CHAPTER 3

THALASSAEMIA CLASSIFICATION

3.1 Introduction

Thalassaemia is a genetic disease that causes a reduction in the life span of a red blood cell (Weatherall and Clegg, 2001). The disease is a result of an abnormality in the genes that regulate the formation of a protein called globin, which is a major component of haemoglobin (Hb). Each red blood cell contains approximately 300 million molecules of haemoglobin. Hence, a change in the structure of globin affects the structure and functionality of a red blood cell. A globin molecule contains two parts: α-globin and β-globin. The α-globin contains 141 amino acids, which are regulated by genes on chromosome 16. The β-globin consists of 146 amino acids, which are governed by genes on chromosome 11. Since the regulatory genes reside on two autosomes, the transmission mode of abnormal genes is autosomal recessive. Hence, a person must have two copies of a recessive gene on the same chromosome in order to have the disease. In order to make the diagnosis, the blood characteristics must be analysed. A complete blood count (CBC) and haemoglobin typing are the primary screening tests for a laboratory diagnosis of thalassaemia. However, there is still a limitation in the analysis of data due to a large number of possible candidate characteristics. In addition, there are various types of thalassaemia and thalassaemia trait. (Persons with thalassaemia trait do not have the disease but inherit genes that cause the disease.) As a result, a manual diagnostic process can only be carried out by specialists (Jimenez, Minchinela and Ros, 1995; Demir et al., 2002; Ntaios et al., 2007).

Early attempts to formulate an automated diagnostic tool concentrate on analysing CBC data with image analysis (Lund and Barnes, 1972), statistical (Engle et al., 1976) and clustering techniques (Barosi et al., 1985). Later, the implementation protocol *has shifted to the expert systems, in which both rule-based (Quaglini et al., 1986,* 1988; Lanzola et al., 1990) and hybrid neural network/rule-based systems (Birndorf et al., 1996) have been successfully tested in clinical trials. Nonetheless, these tools broadly differentiate between a wide range of blood-related diseases including various types of anaemia. In order to narrow the diagnostic target down to the differentiation between thalassaemic patients, persons with thalassaemia trait and normal subjects, an alternative automated diagnostic tool is required. Recently, a successful implementation of a multilayer perceptron (Amendolia et al., 2002, 2003; Wongseree et al., 2007), a k-nearest neighbour technique (Amendolia et al., 2003), a support vector machine (Amendolia et al., 2003) and a genetic programming based decision tree (Wongseree et al., 2007) as a thalassaemic diagnostic tool has been reported. Among these tools, the multilayer perceptron (Rumelhart and McClelland, 1986) emerges as the most suitable tool for thalassaemia classification problems in Thailand which cover higher varieties of haemoglobinopathies than other countries (Fucharoen et al., 1997). In the early investigation, it has been demonstrated that a multilayer perceptron can handle a problem with 13 classes of thalassaemic abnormality and two classes of normal subjects with and without iron deficiency (Wongseree et al., 2007).

Although the use of a neural network with CBC inputs has been proven to be successful, further investigation into automated thalassaemia classification is still required. Specifically, haemoglobin typing data should also be considered as possible inputs (Kutlar, 2007). This is because CBC and haemoglobin typing data represent different aspects of blood characteristics. CBC information is useful for the diagnosis of various types of anaemia while haemoglobin typing data alone can confirm the haemoglobinopathies. In this thesis, the possibility of using haemoglobin typing data in automatic thalassaemia classification is investigated. The choices of classifier for the task include a multilayer perceptron, a C4.5 decision tree (Quinlan, 1993) and random forests (Breiman, 2001). A multilayer perceptron is selected for this study because it has been proven to be a suitable classifier in early investigations (Amendolia et al., 2002, 2003; Wongseree et al., 2007). In contrast, both C4.5 and random forests have never been used in thalassaemia classification problems. Nonetheless, both classifiers have been successfully applied to many chemometric applications. For instance, C4.5 has been used for ion chromatography detection (Mulholland et al., 1995a,b) while random forests have been implemented for prediction of drug's chromatographic retention time (Hancock et al., 2005), near-infrared spectrum analysis of red grape homogenates (Donald et al., 2006a), and classification of prostate cancer (Donald et al., 2006b) and agro-industrial products (Granitto et al., 2006).

With the availability of haemoglobin typing data and selected choices of classifier, an investigation can be conducted as follows. Firstly, the data are pre-processed via input attribute (feature) discretisation and reduction. It has been reported that a proper discretisation of continuous-valued attributes can improve the classification performance of decision tree classifiers (Fayyad and Irani, 1993). In addition, an early investigation indicates that measured blood characteristics usually contain non-informative attributes which can be eliminated without affecting the classification outcome (Wongseree et al., 2007). The attributes are thus discretised via an information-theoretic technique (Fayyad and Irani, 1993) while redundant attributes are eliminated using a correlation-based feature selection technique (Hall and Holmes, 2003). As a result, two reduced-attribute data sets are available for classifier benchmarking: continuous- and discrete-valued data sets. Since the attribute discretisation is also required prior to the continuous-valued attribute reduction in correlation analysis (Hall and Holmes, 2003), eliminated attributes from the original continuous-valued data coincide with those from the derived discrete-valued data. The use of both continuous- and discrete-valued data in the benchmarking of neural networks is necessary since there is not enough evidence which suggests that one attribute representation is better than the other. After the classifier performance evaluation for both cases of attribute representation is completed, the best classifier together with the suitable attribute format is picked for a clinical trial involving a separate data set. Finally, classification analysis of clinical trial results is carried out to

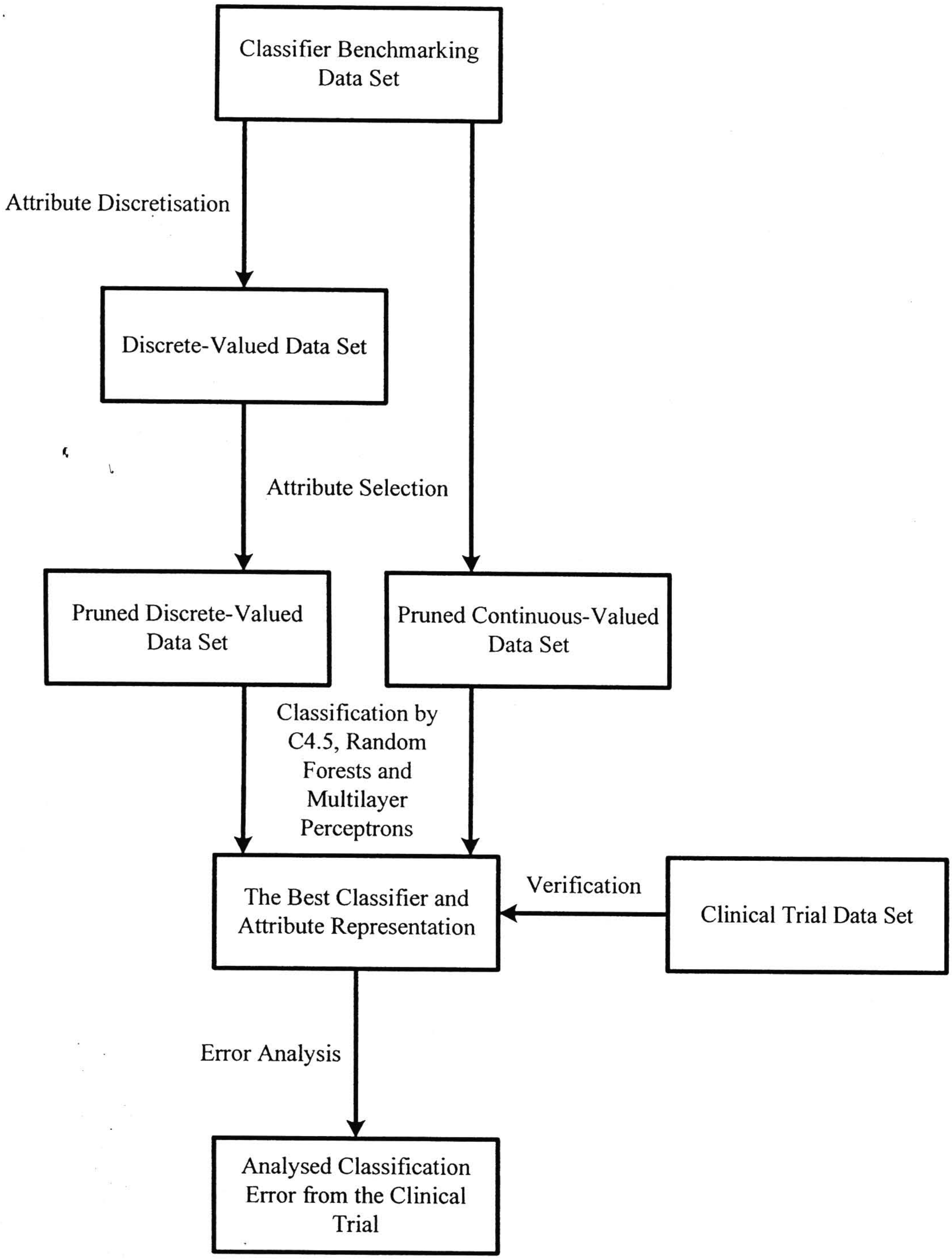

Classifier Benchmarking Data Set
Attribute Discretisation
Discrete-Valued Data Set
Attribute Selection
Pruned Discrete-Valued Data Set
Pruned Continuous-Valued Data Set
Classification by C4.5, Random Forests and Multilayer Perceptrons
The Best Classifier and Attribute Representation
Verification
Clinical Trial Data Set
Error Analysis
Analysed Classification Error from the Clinical Trial

FIGURE 3-1 Schematic diagram for the methodology employed in the investigation.

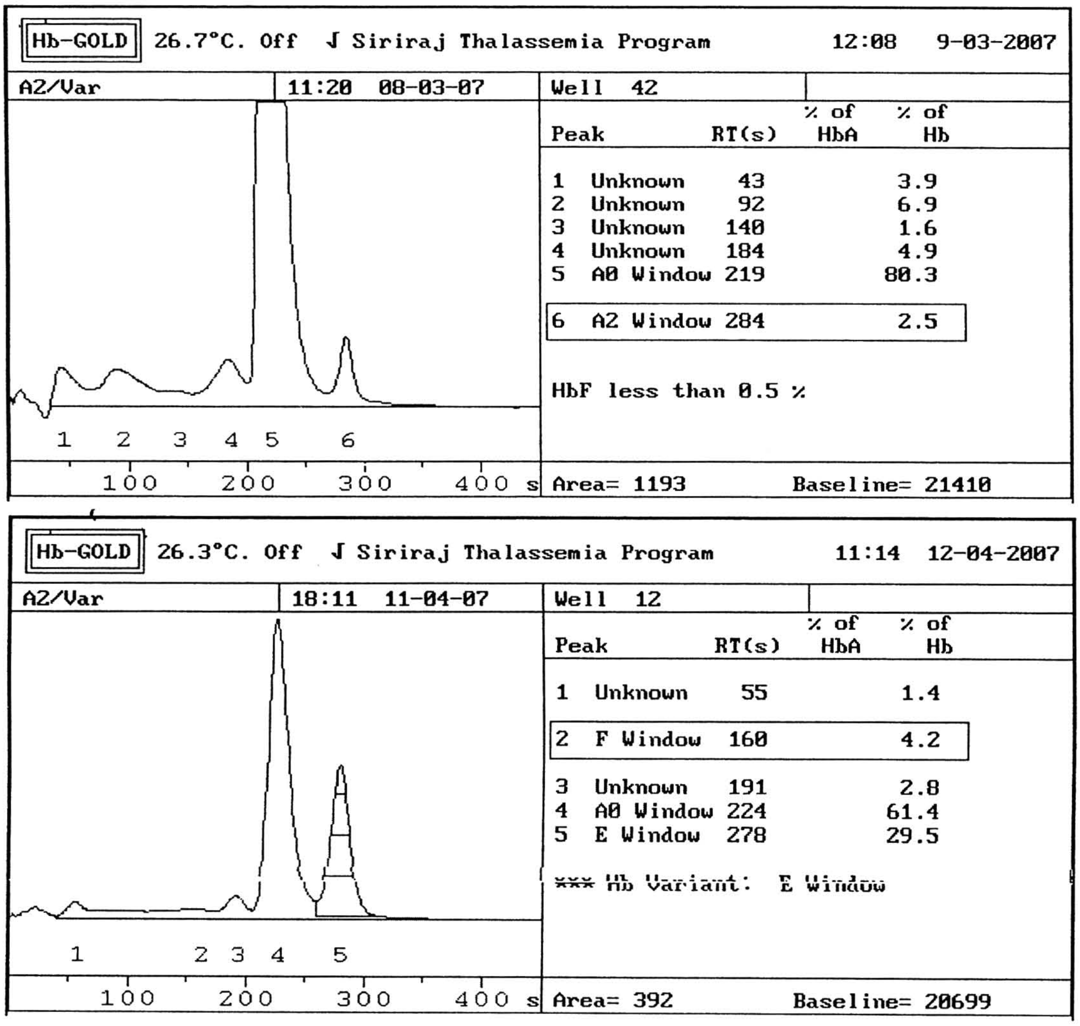
determine the feasibility of the selected classifier. Every step in the procedure described above is illustrated in Figure 3-1 and implemented using a WEKA package (Witten and Frank, 2005).

The organisation of this chapter is as follows. In section 3.2, materials and methods are explained. These include the description of haemoglobin typing data, an information-theoretic attribute discretisation technique, a correlation-based feature selection technique, a multilayer perceptron, C4.5, and random forests. Results from attribute discretisation and redundancy elimination, classifier benchmarking and a clinical trial are then discussed in section 3.3. Finally, the conclusions and further works are given in section 3.4.

3.2 Materials and Methods

3.2.1 Haemoglobin Typing Data Sets

A blood specimen generally contains more than one type of haemoglobin. With the use of high performance liquid chromatography (HPLC) (Clarke and Higgins, 2000), the haemoglobin contents in each blood specimen can be characterised. Various types of thalassaemia are identifiable through the difference in proportion of haemoglobin contents (Ou and Rognerud, 2001; Old, 2003; Colah et al., 2007). Haemoglobin typing results are usually obtained in the form of elution chromatograms (Joutovsky, Hadzi-Nesic and Nardi, 2004). Typical elution chromatograms of a normal specimen and a specimen from a person with Hb E trait are illustrated in Figure 3-2. The normal specimen is mostly made up from Hb A_0 while the specimen from a person with Hb E trait consists of Hb F, Hb A_0 and Hb E. Since different types of haemoglobin are detectable in the form of elution peaks at different retention time, a chromatogram can be divided into multiple sections where each section occupies a non-overlapping range of retention time. Each chromatogram section would represent a unique input feature or attribute for thalassaemia classification in which the percentage of haemoglobin in the elution profile corresponds to the attribute value. In this investigation, a chromatogram is divided into eight sections. The attribute set and the associated types of haemoglobin are summarised in Table 3-1. In Table 3-1, eight attributes are defined according to the possible range of retention time in a haemoglobin chromatogram. Some attributes are



```
Hb-GOLD  26.7°C. Off  √ Siriraj Thalassemia Program          12:08    9-03-2007
A2/Var              11:20  08-03-07     Well  42
                                                            % of     % of
                                        Peak          RT(s)  HbA       Hb

                                        1  Unknown     43               3.9
                                        2  Unknown     92               6.9
                                        3  Unknown    140               1.6
                                        4  Unknown    184               4.9
                                        5  A0 Window  219              80.3

                                        6  A2 Window  284               2.5

                                        HbF less than 0.5 %

                                        Area= 1193             Baseline= 21410
```

```
Hb-GOLD  26.3°C. Off  √ Siriraj Thalassemia Program          11:14  12-04-2007
A2/Var              18:11  11-04-07     Well  12
                                                            % of     % of
                                        Peak          RT(s)  HbA       Hb

                                        1  Unknown     55               1.4

                                        2  F Window   160               4.2

                                        3  Unknown    191               2.8
                                        4  A0 Window  224              61.4
                                        5  E Window   278              29.5

                                        *** Hb Variant:  E Window

                                        Area= 392              Baseline= 20699
```

FIGURE 3-2 Elution chromatograms of (a) a normal specimen and (b) a specimen from a person with Hb E trait that are obtained from an Hb Gold HPLC system. RT(s) represents the retention time in seconds for each fraction of elute. % of Hb represents the percentage of haemoglobin in the elution peak.

related to known types of haemoglobin while the others are corresponded to unknown haemoglobin. Two confirmed diagnosis data sets are acquired for this investigation. The first data set is created for the evaluation of classifier performance while the second set is used in a clinical trial. The data set for classifier evaluation consists of 150 samples which represent the majority of blood specimens from adults that need to be screened for thalassaemia. On the other hand, the data set for clinical trial contains 1,000 samples and represents a typical distribution of specimens which are submitted

TABLE 3-1 Input features or attributes for thalassaemia classification.

Attribute	Type of haemoglobin	Retention time (s)
1	Hb Bart's	0– 68
2	Hb A_{1c}, Hb F	69–160
3	Unknown	161–199
4	Hb A_0	200–230
5	Unknown	231–249
6	Hb E	250–280
7	Hb A_2	281–289
8	Hb D, Hb S, Hb Constant Spring, Hb C	290–320

Each attribute represents different type of haemoglobin and occupies a unique range of retention time.

for screening during a fixed time period. This data set is collected from Siriraj Hospital, Bangkok, Thailand during 1 August 2007 and 31 October 2007. The data acquisition has been conducted in accordance with the Faculty of Medicine Siriraj Hospital Ethics Committee's guidelines and in accordance with the Helsinki Declaration. In addition, informed consent has been obtained from all individuals. The description of these two sample sets is summarised in Table 3-2. From Table 3-2, the samples are made up from seven groups of thalassaemic patients, five groups of persons with thalassaemia trait, one group of persons with abnormal haemoglobin and one group of normal subjects. It is noticed that some types of thalassaemia in the data set for classifier benchmarking are not presented in the specimens collected for the clinical trial. Further, samples from persons with α-thalassaemia 1 and α-thalassaemia 2 traits are not included in this study. This is because haemoglobin typing characteristics cannot be used to differentiate between these two groups and the normal subject group. Generally, CBC and genotyping confirmation is needed for the diagnosis of these two types of thalassaemia trait (Old, 2003).

3.2.2 Attribute Discretisation

When attributes in the classification problem of interest are continuous-valued attributes, it is possible to transform these attributes into discrete-valued attributes. This transformation can be viewed as a form of data pre-processing procedure. The attribute discretisation technique that is selected for the current application is proposed by (Fayyad and Irani, 1993). The technique is an information-theoretic technique that employs entropy-based splitting and minimum description length stopping criteria. A chosen cut point within the range of each attribute values is guaranteed to lie between two class boundaries. A candidate cut point is introduced recursively to each sample subset and is acceptable if a significant information gain—the difference between the information value with and without the split is achieved. For a sample set S, which contains samples from m classes denoted by $C_1, ..., C_m$, the class entropy of S is defined as

$$Ent(S) = -\sum_{i=1}^{m} p(C_i, S)\log_2(p(C_i, S)) \qquad \text{Eq.3-1}$$

where $p(C_i, S)$ is the proportion of samples in S that belong to class C_i.A cut point T, which is introduced to an attribute A in the sample set S, will create a partition that has a class information entropy

$$E(A, T, S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2) \qquad \text{Eq.3-2}$$

where S_1 and S_2 are sample subset of S and $S_1 + S_2 = S$. According to the minimum description length stopping criterion, a cut point T is accepted if and only if

$$Gain(A, T, S) > \frac{\log_2(|S| - 1)}{|S|} + \frac{\log_2(3^m - 2) - [mEnt(S) - m_1 Ent(S_1) - m_2 Ent(S_2)]}{|S|} \qquad \text{Eq.3-3}$$

where m_1 and m_2 are the number of classes in the subset S_1 and S_2, respectively and $Gain(A, T, S)$ is the information gain of the cut point, which is defined by

$$Gain(A, T, S) = Ent(S) - \frac{|S_1|}{|S|}Ent(S_1) - \frac{|S_2|}{|S|}Ent(S_2) \qquad \text{Eq.3-4}$$

3.2.3 Attribute Selection

Attribute selection is the process of identifying and removing irrelevant and redundant information from input features. This procedure has been proven to help improving the classification robustness in thalassaemia classification (Amendolia et al., 2003; Wongseree et al., 2007). The chosen attribute selection method for the current investigation is a correlation-based feature selection technique (Hall and Holmes, 2003). Detailed description of the technique has been given in Chapter 2.

3.2.4 C4.5 Decision Tree

C4.5 decision tree is one of the most widely used and practiced tools for inductive inference (Quinlan, 1993). A decision tree is generally constructed in a top-down manner. The tree construction begins at the root node where each input feature or attribute is evaluated using a statistical test to determine how well it alone classifies the training samples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for either each possible value of this attribute if the attribute value is discrete or each possible discretised interval of this attribute if the attribute value is continuous. Next, the training samples are sorted to the appropriate descendant node. The entire process is subsequently repeated using the training samples associated with each descendant node to select the best attribute to test at that point in the tree. This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices. Although a new node can always be added to the tree until all samples which are assigned to one node belong to the same class, C4.5 does not allow the tree to grow to its maximum depth. As a result, a node is only introduced to the tree only when there are a sufficient number of samples left from sorting. After the complete tree has been constructed, a tree pruning is usually carried out to avoid data over-fitting.

The statistical test for assigning an attribute to each node in C4.5 also employs an entropy-based measure. The chosen attribute is the one with the highest information gain ratio among available attributes at the tree construction step considered. The information gain ratio $GainRatio(A, S)$ of an attribute A relative to a sample set S is defined as

$$GainRatio(A, S) = \frac{Gain(A, S)}{SplitInformation(A, S)} \qquad \text{Eq.3-5}$$

where

$$Gain(A, S) = Ent(S) - \sum_{v \in V} \frac{|S_v|}{|S|} Ent(S_v) \qquad \text{Eq.3-6}$$

and

$$SplitInformation(A, S) = -\sum_{v \in V} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}. \qquad \text{Eq.3-7}$$

V is the set of all possible values of attribute A and S_v is the subset of S for which attribute A has value v. This information gain ratio can be calculated straightaway for discrete-valued attributes. For continuous-valued attributes, it is necessary to partition the attribute value into a discrete set of intervals prior to the calculation of information gain ratio. Quinlan (1993) suggests that an appropriate threshold can be used to partition a continuous-valued attribute into two intervals. Let $\{v_1, v_2, ..., v_{|S|}\}$ be the sorted attribute values of attribute A. Candidate values that need to be considered are only the midpoints between v_i and v_{i+1}. The chosen midpoint is the value that leads to the partition that gives the highest information gain ratio according to equation 3-5. After the best midpoint is selected, C4.5 will pick the largest attribute value in A that does not exceed this midpoint as the threshold. Quinlan (1993) explains that this strategy ensures that all threshold values appearing in the tree actually occur in the data.

3.2.5 Random Forests

Random forests refer to a collection or ensemble of decision trees (Breiman, 2001). The technique takes a majority vote result from all trees as the class decision. Hence, the tree structures should be significantly diversified in order for the majority-vote concept to be applicable. This can be achieved by replacing the greedy search strategy for attribute selection as implemented in C4.5 with a stochastic attribute selection procedure. Breiman (2001) suggests that an attribute for each node in a tree

TABLE 3-2 Two data sets for thalassaemia classification.

Class	Description	Category	Number of samples	
			Classifier benchmarking	Clinical trial
1	Normal subject	Normal	15	281
2	Hb Constant Sprint trait	Trait	9	35
3	Hb E trait	Trait	15	500
4	α-Thalassaemia 1 trait + Hb E trait	Trait	15	43
5	Homozygous Hb E	Trait	15	64
6	β-Thalassaemia trait	Trait	15	44
7	Abnormal haemoglobin	N/A	6	6
8	Hb H disease	Disease	12	4
9	Hb H disease with Constant Spring	Disease	12	0
10	EA Bart's disease	Disease	9	3
11	HPFH disease	Disease	10	2
12	Homozygous β-Thalassaemia	Disease	5	0
13	β^0-Thalassaemia/Hb E	Disease	6	14
14	β^+-Thalassaemia/Hb E	Disease	6	4
	Total		150	1,000

The first set contains 150 samples and is used for classification benchmarking. The second set consists of 1,000 samples which are collected for a clinical trial.

can be randomly selected from a small group of input features. Further, empirical studies indicate that a feature group size of one is sufficient. As a result, an attribute is randomly selected from available attributes for each node in this investigation. Another main difference between C4.5 and random forests is that each tree in random forests is allowed to grow to its maximum size. This would not lead to data over-fitting since the overall class decision would rely on outcomes from multiple trees within the forest (Breiman, 2001).

Similar to C4.5, random forests can handle problems with discrete-valued attributes straightaway. Again, continuous-valued attributes must be split into discrete intervals during tree construction. Similar to C4.5 which uses an information gain ratio to determine the best split location on the attribute value range, Breiman et al. (1984) introduces a *Gini* index for the same task. The *Gini* index is an "impurity" measure that directly relates to the proportion of classes in a sample set. The index reaches the value of zero when only one class is present in the collection and attains the maximum value when class sizes in the collection are equal. Using the same notation for equation 3-1, the $Gini$ index of a sample set S is defined by

$$Gini(S) = \sum_{i \neq j} p(C_i, S)p(C_j, S) = 1 - \sum_{i} p(C_i, S)^2. \qquad \text{Eq.3-8}$$

The best split location on attribute A is the one that most decreases the $Gini$ index. This is achieved when

$$\Delta Gini(A, S) = Gini(S) - \frac{|S_1|}{|S|} Gini(S_1) - \frac{|S_2|}{|S|} Gini(S_2) \qquad \text{Eq.3-9}$$

is minimal. S_1 and S_2 denotes the sample subsets of S after the split and $S_1 + S_2 = S$.

3.2.6 Neural Network

A neural network is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach. The neural network that is selected for this implementation is a multilayer perceptron (Rumelhart and McClelland, 1986). The model of a neuron is illustrated in Figure 3-3(a). From Figure 3-3(a), q input signals are received by the neuron. These inputs are weighted and linearly summed together. The threshold, which can be treated as an extra connection weight, is then applied to the weighted-sum result. Thus, the linear combiner output (z) or input to the activation function is given by

$$z = \sum_{i=0}^{q} w_i u_i \qquad \text{Eq.3-10}$$

where u_i is the ith input to the neuron and w_i the connection weight on input u_i. In addition, $u_0 = -1$ and w_0 is the threshold. The neuron output ($h(z)$) is the output from the activation function and is given by

$$h(z) = \frac{1}{1 + \exp(-z)}. \qquad \text{Eq.3-11}$$

The output signal from each neuron is thus limited by a logistic sigmoid function. This described neuron model is used throughout the multilayer feed-forward network depicted in Figure 3-3(b).

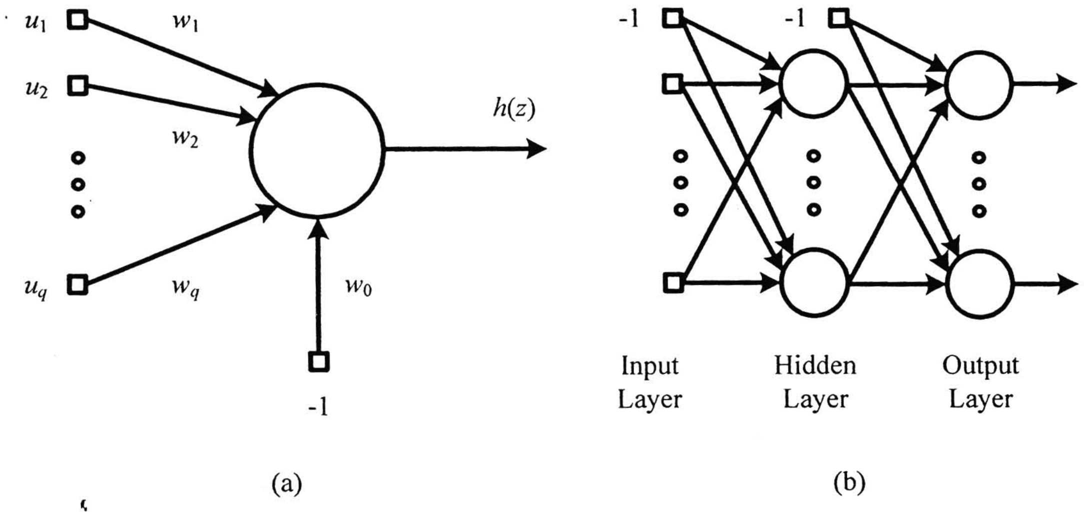


FIGURE 3-3 Schematic diagram of a multilayer perceptron: (a) computational model of a neuron and (b) feed-forward network structure that contains one hidden layer.

In this implementation, the number of network inputs is equal to the number of attributes or input features while the number of network outputs is equal to the number of output classes. Each output neuron represents one of possible output classes with the highest-valued output taken as the network prediction. This is often referred to as a 1-of-n output encoding technique (Mitchell, 1997). The network has a single hidden layer a layer of neurons that receive attributes as inputs and send signals to output neurons in which the number of neurons in the hidden layer can vary. As a result, the observation displays the effect of network non-linearity on the classification performance.

The multilayer perceptron will be trained by a back-propagation algorithm. The algorithm uses a sample-by-sample updating rule for adjusting connection weights in the network. In one algorithm iteration, a training sample is presented to the network. The signal is then fed in a forward manner through the network until the network output is obtained. The error between the actual and aimed network outputs is calculated and used to adjust the connection weights. Basically, the adjustment procedure, derived from a gradient descent method, is used to reduce the error magnitude. The procedure is firstly applied to the connection weights in the output layer, followed by the connection weights in the hidden layer. The iteration is completed after all connection weights in the network have been adjusted.

3.3 Results and Discussions

3.3.1 Attribute Discretisation and Selection

The original haemoglobin typing data contains eight attributes as shown in Table 3-1. The attributes in classifier evaluation samples have been discretised using the information-theoretic technique (Fayyad and Irani, 1993) described in chapter 2. The discrete interval of each attribute is illustrated in Table 3-3. After applying the correlation-based feature selection technique (Hall and Holmes, 2003) described in section 2.3 to the discretised attribute set, it is found that attributes 3 and 5 are redundant and can be omitted from the classification task. This decision is based on the merit scores shown in Table 3-4. These two attributes are related to unknown haemoglobin. This implies that each thalassaemia abnormality can be described by a unique combination of known types of haemoglobin.

3.3.2 Classifier Performance Evaluation

Three candidate classifiers—a multilayer perceptron, C4.5 and random forests—are benchmarked using the reduced-attribute evaluation data in stratified 10-fold cross-validation experiments (Kohavi, 1995). All three classifiers can take both continuous- and discrete-valued attributes. The best network size for multilayer perceptron is identified by varying the number of hidden nodes. The numbers of hidden nodes used in the trial are 10, 15, 20 and 25. The training process of each multilayer perceptron is terminated prior to the occurrence of data over-fitting. The appropriate number of training epochs is identified via a training and validation approach (Mitchell, 1997). A validation data set is generally employed to detect the point where the training error continues to decrease while the validation error starts to increase. This is the point where data over-fitting usually occurs. In this investigation, 10% of samples in the data set are used as the validation set.

TABLE 3-3 Discrete intervals of the attributes.

Attribute	Type of haemoglobin	Interval	% of haemoglobin
1	Hb Bart's	1	$\mathrm{Hb} \leq 1.85$
		2	$1.85 < \mathrm{Hb} \leq 4.70$
		3	$4.70 < \mathrm{Hb} \leq 7.95$
		4	$7.95 < \mathrm{Hb} \leq 10.40$
		5	$10.40 < \mathrm{Hb} \leq 23.00$
		6	$\mathrm{Hb} > 23.00$
2	Hb A$_{1c}$, Hb F	1	$\mathrm{Hb} \leq 0.35$
		2	$0.35 < \mathrm{Hb} \leq 3.50$
		3	$3.50 < \mathrm{Hb} \leq 15.25$
		4	$15.25 < \mathrm{Hb} \leq 24.32$
		5	$24.32 < \mathrm{Hb} \leq 67.75$
		6	$\mathrm{Hb} > 67.75$
3	Unknown	1	$\mathrm{Hb} \geq 0.00$
4	Hb A$_0$	1	$\mathrm{Hb} \leq 13.05$
		2	$13.05 < \mathrm{Hb} \leq 50.70$
		3	$50.70 < \mathrm{Hb} \leq 62.35$
		4	$62.35 < \mathrm{Hb} \leq 65.95$
		5	$65.95 < \mathrm{Hb} \leq 73.80$
		6	$\mathrm{Hb} > 73.80$
5	Unknown	1	$\mathrm{Hb} \leq 0.50$
		2	$\mathrm{Hb} > 0.50$
6	Hb E	1	$\mathrm{Hb} \leq 6.60$
		2	$6.60 < \mathrm{Hb} \leq 20.00$
		3	$20.00 < \mathrm{Hb} \leq 25.55$
		4	$25.55 < \mathrm{Hb} \leq 37.90$
		5	$37.90 < \mathrm{Hb} \leq 81.00$
		6	$\mathrm{Hb} > 81.00$
7	Hb A$_2$	1	$\mathrm{Hb} \leq 0.40$
		2	$0.40 < \mathrm{Hb} \leq 2.15$
		3	$2.15 < \mathrm{Hb} \leq 3.65$
		4	$\mathrm{Hb} > 3.65$
8	Hb D, Hb S, Hb Constant Spring, Hb C	1	$\mathrm{Hb} \leq 0.05$
		2	$0.05 < \mathrm{Hb} \leq 12.20$
		3	$\mathrm{Hb} > 12.20$

TABLE 3-4 Merit scores for subsets of discrete-valued attributes.

Number of attributes	Attribute subset	Merit score
1	{ Attr_6 }	0.6905
2	{ Attr_2, Attr_6 }	0.7667
3	{ Attr_2, Attr_4, Attr_6 }	0.8178
4	{ Attr_2, Attr_4, Attr_6, Attr_7 }	0.8436
5	{Attr_1, Attr_2, Attr_4, Attr_6, Attr_7 }	0.8521
6	{Attr_1, Attr_2, Attr_4, Attr_6, Attr_7, Attr_8}	0.8558
7	{Attr_1, Attr_2, Attr_4, Attr_5, Attr_6, Attr_7, Attr_8}	0.8313
8	{Attr_1, Attr_2, Attr_3, Attr_4, Attr_5, Attr_6, Attr_7, Attr_8}	0.5994

The merit scores indicate that attributes 3 and 5 are redundant and can be eliminated. Since the total number of attributes is relatively small, an exhaustive search for the best attribute subset is also carried out. The exhaustive search confirms that the displayed six-attribute subset is the optimal subset.

In contrast to parametric techniques such as neural networks, random forests avoid the occurrence of data over-fitting by constructing multiple trees for the ensemble. In this implementation, the ensemble is made up from ten trees. Although this setting is significantly less than the number of trees recommended by Breiman (2001), which is between 100 and 200, empirical studies indicate that the difference in the classification accuracy of random forests with 10 and 100 trees in this application is negligible. Similar to random forests, C4.5 also has a built-in mechanism for data over-fitting avoidance. However, since C4.5 produces a single decision tree, data over-fitting is avoided by employing a rule post-pruning strategy (Quinlan, 1993).

The classification performance of the multilayer perceptron, C4.5 and random forests on reduced-attribute data obtained from stratified 10-fold cross-validation is summarised in Table 3-5. The results from each classifier are obtained from 30 runs where new cross-validation folds are generated for each run. The results suggest that for this task, discrete-valued attributes appear to be better than continuous-valued attributes at representing the problem inputs. In addition, decision trees also have higher classification accuracy than multilayer perceptrons in which C4.5 possesses the highest performance. These results can be interpreted as follows.

Firstly, consider the results from multilayer perceptrons with different number of hidden nodes. In the case of continuous-valued attribute, an increase in the number of hidden nodes leads to a significant increase in the classification accuracy. Hence, the most suitable number of hidden nodes is 25. The statistical significance is determined via a t-test at a 95% confidence level where the results from different neural network settings are compared in a pair-wise manner. In contrast, the appropriate number of hidden nodes for the network that takes discrete-valued attributes is 15. This is because an increase in the number of hidden nodes does not produce a statistically significant change in classification accuracy after the number of hidden nodes is greater than 15. Nonetheless, further exploration of appropriate number of hidden nodes for the network with continuous-valued attributes appears to be unnecessary since a change in the classification accuracy is more driven by a transition from continuous-valued attributes to discrete-valued attributes.

The classification accuracy of decision trees is now compared. Random forests outperform C4.5 when continuous-valued attributes are used. However, C4.5 has higher classification accuracy than random forests in the case of discrete-valued attribute. Further inspection on the effect of attribute discretisation reveals that changing attribute representation can significantly improve classification performance of both neural networks and decision trees. In addition, this improvement is more prominent in C4.5 than random forests. This makes C4.5 with discrete-valued attributes the best decision tree classifier in this application.

Finally, the classification accuracy of multilayer perceptrons and C4.5 is compared. With the use of discrete-valued attributes, C4.5 outperforms the multilayer perceptron with 15 hidden nodes by 4%. This helps to confirm that the best classifier for the task at hand is C4.5. In addition to the superiority in accuracy, a single decision tree produced by C4.5 can give further insight into the relationship between attributes and output classes. The C4.5 decision tree, which is constructed using all classifier benchmarking samples, is illustrated in Figure 3-4. In Figure 3-4, the tree has the maximum depth of four in which attribute 6, which represents Hb E, is located at the root node. This attribute alone can be used to identify four classes including an EA Bart's disease, a person with both α-thalassaemia 1 trait and Hb E trait, a person with Hb E trait and a person with homozygous Hb E. A direct relationship between the Hb E attribute and three output classes among the mentioned classes is obvious. Together with the use of other attributes, a similar explanation can be deduced from the tree. It is noticed that the decision tree can be further simplified by merging a number of adjacent leave nodes that lead to the same class prediction into one node. These extra leave nodes are the result of the attribute discretisation prior to the tree construction.

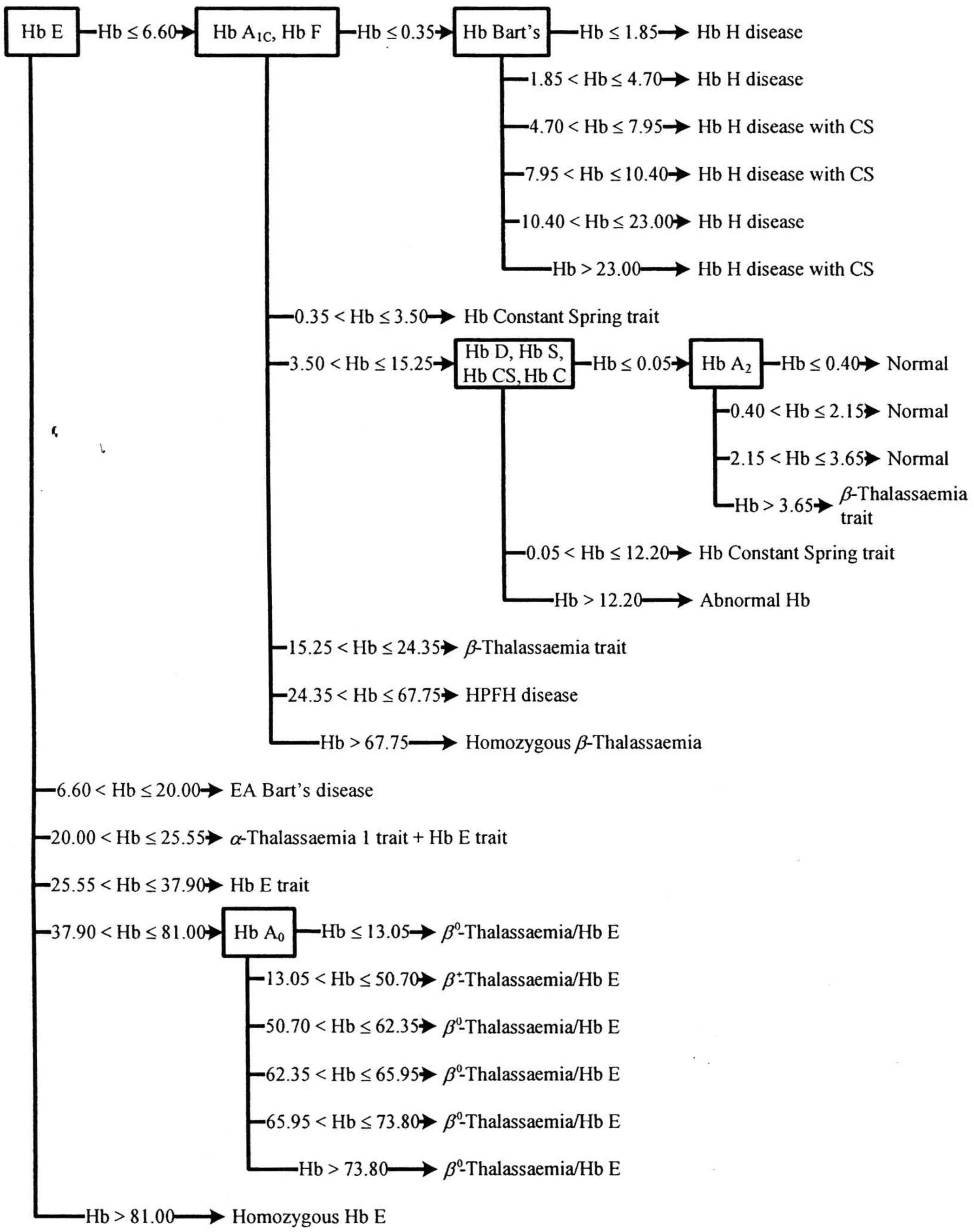


FIGURE 3-4 C4.5 decision tree which is constructed using discretised attributes. A set of screening rules can be extracted from the decision tree. For example, if the percentage of Hb E in a blood specimen is less than or equal to 6.60 while the combined percentage of Hb A$_{1C}$ and Hb F from the same specimen is between 24.35 and 67.75, then it is most likely that the specimen is taken from an HPFH patient.

TABLE 3-5 Classification performance of the multilayer perceptron, C4.5 and random forests.

Index	Attribute	C4.5	Random forests	Multilayer perceptron			
				10 nodes	15 Nodes	20 nodes	25 nodes
Accuracy (%)	Continuous	93.13 (0.82)	94.40 (1.01)	90.13 (1.17)	92.02 (1.34)	91.89 (1.08)	93.07 (1.16)
	Discrete	97.24 (0.89)	96.00 (1.15)	92.56 (0.94)	93.24 (0.69)	93.38 (0.78)	93.44 (0.76)
Sensitivity (%)	Continuous	93.13 (0.82)	94.40 (1.01)	90.13 (1.17)	92.02 (1.34)	91.89 (1.08)	93.07 (1.16)
	Discrete	97.24 (0.89)	96.00 (1.15)	92.56 (0.94)	93.24 (0.69)	93.38 (0.78)	93.44 (0.76)
Specificity (%)	Continuous	99.40 (0.10)	99.50 (0.13)	99.14 (0.13)	99.32 (0.13)	99.31 (0.13)	99.40 (0.11)
	Discrete	99.78 (0.07)	99.68 (0.11)	99.37 (0.08)	99.45 (0.10)	99.46 (0.09)	99.42 (0.08)

The results are averaged over 30 runs of stratified 10-fold cross-validation. The numbers in the brackets are standard deviations. The discretisation of attribute values leads to a statistically significant improvement in classification accuracy of all three classifiers ($p < 0.05$).

TABLE 3-6 Clustering of persons with abnormal haemoglobin, persons with thalassaemia trait, thalassaemic patients and normal subjects into three super-groups.

Minor trait/normal	Major trait	Disease
Normal subject	α-Thalassaemia 1 trait + Hb E trait	Hb H disease
Hb Constant Spring trait	Homozygous Hb E	Hb H disease with Constant Spring
Hb E trait	β-Thalassaemia trait	EA Bart's disease
	Abnormal haemoglobin	HPFH disease
		Homozygous β-Thalassaemia
		β^0-Thalassaemia/Hb E
		β^+-Thalassaemia/Hb E

Persons with abnormal haemoglobin are treated as persons with major thalassaemia trait due to their similarity in the blood characteristics.

3.3.3 Clinical Trial

In the previous sub-section, C4.5 with discrete-valued attributes is proven to be the best approach for thalassaemia classification. In addition, the decision tree generated by C4.5 can also be represented by a set of decision rules; this is convenient for diagnostic interpretation. C4.5 has subsequently been used in a clinical trial involving 1,000 samples. The distribution of classes within the sample set has been given in Table 3-2. The C4.5 decision tree illustrated in Figure 3-4 is directly applied to the clinical trial data set where classification accuracy of 93.1% (sensitivity = 93.1% and specificity = 99.5%) has been achieved. In addition, the clinical trial data set with attribute discretisation and reduction is randomly split five times into 75% for training and 25% for testing of a C4.5 decision tree. The classification accuracy of 95.0% (sensitivity = 95.0% and specificity = 99.4%) has subsequently been achieved. A higher accuracy for this latter case is to be expected since the clinical trial data are used both to train and to test the C4.5 decision tree. The classification errors can be divided into three main categories: a misclassification within the same super-group, a false prediction of high severity and a false prediction of low severity. The categorisation of errors in this manner is crucial since the problem covers multiple types of abnormality. These error categories can be explained as follows. The misclassification within the same super-group occurs when either a person with thalassaemia trait is identified as being a person with another type of thalassaemia trait, or a thalassaemic patient is misdiagnosed as being another type of patient. In this study, the persons with thalassaemia trait, thalassaemic patients and normal subjects are clustered into three super-groups based upon the severity of the exhibited thalassaemic characteristics. The details of all three super-groups are given in Table 3-6. From Table 3-6, the super-groups are made up from the minor trait/normal super-group, the major trait super-group and the disease super-group. It is noticed that normal subjects are gathered into the minor trait/normal super-group that also contains persons with Hb Constrant Spring and Hb E traits. They are grouped together since the blood characteristics of the sample within this super-group are very similar. Using the same argument, homozygous Hb E samples are placed in the same super-group as mixed α-thalassaemia 1 and Hb E trait, β-thalassaemia trait and abnormal haemoglobin samples.

The last two types of classification error are the false predictions of low and high severity. In this study, the false prediction of high severity refers to the situation when a sample is misidentified as belonging to a super-group with a higher severity of thalassaemic characteristics. On the other hand, the false prediction of low severity refers to the case when a sample is misclassified as being a member of a super-group with a lower severity of thalassaemic characteristics.

The number of misclassified samples, which are extracted from a confusion matrix, is given in Tables 3-7 and 3-8. It can be clearly seen that most of classification errors stem from samples in the minor trait/normal and major trait super-groups. The summary of classification errors from Tables 3-7 and 3-8 can be expressed as percentages as given in Table 3-9. It is noticeable that the misclassification within the disease super-group and the false predictions of severity levels for samples from thalassaemic patients are low. This agrees with the previous observation regarding the primary cause of errors.

3.4 Conclusions

In this thesis, a thalassaemia classification problem is investigated. The objective is to identify automatically whether the human subject is a person with abnormal haemoglobin, a person with thalassaemia trait, a thalassaemic patient, or a normal person using haemoglobin typing data from HPLC. The derived data sets contain eight input features or attributes and 14 distinct classes. Each attribute reflects the percentage of haemoglobin at a specific chromatographic retention time. In other words, the attribute set covers multiple types of haemoglobin. The investigation is divided into two main parts: a classifier selection and a clinical trial. Candidate classifiers for the task include a multilayer perceptron, a C4.5 decision tree and random forests.

TABLE 3-7 Detailed classification errors from applying the C4.5 decision tree in Figure 3-4 to 1,000 samples in the clinical trial.

Actual class		Identified class (Number of misclassified samples)													
		Minor/normal			Major trait				Disease						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
	1		19				2		6	2					
Normal	2						2		1	1					
	3				9										
	4		1							1	3				
Major	5													5	
	6	6	4					1		1					
	7		1	1			2				1				1
	8														
	9														
	10														
Disease	11						1								
	12														
	13														
	14														

The description of each class has been given in Table 3-2. Almost all classification errors can be traced back to samples in the minor trait/normal and major trait supergroups.

TABLE 3-8 Detailed classification errors from the C4.5 decision tree which is trained and tested with 1,000 samples in the clinical trial.

Actual class		Identified Class (Number of misclassified samples)													
		Minor/normal			Major trait				Disease						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Normal	1		4				4		2						
	2														
	3				10										
Major	4	2									3				
	5													3	
	6	5	3												
	7			1			4				1				
Disease	8	4													
	9														
	10				2										
	11	2													
	12														
	13														
	14														

The description of each class has been given in Table 3-2. The numbers of misclassified samples are displayed in a manner that they can be directly compared against the numbers in Table 3-7. Similar to the results in Table 3-7, almost all classification errors can be traced back to samples in the minor trait/normal and major trait super-groups.

TABLE 3-9 Summary of classification errors from the clinical trial.

Type of classification error	Classification error (%)	
	Table 3-7	Table 3-8
Misclassification within the same super-group	2.2	0.8
Misclassification within the minor trait/normal super-group	1.9	0.4
Misclassification within the major trait super-group	0.3	0.4
Misclassification within the disease super-group	0.0	0.0
False prediction of high severity	3.3	2.3
Minor trait/normal identified as major trait	1.1	1.4
Major trait identified as disease	1.2	0.7
Minor trait/normal identified as disease	1.0	0.2
False prediction of low severity	1.4	1.9
Disease identified as major trait	0.1	0.2
Major trait identified as minor trait/normal	1.3	1.1
Disease identified as minor trait/normal	0.0	0.6
Total	6.9	5.0

The errors are described in terms of misclassification within the same super-group, false prediction of high severity and false prediction of low severity.

The study involving stratified 10-fold cross-validation reveals that C4.5 is the most suitable classifier for the data that have been pre-processed by attribute discretisation and reduction. Subsequently, C4.5 is applied in the clinical trial and further analysis of the classification error indicates that the misclassification among disease classes and the false predictions of severity levels for samples from thalassaemic patients are low. This helps emphasise the suitability of C4.5 as an automated thalassaemic classification tool.

In order for the proposed automated classification procedure for thalassaemia screening to be applicable in clinical settings, a larger clinical trial may still be required. However, since C4.5 is the chosen classifier, an additional trial can be easily carried out. This is because a set of decision rules can be extracted from the tree illustrated in Figure 3-4. The decision rules can subsequently be implemented in many user-friendly computer programs including spreadsheets and databases. This is an additional advantage of using C4.5 over a multilayer perceptron and random forests.

With the availability of knowledge regarding the relationship between haemoglobin typing inputs and thalassaemic class outputs and that between CBC inputs and similar outputs (Wongseree et al., 2007), the most obvious further study is to employ both types of inputs in the classification task. Since CBC and haemoglobin typing are always carried out for a laboratory diagnosis of thalassaemia, it is not difficult to acquire both types of data from the same blood specimen. In addition to thalassaemia classification, another possible further work is to apply the procedure for extracting informative features from chromatograms as described in this thesis to other pattern recognition problems. Examples of classification problems that involves the inspection of chromatograms include a diagnosis of liver and bile diseases (Zhao et al., 1999), determination of herb's origins (Chuang et al., 2007), differentiation of tea varieties (Alcazar et al., 2007) and ink identification for forensic purposes (Kher et al., 2006). These examples illustrate that a wide range of chemometric applications can benefit from the feature extraction procedure explained in this investigation.