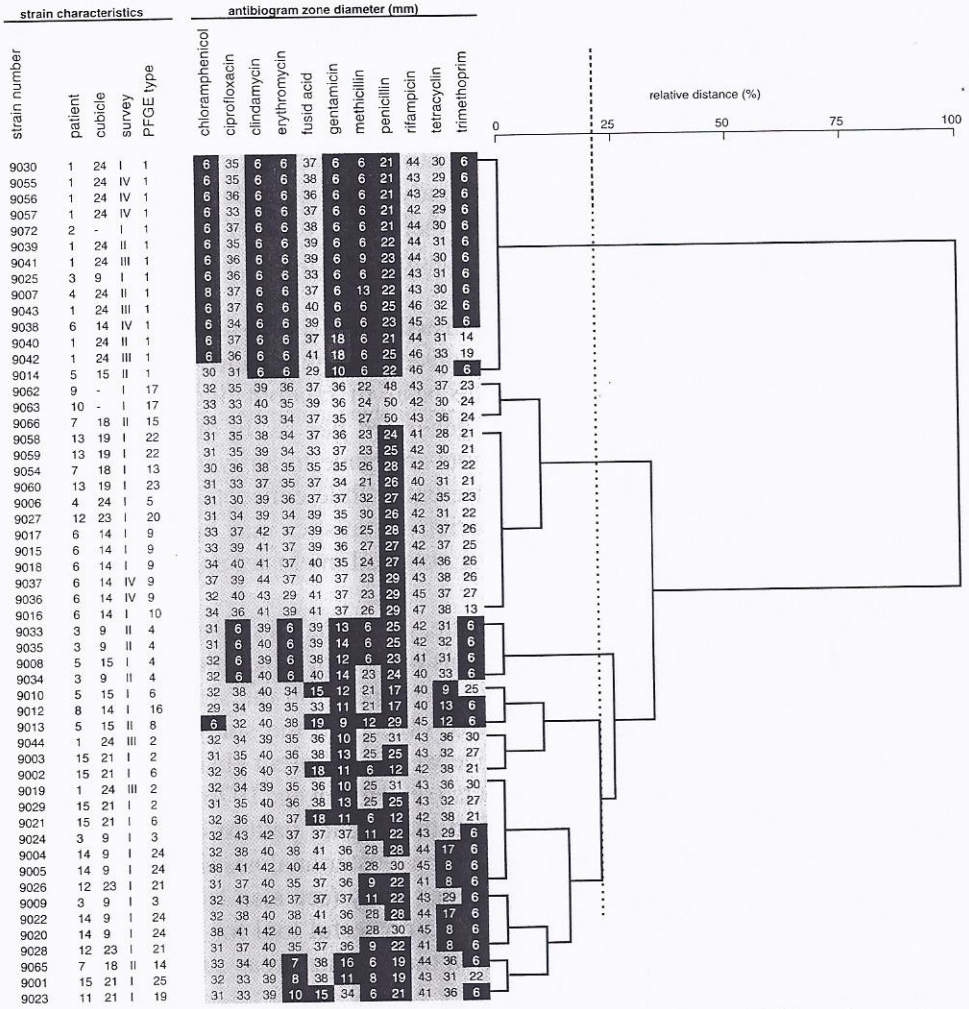

Fig. 2.6. Strain characteristics (left), antibiogram susceptibility profiles (middle), and grouping of 53 *Staphylococcus epidermidis* isolates of neonates on the basis of zone diameters (right). Squared Euclidian distance was calculated between all possible pairs of zones, and grouping was performed using Ward's method. The dotted line denotes the distance at which four clusters are delineated. The inhibition zones were used for classifying isolates into "susceptible" (green), "intermediate resistant" (blue), or "resistant" (red) categories for each antibiotic, using the standard Dutch criteria for susceptibility determination. Taken from Sloos *et al.* (1998).

These goals are illustrated, respectively, by the studies of Coenye *et al.* (2000), Sloos *et al.* (1998) and Blanc *et al.* (1996), as described above.

Although pattern detection is sometimes regarded as yet another form of statistics, there are important conceptual differences (Hogeweg, 1976a):

1. In statistics, deviations from randomness in the data set are looked for, while in pattern detection the structure in the data set is sought. Note that a random data set can also have structure!

2. In statistics, attempts are made to make sample-independent statements. The data under consideration are assumed to be a random sample of the whole population, and the objective is to make statements about the whole population by looking at a representative sample of the population. Ideally, these statements should not change if a different random sample is taken from the population. In pattern detection, the data set under study is not considered a sample from a larger population but is considered all there is. A different structure may be found if new data is added (e.g., in taxonomy when new species are discovered).
3. In statistics, groups (and an underlying distribution) are pre-supposed and tests are made to determine whether these groups differ significantly from each other (i.e., more than can be expected on the basis of random fluctuations alone), while in pattern detection, groups are generated *per se*. In other words, concepts are *tested* in statistics (i.e., attempts are made to answer the question as to whether pre-supposed groups are different), while concepts (i.e., groupings) are *generated* in pattern detection. Descriptive statistics may be used in pattern detection for characterising the grouping obtained in cluster analysis.

Cluster analysis can best be seen as a heuristic, rather than a statistical, method for exploring the diversity in a data set by means of pattern generation. The result of a cluster analysis study can, and usually does, depend on the similarity measure used, the clustering method used, the set of objects in the study, the characters used to describe the objects, and the relative weight different characters are given in calculating the similarity between objects (see Hogeweg, 1976b; Van Ooyen & Hogeweg, 1990). Rather than trying to find the "right" pattern or classification, the differences in the patterns as revealed by the cluster analysis should be used to gain further understanding of the objects under study (see also Hogeweg, 1976a). Used in this heuristic way, cluster analysis is a powerful tool for data exploration in taxonomy and epidemiology, as well as in many other areas such as functional genomics.

REFERENCES

Aldenderfer, M.S. & Blashfield, R.K. (1984). *Cluster analysis*. Sage Publications, Newbury Park.

Applied Maths (1998). *GelCompar (comparative analysis of electrophoresis patterns) reference manual, version 4.1*. Applied Maths, Kortrijk.

Blanc, D.S., Lugeon, C., Wenger, A., Siegrist, H.H. & Francioli, P. (1994). Quantitative antibiogram typing using inhibition zone diameters compared with ribotyping for epidemiological typing of methicillin-resistant *Staphylococcus aureus*. *Journal of Clinical Microbiology* **32**, 2505–2509.

Blanc, D.S., Petignat, C., Moreillon, P., Wenger, A., Bille, J. & Francioli, P. (1996). Quantitative antibiogram as a typing method for the prospective epidemiological surveillance and control of MRSA: comparison with molecular typing. *Infection Control and Hospital Epidemiology* **17**, 654–659.

Bock, H.H. (1974). *Automatische klassifikation*. Vandenhoeck & Ruprecht, Gotingen.

Coenye, T., Falsen, E., Hoste, B., Ohlen, M., Goris, J., Govan, J.R.W., Gillis, M. & Vandamme, P. (2000). Description of *Pandoraea* gen. nov. with *Pandoraea apista* sp. nov., *Pandoraea pulmo-*

nicola sp. nov., *Pandoraea pnomenusa* sp. nov., *Pandoraea sputorum* sp. nov. and *Pandoraea norimbergensis* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **50**, 887–899.

Everitt, B. (1993). *Cluster analysis*, 3rd edn. Arnold, London.

Hogeweg, P. (1976a). *Topics in biological pattern analysis*. PhD Thesis, University of Utrecht.

Hogeweg, P. (1976b). Iterative character weighing in numerical taxonomy. *Computers in Biology and Medicine* **6**, 199–211.

Sloos, J.H., Horrevorts, A.M., Van Boven, C.P.A. & Dijkshoorn, L. (1998). Identification of multiresistant *Staphylococcus epidermidis* in neonates of a secondary care hospital using pulsed field gel electrophoresis and quantitative antibiogram typing. *Journal of Clinical Pathology* **51**, 62–67.

Sneath, P.H.A. &. Sokal, R.R (1973). *Numerical taxonomy*. Freeman, San Francisco.

Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of numerical taxonomy*. Freeman, San Francisco.

Van Ooyen, A. & Hogeweg, P. (1990). Iterative character weighting based on mutation frequency: a new method for constructing phyletic trees. *Journal of Molecular Evolution* **31**, 330–342.