Chapter 8
Cluster and Discriminant Analysis

8.1 Introduction

Under multivariate analysis, two very important techniques are clustering and classification. Under the problem of clustering, we try to find out the unknown number of homogeneous inherent groups in a data set as well as the structure of the groups. But under classification, the basic problem is discrimination of objects into some known groups. One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. Classification is fundamental to most branches of science.

Cluster analysis has a variety of objectives. It is focussed on segmenting a collection of items (also called observations, individuals, cases, or data rows) into subsets such that those within each cluster are more closely related to one another than objects assigned to different clusters. The main focus in cluster analysis is on the notion of degree of similarity (or dissimilarity) among the individual objects being clustered. The two major methods of clustering are hierarchical clustering and k-means clustering. Most of the clustering methods are exploratory in nature and do not need any model assumption.

Different statistical techniques are available for clustering and classification (Fraix-Burnet et al. 2015; De et al. 2013 and references there in). But depending on the nature of the different types of data, several problems often arise and in some cases a proper solution is still not available.

Sometimes the data set under consideration has a distributional form (usually normal), and sometimes it is of non-normal nature. Based on the above point, there is a justification needed about which clustering or classification technique should be used so that it reflects the proper nature of the data set provided. This problem is more relevant for classification as most of the classification methods are model

Sections of this chapter draw from one of the authors' published work, 'Statistical Methods for Astronomical Data Analysis,' authored by Asis Kumar Chattopadhyay and Tanuka Chattopadhyay, and published in 2014 by Springer Science+Business Media, New York.

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_8

based. For clustering, most of the methods are nonparametric in nature and as such the above problem is not very serious. But here also basic assumption is that the nature of the variables under study is continuous, whereas under practical situations, these may be categorical like binary, nominal, ordinal, and even directional (particularly for environmental and astronomical data). Under such situations, standard similarity/dissimilarity measures will not work.

The clustering techniques which require an inherent model assumption are known as model-based methods, whereas the clustering technique where no modeling assumption or distributional form is needed may be termed as non-model-based methods. Hence based on the nature of data set, one has to decide about proper application of the two types of techniques.

At present, big data issues related to data size are quite common. In statistical terms, these problems may be tackled in terms of both the number of observations and the variables considered. Many standard clustering techniques fail to deal with such big data sets. Thus, some dimension reduction methods may be applied at first and then clustering may be performed on the reduced data set. Some data mining techniques are very helpful under such situations.

Finally and most importantly, after all these considerations, the similarity of grouping of objects obtained from different methods should be checked in terms of some physical properties.

8.2 Hierarchical Clustering Technique

There are two major methods of clustering, viz. hierarchical clustering and k-means clustering. In hierarchical clustering, the items are not partitioned into clusters in a single step. Instead, a series of partitions takes place, which runs from a single cluster containing all objects to n clusters each containing a single object. Hierarchical clustering is subdivided into agglomerative methods, which proceed by series of combinations of the n objects into groups, and divisive methods, which separate n objects successively into smaller groups. Agglomerative techniques are more commonly used. Hierarchical clustering may be represented by a two-dimensional diagram known as dendrogram which illustrates the additions or divisions made at each successive stage of analysis.

8.2.1 Agglomerative Methods

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, G_n; G_{n-1}; :::::::; G_1. The first G_n consists of n single-object 'clusters,' and the last G_1 consists of single group containing all n cases. The structure of the groups is not unique and depends on several factors like choice of the dissimilarity/similarity measure, choice of the linkage measure.

At each particular stage, the method adds together the two clusters which are most similar. At the first stage, we join together two objects that are closest together, since at the initial stage each cluster has only one object. Differences between methods arise because of the different ways of defining dissimilarity or similarity between clusters.

Hierarchical clustering is largely dependent on the selection of such a measure. A simple measure is Manhattan distance, equal to the sum of absolute distances for each variable. The name comes from the fact that in a two-variable case, the variables can be plotted on a grid that can be compared to city streets, and the distance between two points is the number of blocks a person would walk.

The most popular measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle. Besides Manhattan and Euclidian distances, there are other dissimilarity measures also based on the correlation coefficients between two observations on the basis of several variables.

Alternatively, one may use a similarity measure which is complementary in nature and under the normalized set up, it may be obtained by subtracting the dissimilarity measure from one.

8.2.2 Similarity for Any Type of Data

The above-mentioned dissimilarity/similarity measures are applicable to continuous-type data only. But generally, we work with mixed-type data sets those include different types like continuous, discrete, binary, nominal, ordinal. Gower (1971) has proposed a general measure as follows:

The Gower's Coefficient of Similarity:

Two individuals i and j may be compared on a character k and assigned a score $sijk$. There are many ways of calculating $sijk$, some of which are described below.

Corresponding to n individuals and p variables, Gower's similarity index S_{ij} is defined as

$$S_{ij} = \Sigma_{k=1}^{p} s_{ijk} / \Sigma_{k=1}^{p} \delta_{ijk} (i, j = 1, 2, \ldots n)$$

$$\begin{aligned} \text{where } \delta_{ijk} &= 1 \text{ when character } k \text{ can be compared} \\ &\quad \text{for observations } i \text{ and } j \\ &= 0 \text{ otherwise} \end{aligned}$$

For continuous (quantitative) variables with values $x_{1k}, x_{2k}, \ldots, x_{nk}$ for the kth variable

$$s_{ijk} = 1- \mid x_{ik} - x_{jk} \mid / R_k$$

where R_k is the range of the variable k and may be the total range in population or the range in the sample.

For a categorical (qualitative) character with m categories ($m = 2$ for binary variable)

$$\begin{aligned} s_{ijk} &= 0 \text{ if } i \text{ and } j \text{ are totally different} \\ &= q \text{ (positive fraction) if there is some degree of agreement} \\ &= 1 \text{ when } i \text{ and } j \text{ are same} \end{aligned}$$

8.2.3 *Linkage Measures*

To calculate distance between two clusters, it is required to define two representative points from the two clusters (Chattopadhyay and Chattopadhyay 2014). Different methods have been proposed for this purpose. Some of them are listed below.[1]

Single linkage: One of the simplest methods is single linkage, also known as the nearest neighbor technique. The defining feature of the method is that distance between clusters is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each cluster are considered.

In the single linkage method, d_{rs} is computed as $d_{rs} = \text{Min}\, d_{ij}$, where object i is in cluster r and object j is in cluster s and d_{ij} is the distance between the objects I and j. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s. The minimum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the shortest link between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Complete linkage: The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between clusters is now defined as the distance between the most distant pair of objects, one from each cluster. In the complete linkage method, $d - rs$ is computed as $d_{rs} = \text{Max}\, d_{ij}$, where object i is in cluster r and object j is cluster s. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s and the maximum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the largest distance between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Average linkage: Here the distance between two clusters is defined as the average of distances between all pairs of observations, where each pair is composed of one object from each group. In the average linkage method, d_{rs} is computed as

[1] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

$d_{rs} = Trs/(Nr \times Ns)$ where Trs is the sum of all pair-wise distances between cluster r and cluster s. Nr and Ns are the sizes of the clusters r and s, respectively. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is the minimum, are merged.

Minimax Linkage: This was introduced by Bien and Tibshirani (2011). For any point x and cluster G, define

$$d_{\max}(x, G) = \max_{y \in G} d(x, y)$$

as the distance to the farthest point in G from x. Define the minimax radius of the cluster G as

$$r(G) = \min_{x \in G} d_{\max}\,(x, G)$$

that is, find the point $x \in G$ from which all points in G are as close as possible. This minimizing point is called the prototype for G. It may be noted that a closed ball of radius $r(G)$ centered at the prototype covers all of G. Finally, we define the minimax linkage between two clusters G and H as

$$d(G, H) = r(GUH)$$

that is, we measure the distance between clusters G and H by the minimax radius of the resulting merged cluster.

8.2.4 Optimum Number of Clusters

Usually, the number of clusters is determined from the dendrogram and validated by the physical properties. We specify a horizontal line for a particular similarity/dissimilarity value, and the clusters below this line are selected as optimum. But some mathematical rules (thumb rules) are also available which are based on between cluster and within cluster sum of squares values. If we denote by k, the number of clusters and define by W(k) the sum of the within cluster sum of squares for k clusters then the values of W(k) will gradually decrease with increase in k and that ‘k’ may be taken as optimum where W(k) stabilizes. For detailed discussion, one may follow the link http://www.cc.gatech.edu/~hpark/papers/cluster_JOGO.pdf.

8.2.5 Clustering of Variables

The hierarchical clustering method can also be used for clustering of variables on the basis of the observations. Here instead of the distance matrix, one may start with the correlation matrix (higher correlation indicating similarity of variables).

The linkage measures as listed in the previous section will not be applicable for variable clustering. In order to measure similarity/dissimilarity between two clusters of variables, one may either use the correlation between first principal components corresponding to the two clusters or the canonical correlations.

8.3 Partitioning Clustering-k-Means Method

The k-means algorithm (MacQueen 1967) assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. This method can be used for clustering of objects and not variables.

This method starts with a value of k. We will discuss later the method of selection of the value of k. Then we randomly generate k clusters and determine the cluster centers, or directly generate k seed points as cluster centers. Assign each point to the nearest cluster center in terms of Euclidian distance. Re-compute the new cluster centers. Repeat until some convergence criterion is met, i.e., there is no reassignment. The main advantages of this algorithm are its simplicity and speed which allows it to run on large data sets. Its disadvantage is that it is highly dependent on the initial choice of clusters. It does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It maximizes inter-cluster variance and minimizes intra-cluster variance.

The advantages of partitioning method are as follows (Chattopadhyay and Chattopadhyay 2014):

(a) A partitioning method tries to select best clustering with k groups which is not the goal of hierarchical method.
(b) A hierarchical method can never repair what was done in previous steps.
(c) Partitioning methods are designed to group items rather than variables into a collection of k clusters.
(d) Since a matrix of distances (similarities) does not have to be determined and the basic data do not have to be stored during the computer run, partitioning methods can be applied to much larger data sets.

For k-means algorithms, the optimum value of k can be obtained in different ways.

On the basis of the method proposed by Sugar and James (2003), by using k-means algorithm first determine the structures of clusters for varying number of clusters taking $k = 2, 3, 4$, etc. For each such cluster formation, compute the values of a distance measure

$$dK = (1/p)\min_x E[(x_k - c_k)'(x_k - c_k)]$$

which is defined as the distance of the x_k vector (values of the parameters) from the center c_k (which is estimated as mean value), p is the order of the x_k vector.

Then the algorithm for determining the optimum number of clusters is as follows. Let us denote by d_k' the estimate of d_k at the kth point which is actually the sum of within cluster sum of squares over all k clusters. Then d_k' is the minimum achievable distortion associated with fitting k centers to the data. A natural way of choosing the number of clusters is plot d_k' versus k and look for the resulting distortion curve. This curve is always monotonic decreasing. Initially, one would expect much smaller drops, i.e., a leveling off for k greater than the true number of clusters because past this point adding more centers simply partitions within groups rather than between groups.

According to Sugar and James (2003) for a large number of items the distortion curve when transformed to an appropriate negative power, will exhibit a sharp "jump" (if we plot k versus transformed d_k'). Then calculate the jumps in the transformed distortion as

$$J_k = (d_k'^{-(p/2)} - d_{k-1}'^{-(p/2)})$$

Another way of choosing the number of clusters is plot J_k versus k and look for the resulting jump curve. The optimum number of clusters is the value of k at which the distortion curve levels off as well as its value associated with the largest jump.

The k-means clustering technique depends on the choice of initial cluster centers (Chattopadhyay et al. 2012). But this effect can be minimized if one chooses the cluster centers through group average method (Milligan 1980). As a result, the formation of the final groups will not depend heavily on the initial choice and hence will remain almost the same according to physical properties irrespective of initial centers. In MINITAB package, the k-means method is almost free from the effect of initial choice of centers as they have used the group average method.

8.4 Classification and Discrimination

Discriminant[2] analysis and classification are multivariate techniques concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. Once the optimum clustering is obtained by applying the method discussed under previous section, one can verify the acceptability of the classification by computing classification/misclassification probabilities for the different objects. Although the k-means clustering method is purely a data analytic method, for classification it may be necessary to assume that the underlying distribution is multivariate normal. The method can be illustrated as follows for two populations (clusters). The method can be easily generalized for more than two underlying populations.

[2] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical Methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

Let $f_1(x)$ and $f_2(x)$ be the probability density functions associated with the $p \times 1$ random vector X for the populations π_1 and π_2 respectively. Let Ω be the sample space, i.e., collection of all objects. Let us denote by x the observed value of X. Let R1 be that set of x values for which we classify objects as π_1 and $R_2 = \Omega R_1$ be the remaining x values for which we classify objects as π_2. Since every object must be assigned to one and only one of the two groups, the sets R_1 and R_2 are disjoint and exhaustive. The conditional probability of classifying an object as π_2 when in fact it is from π_1 (error probability) is,

$$P(2 \mid 1) = P[X \in R_2 \mid \pi_1] = f_{R2} f_1(x) dx$$

Similarly, the other error probability can be defined. Let p_1 and p_2 be the prior probabilities of π_1 and π_1, respectively, $(p_1 + p_2 = 1)$. Then the overall probabilities of correctly and incorrectly classifying objects can be derived as

P (correctly classified as π_1) = P (Observation actually comes from π_1 and is correctly classified as π_1) $= P[X \in R_1 \mid \pi_2] p_2$.
P (misclassified as π_1) $= P[X \in R_1 \mid \pi_2] p_2$.

The associated cost of misclassification can be defined by a cost matrix

	Classified as	
True population	π_1	π_2
π_1	0	$C(2 \mid 1)$
π_2	$C(1 \mid 2)$	0

For any rule, the average or Expected Cost of Misclassification (ECM) is given by

$$ECM = C(2 \mid 1)P(2 \mid 1)p_1 + C(1 \mid 2P(1 \mid 2)p_2$$

A reasonable classification rule should have ECM as small as possible.

Rule: The regions R_1 and R_2 that minimize the ECM are defined by the value of x for which the following inequalities hold.

$$R_1 : \frac{f_1(x)}{f_2(x)} > \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

If we assume $f_1(x)$ and $f_2(x)$ are multivariate normal with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2, respectively, then a particular object with observation vector x_0 may be classified according to the following rule (under the assumption $\Sigma_1 = \Sigma_2$)

Allocate x_0 to π_1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

allocate x_0 to π_2 otherwise.

If we choose $C(1 \mid 2) = C(2 \mid 1)$ and $p_1 = p_2$, then the estimated minimum ECM rule for two Normal populations will be as follows:

Allocate x_0 to π_1 if

$$(m_1 - m_2)'S_{\text{pooled}} - 1x_0 - \frac{1}{2}(m_1 - m_2)'\Sigma^{-1}(m_1 + m_2) \geq 0$$

where m_1 and m_2 are sample mean vectors of the two populations and S_{pooled} is pooled (combined) sample covariance matrix. Allocate x_0 to π_2 otherwise. The LHS is known as the linear discriminant function. One can easily generalize the method for more than two groups.

8.5 Data

Example 8.5.1 The Fisher's *Iris* data set is a multivariate data set introduced by Fisher (1936). It is also known as Anderson's *Iris* data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris setosa (type-3), Iris versicolor (type-2), and Iris virginica (type-1)). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters (Table 8.1).

We have performed k-means clustering of the data on the basis of the first four variables, viz. sepal length, sepal width, petal length, and petal width. Choosing $k = 3$, we have divided the 150 observations into three groups in order to verify whether we can identify three groups corresponding to three species. From columns 6 and 7, it is clear that k-means method has correctly identified Iris setosa (type-3) species for all the 50 cases, whereas there are some errors corresponding to types 1 and 2. For type 2, three cases and for type 1 fourteen cases had wrongly identified. The summary result for k-means clustering is given below:

Table 8.1 Results of k-means clustering for Iris data

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.1	3.5	1.4	0.2	I. setosa	3	3
4.9	3	1.4	0.2	I. setosa	3	3
4.7	3.2	1.3	0.2	I. setosa	3	3
4.6	3.1	1.5	0.2	I. setosa	3	3
5	3.6	1.4	0.2	I. setosa	3	3
5.4	3.9	1.7	0.4	I. setosa	3	3
4.6	3.4	1.4	0.3	I. setosa	3	3
5	3.4	1.5	0.2	I. setosa	3	3
4.4	2.9	1.4	0.2	I. setosa	3	3
4.9	3.1	1.5	0.1	I. setosa	3	3
5.4	3.7	1.5	0.2	I. setosa	3	3
4.8	3.4	1.6	0.2	I. setosa	3	3
4.8	3	1.4	0.1	I. setosa	3	3
4.3	3	1.1	0.1	I. setosa	3	3
5.8	4	1.2	0.2	I. setosa	3	3
5.7	4.4	1.5	0.4	I. setosa	3	3
5.4	3.9	1.3	0.4	I. setosa	3	3
5.1	3.5	1.4	0.3	I. setosa	3	3
5.7	3.8	1.7	0.3	I. setosa	3	3
5.1	3.8	1.5	0.3	I. setosa	3	3
5.4	3.4	1.7	0.2	I. setosa	3	3
5.1	3.7	1.5	0.4	I. setosa	3	3
4.6	3.6	1	0.2	I. setosa	3	3
5.1	3.3	1.7	0.5	I. setosa	3	3
4.8	3.4	1.9	0.2	I. setosa	3	3
5	3	1.6	0.2	I. setosa	3	3
5	3.4	1.6	0.4	I. setosa	3	3
5.2	3.5	1.5	0.2	I. setosa	3	3
5.2	3.4	1.4	0.2	I. setosa	3	3
4.7	3.2	1.6	0.2	I. setosa	3	3
4.8	3.1	1.6	0.2	I. setosa	3	3
5.4	3.4	1.5	0.4	I. setosa	3	3
5.2	4.1	1.5	0.1	I. setosa	3	3
5.5	4.2	1.4	0.2	I. setosa	3	3
4.9	3.1	1.5	0.2	I. setosa	3	3
5	3.2	1.2	0.2	I. setosa	3	3
5.5	3.5	1.3	0.2	I. setosa	3	3
4.9	3.6	1.4	0.1	I. setosa	3	3
4.4	3	1.3	0.2	I. setosa	3	3

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.1	3.4	1.5	0.2	I. setosa	3	3
5	3.5	1.3	0.3	I. setosa	3	3
4.5	2.3	1.3	0.3	I. setosa	3	3
4.4	3.2	1.3	0.2	I. setosa	3	3
5	3.5	1.6	0.6	I. setosa	3	3
5.1	3.8	1.9	0.4	I. setosa	3	3
4.8	3	1.4	0.3	I. setosa	3	3
5.1	3.8	1.6	0.2	I. setosa	3	3
4.6	3.2	1.4	0.2	I. setosa	3	3
5.3	3.7	1.5	0.2	I. setosa	3	3
5	3.3	1.4	0.2	I. setosa	3	3
7	3.2	4.7	1.4	I. versicolor	2	1
6.4	3.2	4.5	1.5	I. versicolor	2	2
6.9	3.1	4.9	1.5	I. versicolor	2	1
5.5	2.3	4	1.3	I. versicolor	2	2
6.5	2.8	4.6	1.5	I. versicolor	2	2
5.7	2.8	4.5	1.3	I. versicolor	2	2
6.3	3.3	4.7	1.6	I. versicolor	2	2
4.9	2.4	3.3	1	I. versicolor	2	2
6.6	2.9	4.6	1.3	I. versicolor	2	2
5.2	2.7	3.9	1.4	I. versicolor	2	2
5	2	3.5	1	I. versicolor	2	2
5.9	3	4.2	1.5	I. versicolor	2	2
6	2.2	4	1	I. versicolor	2	2
6.1	2.9	4.7	1.4	I. versicolor	2	2
5.6	2.9	3.6	1.3	I. versicolor	2	2
6.7	3.1	4.4	1.4	I. versicolor	2	2
5.6	3	4.5	1.5	I. versicolor	2	2
5.8	2.7	4.1	1	I. versicolor	2	2
6.2	2.2	4.5	1.5	I. versicolor	2	2
5.6	2.5	3.9	1.1	I. versicolor	2	2
5.9	3.2	4.8	1.8	I. versicolor	2	2
6.1	2.8	4	1.3	I. versicolor	2	2
6.3	2.5	4.9	1.5	I. versicolor	2	2
6.1	2.8	4.7	1.2	I. versicolor	2	2
6.4	2.9	4.3	1.3	I. versicolor	2	2
6.6	3	4.4	1.4	I. versicolor	2	2
6.8	2.8	4.8	1.4	I. versicolor	2	2

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
6.7	3	5	1.7	I. versicolor	2	1
6	2.9	4.5	1.5	I. versicolor	2	2
5.7	2.6	3.5	1	I. versicolor	2	2
5.5	2.4	3.8	1.1	I. versicolor	2	2
5.5	2.4	3.7	1	I. versicolor	2	2
5.8	2.7	3.9	1.2	I. versicolor	2	2
6	2.7	5.1	1.6	I. versicolor	2	2
5.4	3	4.5	1.5	I. versicolor	2	2
6	3.4	4.5	1.6	I. versicolor	2	2
6.7	3.1	4.7	1.5	I. versicolor	2	2
6.3	2.3	4.4	1.3	I. versicolor	2	2
5.6	3	4.1	1.3	I. versicolor	2	2
5.5	2.5	4	1.3	I. versicolor	2	2
5.5	2.6	4.4	1.2	I. versicolor	2	2
6.1	3	4.6	1.4	I. versicolor	2	2
5.8	2.6	4	1.2	I. versicolor	2	2
5	2.3	3.3	1	I. versicolor	2	2
5.6	2.7	4.2	1.3	I. versicolor	2	2
5.7	3	4.2	1.2	I. versicolor	2	2
5.7	2.9	4.2	1.3	I. versicolor	2	2
6.2	2.9	4.3	1.3	I. versicolor	2	2
5.1	2.5	3	1.1	I. versicolor	2	2
5.7	2.8	4.1	1.3	I. versicolor	2	2
6.3	3.3	6	2.5	I. virginica	1	1
5.8	2.7	5.1	1.9	I. virginica	1	2
7.1	3	5.9	2.1	I. virginica	1	1
6.3	2.9	5.6	1.8	I. virginica	1	1
6.5	3	5.8	2.2	I. virginica	1	1
7.6	3	6.6	2.1	I. virginica	1	1
4.9	2.5	4.5	1.7	I. virginica	1	2
7.3	2.9	6.3	1.8	I. virginica	1	1
6.7	2.5	5.8	1.8	I. virginica	1	1
7.2	3.6	6.1	2.5	I. virginica	1	1
6.5	3.2	5.1	2	I. virginica	1	1
6.4	2.7	5.3	1.9	I. virginica	1	1
6.8	3	5.5	2.1	I. virginica	1	1
5.7	2.5	5	2	I. virginica	1	2

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.8	2.8	5.1	2.4	I. virginica	1	2
6.4	3.2	5.3	2.3	I. virginica	1	1
6.5	3	5.5	1.8	I. virginica	1	1
7.7	3.8	6.7	2.2	I. virginica	1	1
7.7	2.6	6.9	2.3	I. virginica	1	1
6	2.2	5	1.5	I. virginica	1	2
6.9	3.2	5.7	2.3	I. virginica	1	1
5.6	2.8	4.9	2	I. virginica	1	2
7.7	2.8	6.7	2	I. virginica	1	1
6.3	2.7	4.9	1.8	I. virginica	1	2
6.7	3.3	5.7	2.1	I. virginica	1	1
7.2	3.2	6	1.8	I. virginica	1	1
6.2	2.8	4.8	1.8	I. virginica	1	2
6.1	3	4.9	1.8	I. virginica	1	2
6.4	2.8	5.6	2.1	I. virginica	1	1
7.2	3	5.8	1.6	I. virginica	1	1
7.4	2.8	6.1	1.9	I. virginica	1	1
7.9	3.8	6.4	2	I. virginica	1	1
6.4	2.8	5.6	2.2	I. virginica	1	1
6.3	2.8	5.1	1.5	I. virginica	1	2
6.1	2.6	5.6	1.4	I. virginica	1	1
7.7	3	6.1	2.3	I. virginica	1	1
6.3	3.4	5.6	2.4	I. virginica	1	1
6.4	3.1	5.5	1.8	I. virginica	1	1
6	3	4.8	1.8	I. virginica	1	2
6.9	3.1	5.4	2.1	I. virginica	1	1
6.7	3.1	5.6	2.4	I. virginica	1	1
6.9	3.1	5.1	2.3	I. virginica	1	1
5.8	2.7	5.1	1.9	I. virginica	1	2
6.8	3.2	5.9	2.3	I. virginica	1	1
6.7	3.3	5.7	2.5	I. virginica	1	1
6.7	3	5.2	2.3	I. virginica	1	1
6.3	2.5	5	1.9	I. virginica	1	2
6.5	3	5.2	2	I. virginica	1	1
6.2	3.4	5.4	2.3	I. virginica	1	1
5.9	3	5.1	1.8	I. virginica	1	2

Number of clusters: 3

	Number	Within	Average	Maximum
	of	cluster	distance	distance
	observations	sum of	from	from
		squares	centroid	centroid
Cluster1	39	25.414	0.732	1.552
Cluster2	61	38.291	0.731	1.647
Cluster3	50	15.151	0.482	1.248

We have also performed **linear discriminant analysis** by considering types as the true groups.

Linear Method for Response: Type
Predictors: Sepal le Sepal wi Petal le Petal wi
Summary of Classification

Put into	....True Group....		
Group	1	2	3
1	49	2	0
2	1	48	0
3	0	0	50
Total N	50	50	50

Summary of Classification with Cross-validation

Put into	....True Group....		
Group	1	2	3
1	49	2	0
2	1	48	0
3	0	0	50
Total N	50	50	50
N Correct	49	48	50
Proportion	0.980	0.960	1.000

N = 150 N Correct = 147 Proportion Correct = 0.980
Squared Distance Between Groups

	1	2	3
1	0.000	17.201	179.385
2	17.201	0.000	89.864
3	179.385	89.864	0.000

Linear Discriminant Function for Group

	1	2	3
Constant	−103.27	−71.75	−85.21
Sepal le	12.45	15.70	23.54
Sepal wi	3.69	7.07	23.59
Petal le	12.77	5.21	−16.43
Petal wi	21.08	6.43	−17.40

Variable Pooled Means for Group

	Mean	1	2	3
Sepal le	5.8433	6.5880	5.9360	5.0060
Sepal wi	3.0573	2.9740	2.7700	3.4280
Petal le	3.7580	5.5520	4.2600	.4620
Petal wi	1.1993	2.0260	1.3260	0.2460

Variable Pooled StDev for Group

	StDev	1	2	3
Sepal le	0.5148	0.6359	0.5162	0.3525
Sepal wi	0.3397	0.3225	0.3138	0.3791
Petal le	0.4303	0.5519	0.4699	0.1737
Petal wi	0.2047	0.2747	0.1978	0.1054

Pooled Covariance Matrix
Sepal le Sepal wi Petal le Petal wi
Sepal le 0.26501
Sepal wi 0.09272 0.11539
Petal le 0.16751 0.05524 0.18519
Petal wi 0.03840 0.03271 0.04267 0.04188
Here we see that only three observations are wrongly classified. The corresponding probabilities are given by

Observation	True	Pred	Group	Probability
	Group	Group		Predicted
71 **	2	1	1	0.75
			2	0.25
			3	0.00
84 **	2	1	1	0.86
			2	0.14
			3	0.00
134 **	1	2	1	0.27
			2	0.73
			3	0.00

Example 8.5.2 The following data are related to a survey on environmental pollution level. The following variables were observed in suitable units at 111 selected places. The four variables under study were Ozone content, Radiation, Temperature, and Wind speed in some proper units. We have performed hierarchical clustering with Euclidian distance and single linkage. The data set as well as the cluster membership is shown in the following table.

The summary of results and the dendrogram are given below the table. By considering similarity level at 93, six clusters were found of which three (4, 5, and 6) may omitted as outliers containing 2, 1, and 1 observations. Hence clusters 1, 2, and 3 are the main clusters. Figures corresponding to radiation, temperature, wind speed, ozone content and H-cluster number of 111 places.

Table 8.2 Results of hierarchical clustering for pollution data

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
190	67	7.4	41	1
118	72	8	36	2
149	74	12.6	12	2
313	62	11.5	18	1
299	65	8.6	23	1
99	59	13.8	19	2
19	61	20.1	8	3
256	69	9.7	16	1
290	66	9.2	11	1
274	68	10.9	14	1
65	58	13.2	18	3
334	64	11.5	14	1
307	66	12	34	1
78	57	18.4	6	3
322	68	11.5	30	1
44	62	9.7	11	3
8	59	9.7	1	3
320	73	16.6	11	1
25	61	9.7	4	3
92	61	12	32	2
13	67	12	23	3
252	81	14.9	45	1
223	79	5.7	115	1
279	76	7.4	37	1
127	82	9.7	29	2
291	90	13.8	71	1

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
323	87	11.5	39	1
148	82	8	23	2
191	77	14.9	21	1
284	72	20.7	37	1
37	65	9.2	20	3
120	73	11.5	12	2
137	76	10.3	13	2
269	84	4	135	4
248	85	9.2	49	1
236	81	9.2	32	1
175	83	4.6	64	1
314	83	10.9	40	1
276	88	5.1	77	1
267	92	6.3	97	1
272	92	5.7	97	1
175	89	7.4	85	1
264	73	14.3	10	1
175	81	14.9	27	1
48	80	14.3	7	3
260	81	6.9	48	1
274	82	10.3	35	1
285	84	6.3	61	1
187	87	5.1	79	1
220	85	11.5	63	1
7	74	6.9	16	3
294	86	8.6	80	1
223	85	8	108	1
81	82	8.6	20	3
82	86	12	52	3
213	88	7.4	82	1
275	86	7.4	50	1
253	83	7.4	64	1
254	81	9.2	59	1
83	81	6.9	39	3
24	81	13.8	9	3
77	82	7.4	16	3
255	89	4	122	4
229	90	10.3	89	1
207	90	8	110	1

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
192	86	11.5	44	1
273	82	11.5	28	1
157	80	9.7	65	1
71	77	10.3	22	3
51	79	6.3	59	5
115	76	7.4	23	2
244	78	10.9	31	1
190	78	10.3	44	1
259	77	15.5	21	1
36	72	14.3	9	3
212	79	9.7	45	1
238	81	3.4	168	6
215	86	8	73	1
203	97	9.7	76	1
225	94	2.3	118	1
237	96	6.3	84	1
188	94	6.3	85	1
167	91	6.9	96	1
197	92	5.1	78	1
183	93	2.8	73	1
189	93	4.6	91	1
95	87	7.4	47	3
92	84	15.5	32	3
252	80	10.9	20	1
220	78	10.3	23	1
230	75	10.9	21	1
259	73	9.7	24	1
236	81	14.9	44	1
259	76	15.5	21	1
238	77	6.3	28	1
24	71	10.9	9	3
112	71	11.5	13	2
237	78	6.9	46	1
224	67	13.8	18	1
27	76	10.3	13	3
238	68	10.3	24	1
201	82	8	16	1
238	64	12.6	13	1
14	71	9.2	23	3

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
139	81	10.3	36	2
49	69	10.3	7	3
20	63	16.6	14	3
193	70	6.9	30	1
191	75	14.3	14	1
131	76	8	18	2
223	68	11.5	20	1

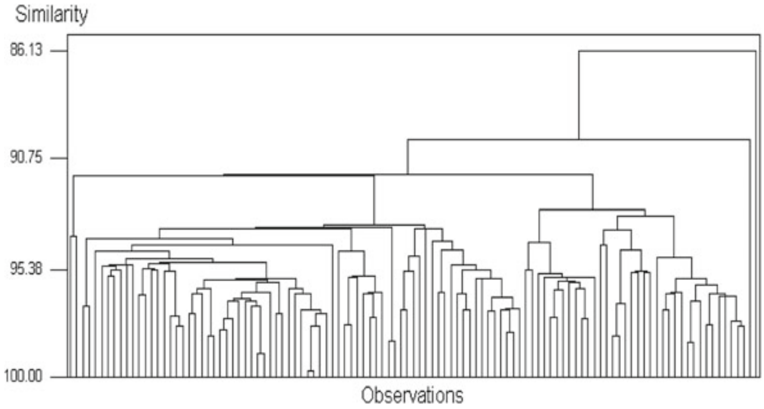

Fig. 8.1 Dendrogram of pollution data

Number of main clusters: 3

```
               Number      Within     Average  Maximum
                 of        cluster    distance distance
            observations   sum of      from     from
                           squares    centroid centroid
Cluster1         71      202337.219   48.851   101.003
Cluster2         12        5151.429   18.929    35.732
Cluster3         24       26269.208   30.505    58.654
```

Cluster Centroids

```
Variable   Cluster1  Cluster2  Cluster3 Grand centroid
Radiatio   240.7606  123.9167  46.6250     184.8018
Temperat    80.1831   73.5833  71.9167      77.7928
Wind spe     9.6577   10.2583  11.5292       9.9387
Ozone Co    49.2535   22.1667  17.7500      42.0991
```

The dendrogram of the pollution data is shown below. The centroids of the first three clusters are widely separated corresponding to all the variables; the 24 places falling in cluster 3 may be considered to be least polluted, whereas the 71 places falling in cluster 1 are most polluted (Fig. 8.1).

References and Suggested Readings

Bien, J., & Tibshirani, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, *106*(495), 1075–1084.

Chattopadhyay, A. K., & Chattopadhyay, T. (2014). *Statistical methods for astronomical data analysis.*, Springer series in astrostatistics New York: Springer.

Chattopadhyay, T., et al. (2012). Uncovering the formation of ultracompact dwarf galaxies by multivariate statistical analysis. *Astrophysical Journal*, *750*, 91.

De, T., Chattopadhyay, T., & Chattopadhyay, A. K. (2013). Comparison among clustering and classification techniques on the basis of galaxy data. *Calcutta Statistical Association Bulletin*, *65*, 257–260.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. (2015). Multivariate approaches to classification in extragalactic astronomy. *Frontiers in Astronomy and Space Science*, *2*, 1–17.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, p. 281).

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342.

Sugar, A. S., & James, G. M. (2003). Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, *98*, 750.