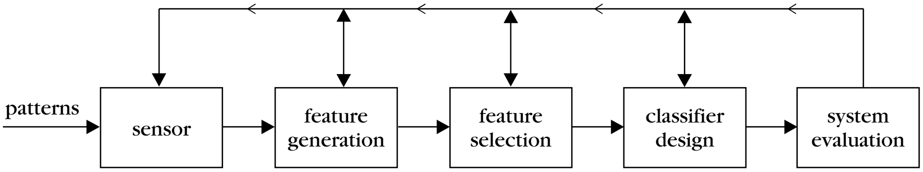

FIGURE 1.3

The basic stages involved in the design of a classification system.

to redesign earlier stages in order to improve the overall performance. Furthermore, there are some methods that combine stages, for example, the feature selection and the classifier design stage, in a common optimization task.

Although the reader has already been exposed to a number of basic problems at the heart of the design of a classification system, there are still a few things to be said.

1.3 SUPERVISED, UNSUPERVISED, AND SEMI-SUPERVISED LEARNING

In the example of Figure 1.1, we assumed that a set of training data were available, and the classifier was designed by exploiting this *a priori* known information. This is known as *supervised pattern recognition* or in the more general context of machine learning as *supervised learning*. However, this is not always the case, and there is another type of pattern recognition tasks for which training data, of known class labels, are not available. In this type of problem, we are given a set of feature vectors $\boldsymbol{x}$ and the goal is to unravel the underlying *similarities* and *cluster* (group) "similar" vectors together. This is known as *unsupervised pattern recognition* or *unsupervised learning* or *clustering*. Such tasks arise in many applications in social sciences and engineering, such as remote sensing, image segmentation, and image and speech coding. Let us pick two such problems.

In *multispectral remote sensing*, the electromagnetic energy emanating from the earth's surface is measured by sensitive scanners located aboard a satellite, an aircraft, or a space station. This energy may be reflected solar energy (passive) or the reflected part of the energy transmitted from the vehicle (active) in order to "interrogate" the earth's surface. The scanners are sensitive to a number of wavelength bands of the electromagnetic radiation. Different properties of the earth's surface contribute to the reflection of the energy in the different bands. For example, in the visible–infrared range properties such as the mineral and moisture contents of soils, the sedimentation of water, and the moisture content of vegetation are the main contributors to the reflected energy. In contrast, at the thermal end of the infrared, it is the thermal capacity and thermal properties of the surface and near subsurface that contribute to the reflection. Thus, each band measures different properties of the same patch of the earth's surface. In this way, images of the earth's surface corresponding to the spatial distribution of the reflected energy in each band can be created. The task now is to exploit this information in order to identify the various ground cover types, that is, built-up land, agricultural land, forest, fire burn, water, and diseased crop. To this end, one feature vector $\boldsymbol{x}$ for each cell from the "sensed" earth's surface is formed. The elements $x_i, i = 1, 2, \ldots, l$, of the vector are the corresponding image pixel intensities in the various spectral bands. In practice, the number of spectral bands varies.

A *clustering* algorithm can be employed to reveal the groups in which feature vectors are clustered in the l-dimensional feature space. Points that correspond to the same ground cover type, such as water, are expected to cluster together and form groups. Once this is done, the analyst can identify the type of each cluster by associating a sample of points in each group with available reference ground data, that is, maps or visits. Figure 1.4 demonstrates the procedure.

Clustering is also widely used in the social sciences in order to study and correlate survey and statistical data and draw useful conclusions, *which will then lead to the right actions*. Let us again resort to a simplified example and assume that we are interested in studying whether there is any relation between a country's gross national product (GNP) and the level of people's illiteracy, on the one hand, and children's mortality rate on the other. In this case, each country is represented by a three-dimensional feature vector whose coordinates are indices measuring the quantities of interest. A clustering algorithm will then reveal a rather compact cluster corresponding to countries that exhibit low GNPs, high illiteracy levels, and high children's mortality expressed as a population percentage.

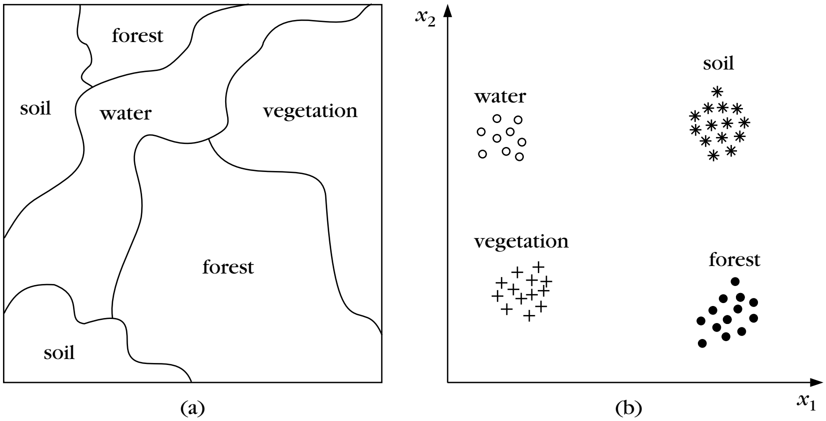

FIGURE 1.4

(a) An illustration of various types of ground cover and (b) clustering of the respective features for multispectral imaging using two bands.

A major issue in unsupervised pattern recognition is that of defining the "similarity" between two feature vectors and choosing an appropriate measure for it. Another issue of importance is choosing an algorithmic scheme that will cluster (group) the vectors on the basis of the adopted similarity measure. In general, different algorithmic schemes may lead to different results, which the expert has to interpret.

Semi-supervised learning/pattern recognition for designing a classification system shares the same goals as the supervised case, however now, the designer has at his or her disposal a set of patterns of unknown class origin, in addition to the training patterns, whose true class is known. We usually refer to the former ones as *unlabeled* and the latter as *labeled* data. Semi-supervised pattern recognition can be of importance when the system designer has access to a rather limited number of labeled data. In such cases, recovering additional information from the unlabeled samples, related to the general structure of the data at hand, can be useful in improving the system design. Semi-supervised learning finds its way also to clustering tasks. In this case, labeled data are used as constraints in the form of *must-links* and *cannot-links*. In other words, the clustering task is constrained to assign certain points in the same cluster or to exclude certain points of being assigned in the same cluster. From this perspective, semi-supervised learning provides an *a priori* knowledge that the clustering algorithm has to respect.

1.4 MATLAB PROGRAMS

At the end of most of the chapters there is a number of MATLAB programs and computer experiments. The MATLAB codes provided are not intended to form part of a software package, but they are to serve a purely pedagogical goal. Most of these codes are given to our students who are asked to play with and discover the "secrets" associated with the corresponding methods. This is also the reason that for most of the cases the data used are simulated data around the Gaussian distribution. They have been produced carefully in order to guide the students in understanding the basic concepts. This is also the reason that the provided codes correspond to those of the techniques and algorithms that, to our opinion, comprise the backbone of each chapter and the student has to understand in a first reading. Whenever the required MATLAB code was available (at the time this book was prepared) in a MATLAB toolbox, we chose to use the associated MATLAB function and explain how to use its arguments. No doubt, each instructor has his or her own preferences, experiences, and unique way of viewing teaching. The provided routines are written in a way that can run on other data sets as well. In a separate accompanying book we provide a more complete list of MATLAB codes embedded in a user-friendly Graphical User Interface (GUI) and also involving more realistic examples using real images and audio signals.