CHAPTER 1

PROLOGUE

Mathematical models play a crucial role in scientific studies. This is because they serve many purposes including function approximation, interpolation and extrapolation. Pattern recognition interests in the mapping between inputs and discrete-valued outputs, which are commonly referred to as classes. In other words, pattern recognition concentrates on creating a mathematical model that captures an input-output relationship which leads to the correct identification of output classes after observing input features or attributes. Since the procedure of capturing the input-output relationship involves many steps, research in pattern recognition can be categorised according to these steps. Examples of these steps are illustrated in Figure 1-1. An information theory has a significant influence on the development of various research areas in pattern recognition. These include attribute discretisation, attribute selection and classification model construction. Attribute discretisation involves a transformation of continuous-valued attributes into discrete-valued attributes (Fayyad and Irani, 1993). It is required when a number of classifiers including a naïve Bayes classifier (Mitchell, 1997) and decision trees (Quinlan, 1993) are employed. Attribute selection interests in the identification of optimal attribute subset that leads to the maximum classification accuracy. Information-theoretic attribute selection techniques include a correlation-based feature selection technique (Hall and Holmes, 2003), a simple symmetrical uncertainty ranking technique (Press et al., 1988) and a newly proposed round robin symmetrical uncertainty ranking technique. Classification model construction concentrates on maximising the capability of identifying the correct class of an unknown pattern based on the available pattern data. A C4.5 decision tree (Quinlan, 1993) and a random forest (Breiman, 2001) are examples of information-theoretic techniques for classification model construction.

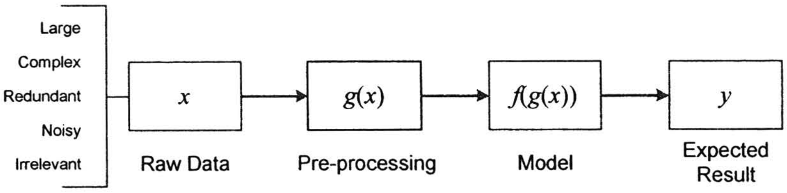

FIGURE 1-1 Steps in pattern recognition. The pre-processing step may involve attribute selection and attribute discretisation.

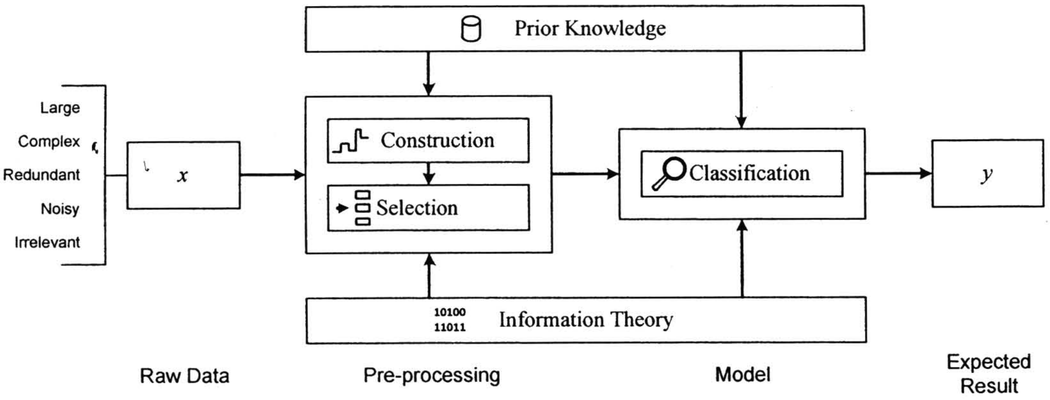

FIGURE 1-2 An information theory influences various steps in pattern recognition.

The influence of information theory on pattern recognition is summarised in Figure 1-2. In this thesis, all three areas of information-theoretic pattern recognition are explored. Specifically, they are applied to two life science problems: identification of ancestry informative markers (AIMs) and thalassaemia classification. AIMs are discrete-valued genetic markers that can be used to identify population labels for a population classification task. Since there are over 3,000,000 markers in the human genome, the problem can be formulated as an attribute selection problem. It will be demonstrated that the protocol involving the round robin symmetrical uncertainty ranking technique and a naïve Bayes classifier, which is shown in Figure 1-3, provides a sufficient means for extracting AIM panels from the genome-wide data. In contrast, the thalassaemia classification problem covers a smaller number of attributes. Moreover, the attributes are continuous-valued attributes. It will be shown that the problem can be solved using a procedure shown in Figure 1-4, which includes an information-theoretic attribute discretisation technique (Fayyad and Irani, 1993), the correlation-based feature selection technique

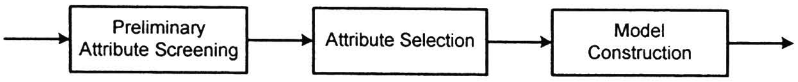

FIGURE 1-3 Schematic diagram of the proposed AIM identification protocol. Details of each step are given in Chapter 2.

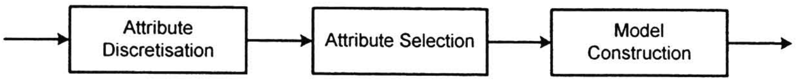

FIGURE 1-4 Schematic diagram of the procedure for solving the thalassaemia classification problem. Details of each step are given in Chapter 3.

and a C4.5 decision tree. The organisation of this thesis is as follows. In Chapter 2, the AIM identification problem is defined. All necessary techniques for solving the problem are also explained. Next, the thalassaemia classification problem and how it can be solved are discussed in Chapter 3. Finally, the epilogue is given in Chapter 4.