THE UNIVERSITY OF HULL

Predicting Cardiovascular Risks using Pattern Recognition and Data Mining

being a Thesis submitted for the Degree of Ph.D

in the University of Hull

by

Thuy Thi Thu Nguyen

August 2009

Acknowledgements

I would like to express gratitude to many people who, in a variety of ways, have helped with the completion of this thesis. Firstly and foremost, I would like to thank my parents and my family, Trung and Dung, for their continued supports and love during my time at Hull University and for helping me to maintain the will to succeed. Secondly, I would like to thank Vietnamese Government in particular to Education and Training Ministry (322 project) for providing the financial supports during my period of study.

I would like to express the special thanks to Darryl N. Davis for his helps in the role of a supervisor from the first day of my research. Darryl encouraged me to achieve to the best of my ability, and he also had to work hard with my poor English writing. Without Darryl, I could not reach to this stage.

I would like to thank my other supervisors: Dr Chandra Kambhampati and Dr Len Bottaci in the role of supervisory panel in the way of my research.

I would like to end by saying thank you to a handful people. Thanks to The Vietnamese Commercial University in particular to Informatics department for their supports in my administration procedures, and their friendship. Thanks to the members of Computer Science for their assistance, suggestions, and criticisms. Finally, thanks also to my friends, who are beside me during the up and down period.

Declaration of Publications

The following papers have been published or accepted for publication during the course of this research and include additional work to the material presented in this thesis.

Davis, N.D., and Thuy, N.T.T (2008). Data Mining and Medical Knowledge Management: Cases and Applications (Book chapter). Editors: Petr Berka, Jan Rauch, & Djamel Abdelkader Zighed, IGI Global Inc. (Accepted)

Thuy N.T.T, and Davis, N. D. (2007). Clustering and Predicting CardioVascular Risk. Proc. In International Conference of Data Mining and Knowledge Engineering, London.

Thuy N.T.T, and Davis, N. D. (2006). Feature Selection and Predicting CardioVascular Risk. Conference of Clinical BioScience Institute, Hull University.

Thuy N.T.T, and Davis, N. D. (2006). Predicting CardioVascular Risks Using POSSUM, PPOSSUM and Neural Net Techniques. Proc. In 8th International Conference on Enterprise Information Systems ICEIS, Paphos, Cyprus.

Thuy N.T.T, and Davis, N. D. (2005). Predicting CardioVascular Risk Using Neural Net Techniques. Poster in conference of Clinical BioScience Institute, Hull University.

Thuy N.T.T (2004). Using Neural Network for cardiovascular data. Workshop in Bradford University.

Contents

List of Tables

List of Figures

Abbreviations

NN(s): Neural Network(s) or Artificial Neural Network.

POSSUM: The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity

PPOSSUM: Portsmouth POSSUM

MLP: MultiLayer Perceptron

RBF: Radial Basis Function

SVM: Support Vector Machine

SOM: Self Organizing Maps

MI: Mutual Information

KMIX: A clustering algorithm derived from Kmeans

WKMIX: A clustering algorithm derived from KMIX

SyntPR: Syntactic pattern recognition

Stat PR: Statistical Pattern, or Statistical Pattern Recognition

ACC: Accuracy

Sen: Sensitivity

Spec: Specificity

PPV: Positive predictive value

NPV: Negative predictive value

Abstract

This thesis presents the use of pattern recognition and data mining techniques into risk prediction models in the clinical domain of cardiovascular medicine. The data is modelled and classified by using a number of alternative pattern recognition and data mining techniques in both supervised and unsupervised learning methods. Specific investigated techniques include multilayer perceptrons, radial basis functions, and support vector machines for supervised classification, and self organizing maps, KMIX and WKMIX algorithms for unsupervised clustering. The Physiological and Operative Severity Score for enUmeration of Mortality and morbidity (POSSUM), and Portsmouth POSSUM (PPOSSUM) are introduced as the risk scoring systems used in British surgery, which provide a tool for predicting risk adjustment and comparative audit. These systems could not detect all possible interactions between predictor variables whereas these may be possible through the use of pattern recognition techniques. The thesis presents KMIX and WKMIX as an improvement of the K-means algorithm; both use Euclidean and Hamming distances to measure the dissimilarity between patterns and their centres. The WKMIX is improved over the KMIX algorithm, and utilises attribute weights derived from mutual information values calculated based on a combination of Baye's theorem, the entropy, and Kullback Leibler divergence.

The research in this thesis suggests that a decision support system, for cardiovascular medicine, can be built utilising the studied risk prediction models and pattern recognition techniques. The same may be true for other medical domains.

Chapter 1

Introduction

This thesis presents an investigation into using pattern recognition and data mining techniques to produce and verify risk prediction models in medicine, in particular in the cardiovascular domain. The necessity for using pattern recognition and data mining techniques to develop and improve risk models arises from the need for clinicians to improve their prediction models for individual patients. This thesis focuses on the task of data modelling and classification using supervised and unsupervised techniques from pattern recognition and data mining. The term data modelling is used here to refer to the filtering of data according to clinical heuristic rules in order to produce the expected outcome set. A particular focus is the use of neural networks, which are being used with a great frequency and success in medical domains. In developing a framework for modelling the given medical data, the research proposes the use of clustering methods to generate new predictive models for use in the cardiovascular domain.

There are popular medical scoring systems used for risk assessment in Britain, namely POSSUM and PPOSSUM (Copeland et al, 1991; Copeland, 2002; and Prytherch et al, 1998). They can produce individual risk scores for patients. Like other logistic regression systems, although the outcome is not derived from linear calculations, the POSSUM and PPOSSUM have linear combinations of variables in the input set. Therefore, according to Tu (1996), these systems are not adept at modelling nonlinear complex interactions in medical domains. Also according to Tu (1996), pattern recognition and data mining, in particular neural network techniques, offer the ability to implicitly detect complex

nonlinear relationships between dependent and independent variables, and the ability to detect all possible interactions between predictor variables. Furthermore, according to Lisboa (2002); and Jain et al (2000), neural network techniques are useful tools for patient diagnosis and prognosis. Therefore, neural network techniques (e.g. multilayer perceptron; radial basis function; and support vector machine) can be seen as candidates for use with the thesis data. A substantive portion of the thesis concentrates on the prediction of cardiovascular risk through the use of clustering techniques such as self organizing maps and the KMIX algorithm. These are regarded as unsupervised pattern recognition techniques. These techniques can help to discover the internal data structure in order to verify the nature of the problems or the difficulty of measuring influential parameters. The selection of domain attributes is another issue in building risk prediction models. By using feature selection methods, in particular mutual information (Shannon, 1948; Kullback and Leibler, 1951), a ranking of domain attributes for risk prediction models can be produced. These features might then be used as attribute weights in the classification process.

1.1. Risk Assessment in Medical Domains

Medical decision support systems are designed, and implemented, to support clinicians in their diagnosis. They typically work through an analysis of medical data and a knowledge base of clinical expertise. The quality of medical diagnostic decisions can be increased by improvements to these systems. Some medical decision support systems are in popular use, such as MYCIN (Shortliffe, 1976); INTERNIST/QMR (Miller et al, 1982); the Framingham study (Framingham Heart Study, 1948); as well as scoring risk systems such as the Consortium for South Eastern Hypertension Control (COSEHC, 2003; Hawkins et al, 2005) and the linear scoring system (Gupta et al, 2005).

The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM), first used by Copeland et al (1991), has been applied to predict individual risk for general surgical patients. This system, which is widely used in Britain, has been devised from both a retrospective and prospective analysis by using 12 factors over 4 grades of physiological scores and 6 factors of operative severity scores. There is another model based on POSSUM, namely Portsmouth POSSUM (PPOSSUM- Prytherch et al, 1998). This was produced because of a claim that POSSUM over predicted death, and in particular with "Low risk" patients (Prytherch et al, 1998). Therefore, the original POSSUM equation was modified leading to the Portsmouth predictor equation for mortality (PPOSSUM), which still utilises the same physiological and operative variables. Given no gold standard for predicting cardiovascular risk, this thesis sets a reference level for risk assessment using the POSSUM and PPOSSUM systems. Their performance is measured by comparison ratios between predicted mortality of all patients and observed dead patients. The POSSUM and PPOSSUM systems can produce individual risks using logistic regression calculations. However, not only are there disadvantages as indicated above, they are also ambiguous in the interpretation of categorical risks over the risk scale. By directly predicting alternative risk categories from a classification process, pattern recognition and data mining techniques might help in providing more reliable diagnoses.

1.2. Pattern Recognition and Data Mining

Pattern recognition (Bishop, 1995; Ripley, 1996) is of interest in many areas such as statistics, probability theory, machine learning, and medicine. The usefulness of pattern recognition is recorded in being able to perform highly sophisticated tasks such as recognising a face, medical diagnosis, and so on. According to Bishop (1995), there are

four approaches that can be applied to medical diagnosis field: template matching; statistical classification; syntactic matching; and neural networks.

The design of neural network was originally motivated by the phenomena of learning and recognition (Hebb,1949; Parisi, 1986; Fausett, 1994; Cross et al, 1995; Haykin, 1999). The neural network techniques can be divided into two alternative ways of learning: supervised and unsupervised. In this thesis, both these approaches are applied to the cardiovascular domain, with the objective of determining the meaningful distinction of alternative outcomes in clinical risk prediction models. The motivation for using these techniques can be seen in their contributions to medical applications so far (see detail in Chapter 3).

Data mining typically deals with data that has been collected for other purposes. In data mining, the emphasis is on the analysis of data sets to find unsuspected relationships and the modelling of the data in novel ways that are both understandable and useful to the data owner. A data mining methodology provides the framework for the processing of the data. Here, aspects of data mining, such as cleaning and filtering, are used in combination with pattern recognition techniques in order to investigate alternative prediction risk models for the cardiovascular domain. Feature selection methods from data mining are used to reduce the attribute-value capacity of the feature set and data set. This eliminates the redundant features without losing the significant characteristics of the data domain.

1.3. Aims and Objectives

This thesis is motivated by the growing interest in using pattern recognition and data mining prediction techniques, such as neural networks, in medical areas (Baxt, 1991; Harrison et al, 1991; Baxt, 1992; Wilson et al, 1995; Weingart et al, 2000; Barach and Small, 2000; Reason, 2000; Lisboa et al, 2000; Jain et al, 2000; Lisboa, 2002; Plaff et al,

2004). The supplied medical data used for assessing risk is, itself, inconsistent over a history of patients at any one clinical site, and not always immediately useable. The motivation for the work in this thesis is in the realization for the necessity of directing attention to the production of risk prediction models for the cardiovascular domain. A specific aim for this thesis is to find solutions for this problem using both supervised and unsupervised pattern recognition and data mining techniques (see Chapter 4). Alternative risk predictions based on the data domain and medical expert knowledge will be compared to other predictive systems such as the POSSUM and PPOSSUM. These alternative risk prediction models are verified using neural network and clustering techniques.

Objectives

This thesis will aim to answer and explore a number of key questions pertinent to the application of pattern recognition and data mining to risk assessment in medical domains. They are:

1. How able are the existing systems in dealing with risk prediction for patients?
2. Are linear model adequate for use with the data domain?
3. What are the different ways to classify the data?
4. Which method of clustering data is appropriate for this medical domain?
5. Can the attribute set be decreased by defining the significant attributes in the data domain?

The objectives for this research can be summarized as follows:

- Demonstrate the existing risk assessment systems, and their limitations.
- Show the use of using pattern recognition and data mining classifications instead of the POSSUM and PPOSSUM systems.

- Show the advantage of using nonlinear models for the classification compared to the use of linear models.
- Produce a data mining methodology of use in applying pattern recognition and data mining to the cardiovascular data.
- Improve the K-means algorithm for a medical data domain, in particular the cardiovascular data, so that it allows multiple data types.
- Investigate alternative risk prediction models using both supervised and unsupervised pattern recognition techniques.
- Investigate the use of mutual information for the reduction of feature selection in data domain.
- Combine the mutual information method and pattern recognition theory for use with a clustering algorithm

1.4. Thesis Structure

The research questions stated above will be dealt with in the next eight chapters of this thesis. Risk assessment in the medical domain is discussed in C**hapter 2**. The popular scoring systems of POSSUM and PPOSSUM are described. This chapter also discuss the disadvantages of these systems as well as other risk assessment systems. C**hapter 3** provides a general background for the thesis. The general theory of pattern recognition is presented, and a detailed literature review of pattern recognition techniques is given. The standard classification measures, which are used for all thesis experiments, are introduced in detail. This chapter also compares linear and nonlinear models in classification problems. **Chapter 4** provides an in-depth investigation of supervised neural network techniques for use in a medical domain, in particular the use of multilayer perceptrons, radial basis

functions, and support vector machines. This chapter also provides the detail of unsupervised pattern recognition techniques as self organizing maps and KMIX, an instance of the K-means algorithm, which can deal with the mixed data types presented in the cardiovascular domain. **Chapter 5** presents the general theory about data mining and methodologies for performing it. The thesis methodology and its framework are shown in this chapter. Information about cardiovascular data domain and the detailed data preparation steps for the thesis experiments, as an application of a data mining methodology, are represented in this chapter. The given data from two clinical sites is analysed and summarized, and the strategy for data preparation, and as used in all thesis experiments, is presented in detail. **Chapter 6** introduces alternative models used as the main thesis experiments in the following chapter. This chapter also demonstrates all thesis case studies for the use of the POSSUM and PPOSSUM systems, supervised and unsupervised pattern recognition techniques with some of the thesis data. **Chapter 7** shows the results and the analysis of these results using standard measures. Feature selection and mutual information is introduced in **Chapter 8.** A combination of Bayes' theory and mutual information is applied in the KMIX clustering algorithm to increase its ability to deal effectively with the cardiovascular data**. Chapter 9** unites the work of the previous chapters in a practical setting. The thesis concludes in this chapter with an analysis and a discussion of the research outlined in the previous chapters. The results from the case studies and thesis experiments using the various risk prediction models are revisited in light of the research questions stated in **Chapter 1**. The final chapter ends with conclusions and suggestions for future work and possible extensions to the research outlined in this thesis.

Chapter 2

Risk Assessment in Medical Domains

2.1. Introduction

Of all the modern technological quests, the search to create artificially intelligent computer systems has been one of the most ambitious and, not surprisingly, controversial, particularly the recent application of artificial intelligence to decision-making areas of medicine (Coiera, 2003). Medical decision support systems or clinical decision support systems play an increasingly important role in medical practice (Marckmann, 2001). They are applied to broad areas of decision-making by clinicians to support their diagnosis based on medical data and domain knowledge. According to Coiera (2003), artificially intelligent computer systems, which are able to store and process vast stores of knowledge, might ably assist clinicians with tasks such as diagnosis, and prediction the patient risk.

This chapter introduces some popular risk assessment and artificially intelligent diagnostic systems such as MYCIN (Shortliffe, 1976); Internist/QMR (Internist/Quick Medical Reference, Miller et al, 1982); the Framingham study (Framingham Heart Study, 1948); the Australian Busselton study (Knuiman et al, 1998); and the German PROCAM study (German Prospective Cardiovascular Münster, Assmann et al, 2002). These systems can be seen as a background for the thesis analysis on risk assessment systems. Some other scoring risk systems such as the Consortium for Southeastern Hypertension Control (COSEHC, 2003; Hawkins et al, 2005) and the linear scoring system (Gupta et al, 2005) are also introduced.

Clinical risk assessment systems such as INdividual Data ANalysis of Antihypertensive intervention trials (INDANA, Pocock et al, 2001), the Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM, Copeland et al, 1991) and the Portsmouth POSSUM (PPOSSUM, Prytherch, 1998) are introduced and discussed in greater detail within this chapter.

2.2. Risk Assessment Systems

The intelligent medical system MYCIN (Shortliffe, 1976) is one of earliest expert systems. It was designed and implemented at Stanford University in the 1970s with the purpose of diagnosing and recommending treatment for certain blood infections. This rule-based expert system is comprised of two major components as follows:

- A knowledge base that stores the information of the domain expert.
- An inference engine to derive knowledge from the presently known knowledge in the first component (Lisboa, 2002).

One of its disadvantages is the time taken to make a decision (Lisboa, 2002). Subsequently, other systems [based on MYCIN such as EMYCIN (Melle, 1979), and PUFF (Aikins et al, 1983)] were introduced with improvements that speed up the decision-making time. These improvements are not discussed in this thesis. More detail can be seen in Lisboa (2002) and Coiera (2003).

Another medical decision system is Internist/QMR (Miller et al, 1982), with a knowledge base that has 956 hypotheses, which works based on diagnostic strategies such as scoring function, partitioning, or questioning. According to Shortliffe et al (1990), the Internist/QMR mirrors hypothetic-deductive reasoning. It can handle coexistent diseases and is remarkably accurate from the start. Hence, it will make strong assumptions about

mutual independence of predictive variables. However, according to Lowe (2003), it was difficult to trace the system's recommendations. Subsequently, its results, and assumptions, were hard to explain to clinicians.

Directing attention to cardiovascular disease, the popular studies are as follows:

- Framingham study (Framingham Heart study, 1948);
- Australian Busselton study (Knuiman et al, 1998);
- and the German PROCAM study (Assmann et al, 2002).

The Framingham study emphasises the relationships between cardiovascular disease and other attributes such as altered blood lipid, blood pressure, body weight, and so on. The detailed review can be seen in Gueli et al (2005). The Busselton study (Knuiman et al, 1998) used epidemiological data from over 8000 patients collected from 1966 to 1981. This model uses logistic regression to predict the 10-year risk of coronary heart disease. Lastly, the German PROCAM study (Assmann et al, 2002) used the epidemiological data from 25000 patients collected from 1979 to 1985. Its logistic model predicts the 8-year risk of a cardiovascular event for patients. Both the Busselton model and German PROCAM model are improved by Twardy et al (2005). They built Bayesian networks using both models as the knowledge engineering component. Cross-validation was then used to evaluate these models. Twardy et al (2005) achieved the same risk results as the use of logistic regression in the original studies. However, they pointed out that their models are easier to interpret, giving a more intuitive causal story of coronary heart disease risk than both original models. The detail about the Busselton model, the German PROCAM model, as well as the use of Bayesian networks for these models can be seen in Twardy et al (2005).

There also exists scoring risk systems, such as the Consortium for Southeastern Hypertension Control (COSEHC, 2003; Hawkins et al, 2005), used to predict an individual's 5-year mortality risk for cardiovascular patients, or a linear scoring system (Gupta et al, 2005) to identify the cardiovascular mortality risk for renal transplant patients. The COSEHC used 17 risk factors divided into three groups as: *non-modifiable; modifiable;* and *non-traditional.* This system produces individual patient risk depending on "5-year cardiovascular mortality risk". For example, individual risk is identified as *"High absolute risk"* when its *"5 year mortality"* risk is exceeding 2.5%. More detail about this system can be seen in Haukins et al (2005). Gupta et al (2005) applied a cardiac risk assessment system devised by one of the authors (Ward, 2005) to identify the cardiovascular mortality risk for renal transplant patients. Four risk groups were identified according to the scores: *Low* (0-4); *Medium* (4-8); *High* (8-12); and *Very High* (>12). The results are then separated by two groups of renal transplant patients. The first group contains patients who subsequently die, and the second contains patients who survive after a renal transplant. Gupta et al (2005) pointed out that the deceased groups had significantly greater cardiovascular scores than the other.

The INdividual Data ANalysis of Antihypertensive (INDANA) intervention study (Pocock et al, 2001) used 47088 cardiovascular cases with 8 randomised controlled trials of antihypertensive treatment to assess the 5-year mortality risk of patients. The risk score is based on 11 factors such as age, sex, diabetes, stroke, heart disease, and so on (see detail in Table 2.1). By using a multivariate Cox model (Cox, 1972; Bennet, 2006) with cardiovascular death as the outcome, an individual score is predicted. This score is then compared to the average risk score in the same age and sex group to assign an appropriate risk for the patient.

Attribute	Example Value	Attribute	Example Value
Sex	Female	*Diabetes*	No
Age	55	*Myocardial Infarction*	No
Smoker	No	*Stroke*	No
Blood Pressure	100mm Hg.	*Left Ventricular Hypertrophy*	No
Cholesterol	4.5 mmol/l.	*Height*	160 cm.
Creatinine	N/A		

Table 2.1: The 11 factors used in the INDANA trial (Pocock et al, 2001) with example values.

For example, a woman's profile can be seen in Table 2.1, for which the system produced a risk score of 19.83. This is compared to the average of risk score of 30.66 ("*High")* in the age range (55-59). The system therefore concludes this woman's risk is "*Low"*. If the value of woman's smoking is changed to "*Yes"*, the risk score will be 30.3. Her risk status is now *"Medium"*. Alternatively, if the stroke's value is changed to "*Yes"*, the risk score increases to 27.91 (less than 30.3 - when Smoker's value is "*Yes"*). The system would then conclude her risk as "*Low"*. From this point, it seems that this system places too much emphasis whether a patient smokes rather than other significant symptoms which can lead to the death of a cardiovascular patient, such as "*stroke"* and "*history of myocardial infraction"* (Kuhan et al, 2001).

The advantage of this system is that it considers a range of personal factors instead of focusing on only treatment (e.g. drugs) and controlling blood pressure (Pocock et al, 2001). However, the individual risk is inferred from the average risk score of the same age and sex group. This can lead to ambiguous interpretations of the status of values such as "*Low", "Medium", "High",* and *"Very High"* in the risk scale. For example, the final score indicated above is 27.91 (when the stroke's value is "*Yes")* labelled as "*Low"*, whereas a nearby risk score (30.3 - when Smoker's value is "*Yes"*) gives the result of "*Medium"*. This score is also quite close to the "*High"* threshold (of 30.66). The same can be said for any

prediction model that uses fixed thresholds over a numeric scale to determine a categorical output.

2.3. POSSUM and PPOSSUM

This section concentrates in depth on the Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM, Copeland et al, 1991) and the Portsmouth-POSSUM (PPOSSUM, Whitley et al, 1996) systems. These systems are used broadly in the Britain for general surgical patients, and particularly with cardiovascular patients (Wijesinghe et al, 1998; Kuhan et al, 2001). The POSSUM and PPOSSUM systems predict mortality, morbidity, and death rate risks for patients based on the scoring system. Furthermore, these tools can compare the predicted deaths and the actual deaths in various ranges (bands) of patient risks. A case study in Chapter 6 will provides a focus for the discussion about these systems.

The POSSUM system was first used by Copeland et al (1991). It has been applied to predict outcomes for general surgical patients. This system has been devised from both retrospective and prospective analyses. The key factor of the POSSUM system is the prediction of individual mortality and morbidity risks based on physiological and operative severity scores. The POSSUM system is built based on an original assessment of 48 physiological factors, and 14 operative and postoperative factors. By using multivariate analysis techniques, these factors were reduced to 12 physiological and 6 operative factors in two parts of physiological assessment and operative severity. In physiological assessment, the following 12 variables are used in the scoring system:

- Age
- Cardiac signs

- Respiratory signs
- Systolic blood pressure
- Pulse
- Coma score
- Serum urea
- Serum sodium
- Serum potassium
- Haemoglobin
- White cell count
- ECG

The 6 variables used in the second part of system (operative factors) are:

- Operative magnitude
- Number of operations within 30 days
- Blood loss
- Peritoneal contamination
- Presence of malignancy
- Timing of operation

The coefficients for the risk factor were divided by a constant and rounded number such as 1, 2, 4, 8, and so on. Tables 2.2 and 2.3 below show the scores of physiological and operative severity for the POSSUM system.

Score	1	2	4	8
Age	<=60	61-70	>=71	…
Cardiac signs *CXR*	Normal Normal	Cardiac drugs or steroids ….	Edema; Wafarin Borderline Cardiomegaly	JVP Cardiomegaly
Respiratory signs *CXR*	Normal Normal	SOB exertion Mild COAD	SOB stairs Mod COAD	SOB rest Any other change
Systolic BP mmHg	110-130	131-170 100-109	>=171 90-99	<=89
Pulse, beats/min	50-80	81-100 40-49	101-120	>=121 <=39
Coma score	15	12-14	9-11	<=8
Urea nitrogen, mmol/L	<7.5	7.6-10	10.1-15	>=15.1
Na,mEq/L	>136	131-135	126-130	<=125
K,mEq/L	3.5-5	3.2-3.4 5.1-5.3	2.9-3.1 5.4-5.9	<=2.8 >=6
Hb, g/dL	13-16	11.5-12.9 16.1-17	10-11.4 17.1-18	<=9.9 >=18.1
WCCx10^{12}/L	4-10	10.1-20 3.1-3.9	>=20.1 <=3	…
ECG	Normal	…	AF(60-90)	Any other change

Table 2.2: Physiological Score (Copeland et al, 1991).

Score	1	2	4	8
Operative magnitude	*Minor*	*Intermediate*	*Major*	*Major+*
No. of operations within 30d	1		2	>2
Blood loss per operation, mL	<100	101-500	501-999	>1000
Peritoneal contamination	No	Serious	Local pus	Free bowel content, pus or blood
Presence of malignancy	No	Primary cancer only	Node metastases	Distant metastases
Timing of operation	Elective		Emergency resuscitation possible, operation<24h	Emergency immediate, operation <2h

Table 2.3: Operative Severity Score (Copeland et al, 1991).

The mechanism to produce the physiological and operative severity scores can be seen in the example 2.1 below.

Example 2.1

Assume that data information is given in Table 2.4 and Table 2.5 below. The physiological (PS) and operative severity (OS) scores are calculated as the sum of respectively attributed scores arrived from Tables 2.2 and 2.3.

Age	*Cardiac signs*	*Respirat-ory signs*	*Systo-lic BP*	*Pulse*	*Coma score*	*Urea*	*Na,mEq/L*	*K, mEq/L*	*Hb, g/dL*	*WCC x10^{12}/L*	*ECG*	***PS Score***
80	JVP	Mild COAD	132	85	15	10.5	140	3.4	10	5.2	Normal	32
50	JVP	Normal	115	85	15	10.5	140	3.4	10	5.2	Normal	27
45	Normal	Normal	115	85	15	7.8	140	3.4	10	5.2	Normal	18
52	Normal	Normal	115	85	15	7	140	3.4	10	5.2	Normal	17
50	steroids	Normal	115	85	15	7	140	3.4	12	5.2	Normal	16

Table 2.4: An example of PS score calculations.

Operative magnitude	*No. operat -ions within 30d*	*Blood loss per operation, mL*	*Peritoneal contaminat -ion*	*Presence of malignancy*	*Timing of operation*	***OS Score***
Minor	2	120	Serious	Node metastase	Elective	14
Major	1	120	Serious	Node metastase	Elective	14
Intermediate	1	520	Serious	No	Elective	11
Major	1	200	Serious	No	Operation <2h	18
Major	1	200	No	No	Elective	10

Table 2.5: An example of OS score calculations.

Given the ratings in Table 2.2, and Table 2.3, the values for entries in Table 2.4, and Table 2.5 can be mapped to an integer scores and summed.

For example, the first entry (PS score) in Table 2.4 is given by:

$$4 + 8 + 2 + 2 + 2 + 1 + 4 + 1 + 2 + 4 + 1 + 1 = 32.$$

The first OS score in Table 2.5 is calculated as:

$$1 + 4 + 2 + 2 + 4 + 1 = 14.$$

The mortality and morbidity risks, based on the physiological and operative severity scores from Tables 2.2 and 2.3, are calculated for each patient. These rates are referred to respective models in chapter 7 (section 7.2.4), where they are the expected outcomes. A logistic regression analysis was performed to yield the equations for the mortality and the morbidity risks.

Mortality rate

Mortality rate is the rate (percentage) of the number of "Very High risk" patients (be called death patients) predicted by the system. Its equation can be calculated as:

$$R1 = 1 / (1 + e^{-z}) \qquad (2.1)$$

where $z = -7.04 + (0.13 * PS) + (0.16 * OS)$; $R1$ is mortality risk; PS is the Physiological Score; and OS is the Operative severity Score.

Morbidity rate

Morbidity rate is the rate (percentage) of the number of "High risk" patients predicted by the system. Its equation can be calculated as:

$$R2 = 1 / (1 + e^{-z}) \qquad (2.2)$$

where $z = -5.91 + (0.16 * PS) + (0.19 * OS)$; and $R2$ is morbidity risk.

According to Wijesinghe et al (1998); Prytherch et al (1998); and Midwinter et al (1999), the POSSUM system over-predicts deaths for cardiovascular patients, especially for the "Low risk" patients. In an effort to counteract the perceived shortcomings of the conventional POSSUM, Whitley et al (1996) devised a new version of POSSUM, called PPOSSUM (Portsmouth Predictor equation for mortality). The PPOSSUM equations were derived from a heterogeneous general surgical population.

In PPOSSUM, the predicted **Death rate** is instead of **Morbidity rate**. This rate is also referred to the "Death rate" model in chapter 7 (section 7.2.4).

Death rate

Death rate is the rate (percentage) of the number of death patients predicted by the system. The Death rate is given by:

$$R3 = 1 / (1 + e^{-z}) \qquad (2.3)$$

where $z = -9.37 + (0.1692 * PS) + (0.150 * OS)$.

The use of the POSSUM and the PPOSSUM calculations can be seen detail in the Example 2.1 below.

Example 2.2

Assume that data scoring information is given in Table 2.6 below. The z values in the logistic equations, labelled as "z1" and "z2", are calculated as follows:

$$z1 = -7.04 + 0.13 * PS + 0.16 * OS$$

$$z2 = -9.37 + 0.169 * PS + 0.15 * OS$$

Assume that R1 and R3 are the "Mortality rate" and "Death rate" of the POSSUM and PPOSSUM systems. They are calculated as follows:

$$R1 = \frac{1}{1 + e^{-z1}}$$

$$R3 = \frac{1}{1 + e^{-z2}}$$

Reg.No	PATIENT_STATUS	PhysiolScore	OpSevScore	z1	z2	R1	R3
006330	Alive	36	14	-0.12	-0.43	88.69	65.05
007931	Alive	27	16	-0.97	-1.84	37.90	15.88
013384	Dead	18	14	-2.46	-3.85	8.54	2.12
017888	Alive	17	14	-2.59	-4.04	7.50	1.75
007931	Alive	16	14	-2.72	-4.23	6.58	1.45
009912	Dead	16	14	-2.72	-4.23	6.58	1.45
017888	Alive	15	14	-2.85	-4.42	5.78	1.20

Table 2.6: An example of POSSUM and PPOSSUM calculation using PS and OS scores.

Obviously, from Table 2.6, whenever the POSSUM system predicts a patient with a high R1 score, this patient will be predicted with high R3 score in the PPOSSUM system as well. For example, a patient, with highest R1 risk at 88.69, will have highest R3 risk at 65.05.

Various risk bands are produced from the individual mortality, morbidity, and death rate risks. Each band denotes patients with the mortality, morbidity or death rate in the same range. The number of predicted mortality is then calculated for each band. These predicted

numbers are compared to the actual deaths to evaluate the system performance. For instance, assume that the POSSUM results in Table 2.6 are divided into two bands of "0-30%", and ">30%" (R1 results). Therefore, the first band of "0-30%" has 5 patients, and the second band has 2 patients. Table 2.7 below shows a comparison of predicted deaths and actual deaths of the POSSUM system for the data in Example 2.2.

According to Wijesinghe et al (1998), a ratio of 1.0 indicates that the scoring system predicts the same as actual deaths; greater than 1.0 means the scoring system under-predicts deaths whereas a ratio less than 1.0 means over-prediction of deaths. Therefore, in the Example 2.2, the POSSUM system over-predicts deaths in the band of ">30%" (ratio of 0 in Table 2.7). The band of "0-30%" under-predicts with 0 predicted deaths, although the ratio cannot be calculated. Overall, the POSSUM system under-predicts deaths compared to actual deaths as shown with the band 0-100%, and the prediction ratio of 2.0.

Range of predicted rate	**Mean predicted risk of Mortality (%)**	**No of operations**	**Predicted deaths**	**Reported deaths**	**The ratio**
0-30%	6.99%	5	0	2	N/A
>30%	63.29%	2	1	0	0
0-100%	23.08%	7	1	2	2.0

Table 2.7: Comparison of observed and predicted death from POSSUM logistic equations.

The argument against the POSSUM and PPOSSUM systems is the predictive accuracy of assessment. Prytherch et al (1998) complained that the POSSUM over-predicted mortality especially for "Low risk" groups. Wijesinghe et al (1998) pointed out that the predictive accuracy for the POSSUM and the PPOSSUM might be better if a correct analysis is used. According to Wijesinghe et al (1998), there are two methods of analysis for the predicted deaths: "linear" and "exponential". Note that the latter is not given a clear description in the

original POSSUM publications (Yii and Ng, 2002). In the linear method, patients were divided into risk groups of 10%, 20%, and so on whereas in the exponential method the cut-off value was used to group patients in a larger range. For example, patients in the range of 50-100% of predicted death can be in one group with a 50% cut-off value. The detail for these methods can be seen in Wijesinghe et al (1998).

2.4. Discussion

The Internist/QMR system (Miller et al, 1982) is based on hypothetic-deductive reasoning. Therefore, it might be useful for the clinicians except its results are difficult to explain. The MYCIN (Shortliffe, 1976) is a rule-based system, but it is limited by its design for patients with blood infections. To be of use for the cardiovascular domain, explicit cardiovascular knowledge needs to be made available. Presently such clinical knowledge is not available.

The system of Gupta et al (2005) is a linear system. It is simple and easy to implement. However, its disadvantage is that it is fails to deal with noisy data such as medical data (Manning et al, 2008). The detail of linear and nonlinear models and their discussions can be seen in Chapter 3. Furthermore, it is limited, because of use with just renal transplant patients.

The Framingham study (Framingham Heart study, 1948); the Australian Busselton study (Knuiman et al, 1998); German PROCAM study (Assmann et al, 2002); the COSEHC (COSEHC, 2003); and the INDANA (Pocockt et al, 2001) used logistic regression methods to predict individual risk. These systems are easy to implement and their results are easily interpreted by clinicians. However, they are designed to their specific purposes. For example, the COSEHC (COSEHC, 2003) is limited for use in the Southeastern United States; the INDANA (Pocock et al, 2001) is limited for use with patients aged less than 74,

and concentrates on patients who smoke. Furthermore, according to Smulders et al (2004), these systems are lacking with no validation with an independent test set. The risk categories are heterogeneous resulting from the random spread of known and unknown risk factors in their scoring systems.

The POSSUM and the PPOSSUM systems, like the above logistic regression systems, can be easily implemented. Furthermore, the performance of these systems can be evaluated by comparing the ratios between predicted deaths and actual deaths. However, the disadvantages of the POSSUM and the PPOSSUM systems are as follows:

- **Ignore the significantly contributed factors**: The systems used significant attributes defined by Copeland et al (1991) to calculate the physiological score and the operative severity score. However, these systems do not consider other attributes, which might be involved to the prediction process to the patient risks. For example, according to Kuhan et al (2003), the attribute of "diabetes" is one of the significant factors for cardiovascular risk.
- **Ambiguous in the evaluation methods for system performance**: This is because the systems' performance depends too much on the "linear" or "exponential" analysis methods for their results. For example, Wijesinghe et al (1998) run an experiment with 312 patients. Its result was 0.59 in the ratio of O/E (Observed per Expected (predicted) mortality of the patients) using the linear analysis whereas this ratio was 1.14 when using an "exponential" analysis. They pointed out that the first analysis yielded spurious results, because of using an inappropriate analysis method. The ambiguous use of linear and exponential analysis methods are also shown in (Yii and Ng, 2002).

- **Ambiguous interpretations for the categorical risks in the risk scale**: Like logistic regression systems, the POSSUM and the PPOSSUM systems use fixed thresholds over a numeric risk scale. Therefore, there are ambiguous interpretations to determine categorical outputs such as "*High*", "*Medium*", or "*Low"* from this risk scale.
- **Lack of the validation for the systems' results**: The POSSUM and the PPOSSUM systems lack an independent test set to validate the systems' results.

From this point, a system to improve on the above disadvantages is needed. Pattern recognition and data mining classifiers might provide suitable candidates capable of producing better results.

2.5. Summary

This chapter provides a literature review of alternative decision support systems in medicine, particularly in the area of cardiovascular disease. The MYCIN (Shortliffe, 1976) and the Internist/QMR (Miller et al, 1982) can be seen as general background for medical diagnostic systems. The Framingham study (Framingham Heart Study, 1948); the Australian Busselton study (Knuiman et al, 1998); the German PROCAM study (Assmann et al, 2002); Bayesian networks models (Twardy et al, 2005); the COSEHC (COSEHC, 2003; Hawkins et al, 2005); and the linear scoring system (Gupta et al, 2005) can be seen as providing more focused background on the cardiovascular area. The INDANA system (Pocock et al, 2001) is discussed in detail as an example to show the limitations in predicting risk for cardiovascular patients.

The main focus in this chapter is the POSSUM and the PPOSSUM scoring systems. These systems can predict outputs of mortality, morbidity and death rate for general surgical

patients via the logistic regression method. However, these systems have some disadvantages as discussed. Consequently, another more appropriate approach needs to be investigated. It is suggested that the areas of pattern recognition and data mining classifiers may help in this.

Chapter 3

Pattern Recognition

3.1. Introduction

This chapter presents a background for pattern recognition, with an emphasis on learning to recognise patterns in data sets. Four basic methods of machine learning are introduced: template matching; statistical classification; syntactic or structure matching; and neural networks. Linear and nonlinear models are also introduced in this chapter. The comparisons between both models lead to a discussion on the use of nonlinear models for noisy data as found in medical domains. Alternative evaluations for classification performance are introduced in this chapter. They are: mean square error; confusion matrix; accuracy; the rates of sensitivity and specificity as well as the positive predictive value and the negative predictive value. These rates are used in all thesis experiments in later chapters for discussions and comparisons. This chapter also presents a brief literature review in the application of pattern recognition techniques in medical domains.

3.2. What is Pattern Recognition?

There are many ways to define what a pattern is, depending on the area of study. In machine learning and data mining, a pattern can be defined as a set of measurements or observations, which can be represented in vector or matrix notation. For example, a pattern can be seen as a vector (patient's record) in the space of a medical data domain, where each data attribute represents separate dimensions.

Recognition is the classification of data according to predefined patterns. In other words, recognition means that "a computer can recognize the patterns of objects as ones that it has seen before" (Anzai, 1992). For example, recognition in respect to a medical database might be the classification of patients' status ("Low Risk" or "High Risk") given their symptoms (predefined patterns).

Pattern recognition is an area of research, which studies the operation and design of systems for recognizing patterns in a data domain. Pattern recognition has been of interest in many areas of study such as:

- Diagnosing diseases (Baxt, 1991; Pedersen et al, 1996; Harrison et al, 1991; Plaff et al, 2004);
- Character recognition and hand writing recognizer (LeCun et al, 1995);
- Speech analysis (Rabiner and Juang, 1993).

More detail about these applications can be seen in Ripley (1996); and Lisboa (2002). The use of pattern recognition techniques is to somehow mimic the decision making of humans. According to Tou and Gonzalez (1974), the fundamental problems in building a pattern recognition system are:

- The sensing problem;
- The pre-processing and feature extraction problem;
- And the determination of optimum decision procedures, which are needed in the identification and classification process.

In the sensing problem, the pattern recognition system is concerned with the representation of input data. This representation should be chosen to aid the measurement of similarity between the current object and previously recognized classes. In the second problem, the system is concerned with the extraction of characteristic features or attributes from input

data and the reduction of the dimension of pattern vectors. In the third problem, the data will be formed as pattern points or measurement vectors in feature space. By using alternative classified techniques, the system will decide to which classes these data belong. The architecture of an archetypal pattern recognition system can be seen in Figure 3.1.

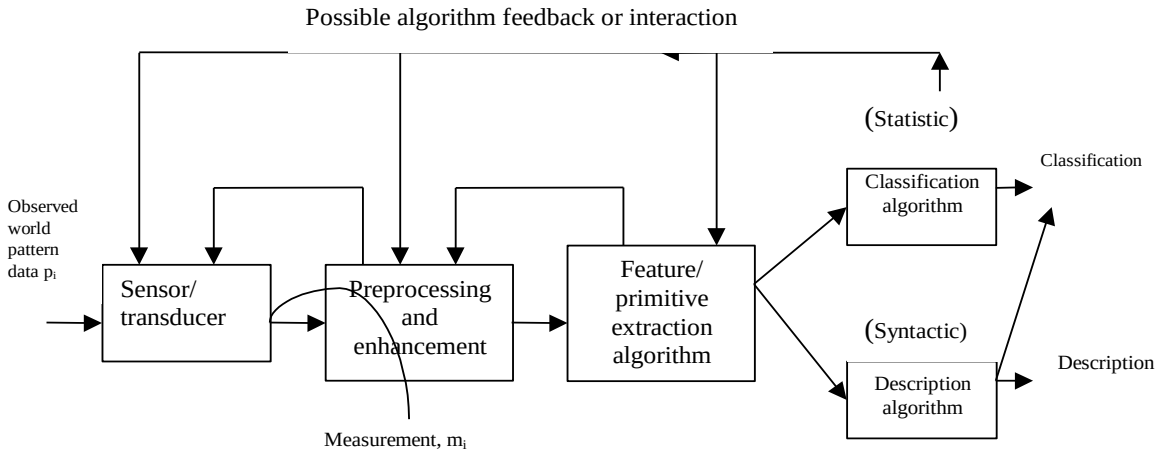

Figure 3.1: Typical pattern recognition system architecture (Schalkoff, 1992).

As an example of this pattern recognition process, cardiovascular diagnosis can be seen as follows: The original collected data is represented in the form of database files; data might be then reduced or cleaned by using alternative data mining techniques; by using feature exaction algorithms, the feature space dimension might be reduced; the system then uses classification algorithms or description algorithms depending on the requested diagnostic purpose. By using classification algorithms, for instance, the system can predict the patient risk. Alternatively, by using description algorithms, the system can describe patterns in different classes based on their characteristics.

Basically, the architecture of the thesis pattern recognition process follows this archetypal process. The detailed steps of the process can be seen in the “Thesis Methodology” section of Chapter 5 (Section 5.2.4).

The main issue in pattern recognition is building of a learning machine to represent the training set; defining methods of recognition; and evaluating the results. According to Jain et al (2000), four main approaches are as follows:

- Template matching;
- Statistical classification;
- Syntactic or structural matching;
- And neural networks.

In fact, these approaches are not necessarily independent in their application. Sometimes a combination might be a better approach to deal with actual pattern recognition issues. A summary of the four approaches to pattern recognition can be seen in Table 3.1.

Approach	**Representation**	**Recognition Function**	**Typical Criterion**
Template Matching	Samples, pixels, curves	Correlation, distance, measure	Classification error
Statistical	Features	Discriminant function	Classification error
Syntactic or structural	Primitives	Rules, grammar	Acceptance error
Neural networks	Samples, pixels, Features	Network function	Mean square error

Table 3.1: Pattern recognition (Jain et al, 2000).

3.3. Methods of Pattern Recognition

3.3.1. Template Matching

Template matching is one of the simplest and earliest approaches to pattern recognition. Matching is a generic operation used to determine the similarity between two entities

(points, curves, or shapes) of the same type. For example, in clustering tasks in Chapter 4, the similarity can be defined by measurements such as Euclidean distance. Other well-known examples of the template matching approach can be seen in symbol recognition problems (LeCun et al, 1995). The example below shows how a template matching method can be used to recognise patient status as “High risk” or “Low risk”.

Example 3.1

Assume that Table 3.2 represents patient information. The set of symptoms are shown in columns as Wound, CNS, Haematology, Carotid, and Cardiac (input columns). The results of diagnosis are shown in the Tract, Graft, and PTA columns (output columns). The column “Row ID” shows patient identifications. The risk for each patient can be calculated as follows. Note that risk here has values labelled as 1 or 0, meaning "High risk" or "Low risk" respectively.

IF ∑(Tract, Graft, PTA) =1 ➔ Risk =1 (High risk)

Other wise, ➔ Risk =0 (Low risk).

Row ID	Wound (i1)	CNS (i2)	Haema -tology (i3)	Carotid (i4)	Cardiac (i5)	GU tract (i6)	Graft (i7)	PTA (i8)	Risk
t_1	1	0	0	0	0	0	0	0	0
t_2	1	0	1	0	0	1	0	0	1
t_3	0	1	0	0	0	0	0	0	0
t_4	1	0	0	0	0	???	???	???	???
t_5	1	1	0	0	1	???	???	???	???

Table 3.2: Information on patients in the cardiovascular domain.

Given a new patient labelled as t_4 in Table 3.2, by comparing symptoms to the given patterns in the database, the system produces t_4 results labelled as (0, 0, 0). This is because

of the exact match to symptoms of patient t_1. Therefore, t_4 risk is labelled as 0 or "Low risk". However, sometimes, the system cannot produce output as it contains insufficient information to match to the given symptoms of a new patient. For example, t_5 cannot be given any output labels as it is only a partial match to any existing patient.

This approach is not be used within this thesis and more detail can be seen in Schalkoff (1992).

3.3.2. Statistical Pattern Approach

This section demonstrates the use of statistical pattern approach for the prediction or classification process. In the statistical pattern approach, each pattern is represented in terms of d- features or measurements. The main purpose of this approach is to choose suitable features in order to assign pattern vectors to different categories. If the patterns can be divided into separate classes, the feature space will be well determined. At this point, there exists a decision boundary in the feature space to separate these classes. Bishop (1995) defined the decision boundaries using a discriminant function based on Bayes' theory. Alternatively, these decision boundaries are built based on a classification approach (Jain et al, 2000). Detail about these methods can be seen in Bishop (1995); and Jain et al (2000). An example of a statistical pattern recognition model can be seen in Figure 3.2. The model contains two modes: the training (or learning) process; and the classification (or testing) process. In the pre-processing module, the pattern is segmented from the background, noise removed then normalized. For example, the pre-processing stage deals with missing values (cleaning task) and transforming original valued types into more appropriate types (normalisation task).

In the training stage, the "feature extraction or selection" module finds appropriate features for the representation of input patterns. The classifier is then trained to partition the feature

space. In the classification stage, the trained classifier assigns each input pattern to one of the pattern classes based on the measured features.

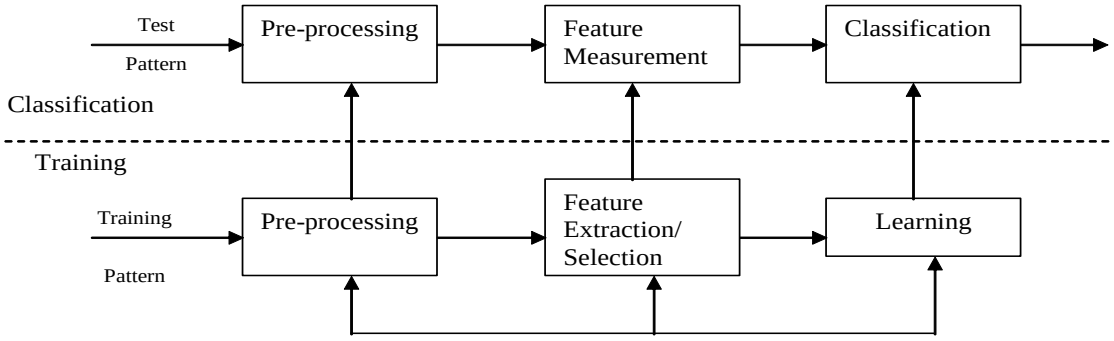

Figure 3.2: Model for statistical pattern recognition (Jain et al, 2000).

3.3.3. Syntactic Pattern Approach

The syntactic pattern approach is based on concepts from formal language theory, and in particular mathematical models of grammar (Gonzalez and Thomason, 1978). Syntactic pattern recognition decomposes the patterns into sub-patterns or primitives. The goal is to classify each pattern as belonging to a specific class. The decomposition of patterns is sometimes referred to as parsing. Schalkoff (1992) suggested two approaches: top down parsing, and bottom up parsing. The syntactic pattern recognition approach can be used for the classification and description purposes.

The elements of classification process are shown in Figure 3.3 (Schalkoff, 1992). This approach has disadvantages in implementation if the data set includes noisy patterns. This is because of difficulties in detecting the primitives, and in the inference of the grammar. Moreover, the explosion of combinatorial possibilities requires a large training data set and much computational effort (Perlovsky, 1998). More detail about this approach can be seen

in Fu (1982) and Schalkoff (1992). Another view of syntactic pattern approach can be seen in the "Decision Trees" section.

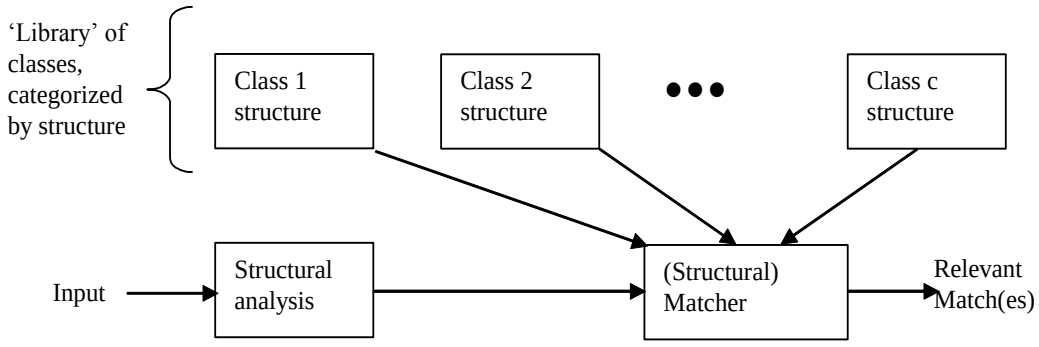

Figure 3.3: Using syntactic pattern approach for classification (Schalkoff, 1992).

3.3.4. Neural Network Pattern Recognition

The general background to neural networks is introduced in this section. Artificial neural networks (neural networks) have a rich history of research, starting with the McCulloch and Pitts (1943) concept of neural networks, and the following popular Hebbian rule (Hebb, 1949). From the first concept of perceptron (Rosenblatt, 1958; Rosenblatt, 1962; Minsky & Papert, 1969), neural networks have developed quickly and been applied in many areas. The three characteristic components of a neural network can be seen as:

- The network topology, or interconnection of neural 'units';
- The characteristics of individual units or artificial neurons;
- And the strategy for pattern learning or training.

The common point of view is that the neural network approach is a black-box strategy, which is trainable Haykin (1999). This means the outputs and the weights can be changed during the learning process. The goal in using neural networks is to build a good

relationship between inputs and outputs. Therefore, its performance seems to be strongly influenced by the quality of the training data and any pre-processing algorithm.

Alternatively, neural networks can be seen as a directed graph with input, output, and hidden nodes where the input nodes are described in an input layer, the output nodes belong to an output layer, and the hidden nodes exist over one or more hidden layers. The number of input nodes is usually the number of input attributes in data domain. The output nodes will determine predictive outcomes. The number of hidden layers and hidden nodes are chosen by heuristic methods. Detailed choices of hidden nodes and layers for the neural network topologies in the thesis can be seen in the specific experiments.

The advantage of the neural network approach is that the predictive outcomes can be improved via the learning process. However, its disadvantage is the difficulty in explaining the prediction results to end users. More detail about this approach can be seen in Chapter 4 with the focus on neural network techniques as multilayer perceptron, radial basis function, and support vector machine. A review about the neural network applications can be seen in section 3.6.1 below, and in Sondak and Sondak(1989); Papik et al (1998); Jain et al (2000); and Lisboa (2002).

3.3.5. Decision Trees

A decision tree is another view of the syntactical pattern recognition approach. This is because a tree includes a number of nodes that have a structure similar to the grammatical analysis in the syntactic pattern recognition method. The fundamental (root) node in the tree is a single node that has no connection from other nodes. A diagrammatic example of a decision tree can be seen in Figure 3.4.

Example 3.2

Assume the transactions of the data set as given in Table 3.2 above. The result of building a decision tree for this data can be seen in Figure 3.4. The tree can be interpreted as a set of rules to define patient risk as follows:

$$\textit{if } S = (i_1 + i_2 + i_3 + i_4 + i_5) >= 5 \quad \rightarrow \textit{ Risk=''High risk''}$$

$$\textit{Otherwise, if } (i_4 + i_5) \geq 1 \quad \rightarrow \textit{ Risk = "High risk"}$$

$$\textit{Otherwise,} \quad \rightarrow \textit{ Risk = "Low risk"}$$

For example, patient t_4 in Table 3.2 will be labelled by this decision tree as "Low risk" as ($S = i_1 + i_2 + i_3 + i_4 + i_5 = 1;\ S < 5$) and ($i_4 + i_5 = 0;\ <1$). Moreover, patient t_5, can be classified as "High risk" according to this tree as ($S = i_1 + i_2 + i_3 + i_4 + i_5 = 3;\ S < 5$) and ($i_4 + i_5 = 1;=1$).

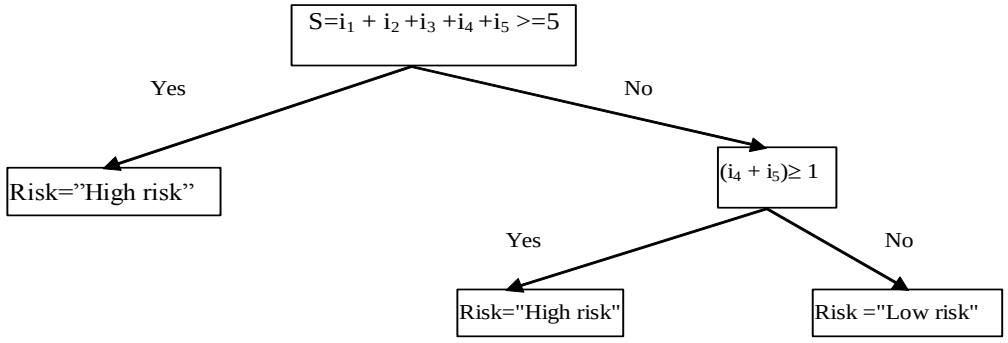

Figure 3.4: A decision tree for the data in Table 3.2.

More detail on building decision trees can be seen in Quinlan (1986); Quinlan (1993) with the ID3 (Iterative Dichotomiser 3) algorithm and its extension, the C4.5 algorithm; or in Breiman et al, (1984); Lewis (2000) with the use of CART (Classification And Regression Tree). The application of the C4.5 algorithm, as the J48 algorithm (Witten and Frank, 2000), is used in Chapter 8 (Table 8.3) for the comparison purpose with neural network techniques.

3.3.6. Discussion

The comparisons of statistical pattern recognition, syntactic pattern recognition, and neural network are shown in Table 3.3. According to Table 3.3, the statistical pattern recognition has difficulty in representing the structure of patterns whereas this structure is easily presented in the syntactic pattern recognition approach. This is also true for the decision tree, where the pattern structure is clearly shown. This helps the end users to interpret the classification problem.

	Statistical Pattern Recognition	**Syntactic Pattern Recognition**	**Neural Network**
Pattern generation	Probability models	Formal grammars	State or weight array
Pattern classification (Recognition/ Description) basis	Estimation or decision theory	Parsing	Based on properties of neural network
Feature organization	Feature vector	Primitives and observed relations	Neural input or stored states
Typical learning (Training) approaches: Supervised *Unsupervised*	Density or distribution estimation (usually parametric) Clustering	Forming grammars (heuristic or grammatical inference) Clustering	Determining system parameters (e.g., weights) Clustering
Limitation	Difficulty in expressing structural information	Difficulty in learning structural rules	Little semantic information from network.

Table 3.3: Comparing statistical pattern recognition, syntactic pattern recognition, and neural network approaches (Schalkoff, 1992).

In fact, it is difficult to classify the boundaries between statistical pattern recognition, syntactic pattern recognition, and neural networks. For example, the classification for Example 3.2 with the statistical approach can be performed by representing the data in a feature space (5 - dimensions), with a decision boundary built to separate the patients into different classes.

Hence, whenever there is a specific pattern recognition problem, an appropriate approach should be chosen based on an analysis of underlying statistical components (statistical pattern recognition), underlying grammatical structure (syntactic pattern recognition), and the suitability of a neural network solution. According to Tsai and Fu (1980), sometimes the neural network approach might be seen as an implementation derived from the statistical pattern recognition and syntactic pattern recognition approaches. Therefore, depending on actual classification situations, the most appropriate approach might be applied.

3.4. Linear and Non Linear Pattern Recognition Techniques

This section introduces linear and nonlinear models. The linear models are compared to nonlinear models via their use as pattern recognition techniques.

3.4.1. Linear Models

Definition: A linear model is a model defined using a linear function. A linear function can be represented in an n-dimensional space as follows:

$$y = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$$

where $x = (x_1, x_2, .., x_n)$ is a vector of a pattern in an n-dimensional space,

$w = (w_1, w_2, .., w_n)$ is a parameter vector of x in the data space

The weights are used to define the decision boundary for the classification problem. For simplicity, y can be presented as:

$$y = w^Tx + b \text{ ; or } y = x^Tw + b$$

Example 3.3

Assume that the data distributions can be represented in the graph in Figure 3.5. Data is classified into two classes of “High risk” and “Low risk” areas. Obviously, this is a linear

classification problem in a 2-dimensional space, because its decision boundaries, to separate the output classes, can be represented as linear functions.

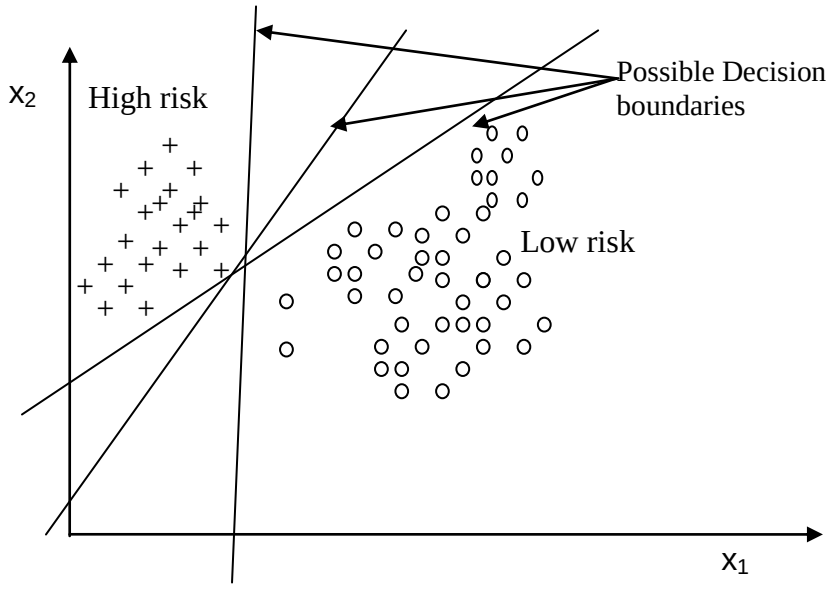

Figure 3.5: An example of a linear classification problem.

3.4.2. Nonlinear Models

Definition: Any model that can not be defined using a linear model can be seen as a nonlinear model. It is defined using nonlinear functions. Nonlinear functions are represented as:

$y = f(w, x)$, where $w=(w_1, w_2,..., w_n)$ is not linear; or f is a nonlinear function.

Example 3.4

Assume that a Gaussian distribution of data as seen in Figure 3.6. This Gaussian function, in a 3-dimensional space, can be seen as

$$y = f(x, \mu, \delta) = f(x_1, x_2, \mu, \delta) = e^{-\frac{1}{2}(\frac{x-\mu}{\delta})^2} ;$$

where $\mu= 0$; and $\delta=0.5$.

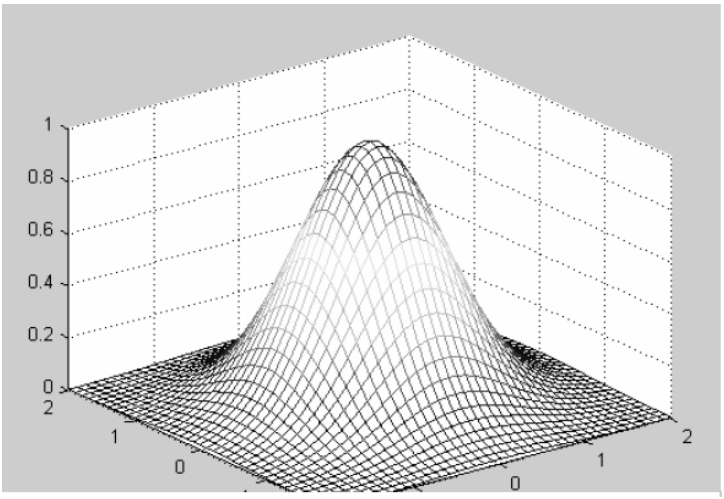

Figure 3.6: Gaussian distributions of data in a 3-dimensional space.

The graph can be redrawn into a 2-dimensional space as seen in Figure 3.7. It is clear that this is an example of a nonlinear classification problem, because the decision boundary between "High risk" and "Low risk" classes is a curve.

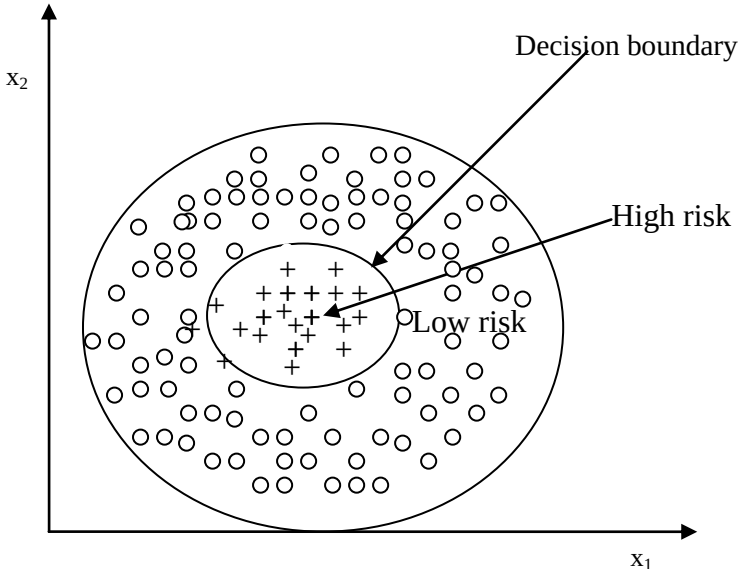

Figure 3.7: An example of a nonlinear problem.

3.4.3. Linear Vs Nonlinear Models

A comparison of the linear and nonlinear models can be seen in Table 3.4.

Linear models	Nonlinear models
Advantage	
• Simple, and easy for the interpretation of models • Clearly separable by decision boundaries between classes. • A basic choice for a mixed models or nonlinear models analysis • Low requirement of computation (Vapnik, 1999; Lotte et al, 2007).	• Can produce good estimates of the unknown parameters in the model with relatively small data sets (NIST/SEMATECH, 2006). • Able to deal with noisy classification problems (Manning et al, 2008). • Might produce confident values (Jain et al, 2000).
Disadvantage	
• Poor prediction for outliers • Poor extrapolation properties (NIST/SEMATECH, 2006) • Very sensitive to the presence of unusual data points in the data used to fit a model • Hard to deal with noisy classification problems (Manning et al, 2008)	• Slow training parameters (Jain et al, 2000) • Complexity to implement models • Difficult to interpret and explain to the end users. • Require an iterative search for the best parameter values.
Prediction and Classification Method Examples	
• Linear scoring system (Gupta et al, 2005). • Single Perceptron (Haykin, 1999).	• INDANA (Pocock et al, 2001). • POSSUM and PPOSSUM systems (Copeland et al, 1991). • Logistic classifier (Anderson, 1982) • Neural networks such as multilayer perceptron, radial basis function, and support vector machine with kernel functions (Jain et al, 2000).

Table 3.4: Comparisons of linear and nonlinear models.

It is clear that linear models are the first choices whenever a new model is generated. However, they struggle to deal with the classification of noisy data whereas nonlinear classifiers usually deal better with noisy data (Manning et al, 2008). Furthermore, according to NIST/SEMATECH (2006), linear models have limited shapes, because they only can be devised using linear functions. This might cause a poor performance for their classification process. On the other hand, nonlinear models can be used with a broad range of linear and nonlinear functions. Therefore, they might demonstrate better classification performance than linear models.

For example, the linear scoring model of Gupta et al (2005) introduced in Chapter 2 uses a global linear function to produce the cardiovascular risk for the patients. However, its performance might be poor with noisy classification problems as indicated by the disadvantages of linear models in Table 3.4. The INdividual Data ANalysis of Antihypertensive (INDANA) intervention trials (Pocock et al, 2001) as well as the POSSUM and the PPOSSUM systems introduced in Chapter 2 use local linear functions to calculate the system scores. These scores are then used with nonlinear functions, which are derived from the logistic regression, to produce individual numerical risks. Therefore, they show advantages in dealing with noisy classification in the cardiovascular risk prediction as indicated in Table 3.4. Similarly, neural network classifiers such as multilayer perceptron, radial basis function, and support vector machine are nonlinear models. Note that a single perceptron, however, can be viewed as a linear classifier (Haykin, 1999). Multilayer perceptron use nonlinear (logistic) activation functions; radial basis functions use Gaussian activation functions in its hidden nodes; and support vector machine use (eigenvalues) kernel activation functions. Therefore, as indicated in Table 3.4 they demonstrate advantages in dealing with noisy classification problems, and might produce better results.

Overall, nonlinear models show advantages in dealing with complicated classification problems. Additionally, the INDANA as well as the POSSUM and the PPOSSUM disadvantages as shown and discussed in Chapter 2, suggest that nonlinear neural network classifiers, such as multilayer perceptron, radial basis functions, and support vector machines, might be more appropriate for the thesis data.

3.5. Evaluating Classifiers

This section introduces the concepts of classification performance evaluations. They are: the mean square error (*MSE*); confusion matrix; accuracy (ACC); sensitivity (S*en*); specificity (*Spec*) rates; and the positive predictive value (PPV) and the negative predictive value (NPV). These evaluations are used in all thesis case studies and experiments.

3.5.1. Mean Square Error

Assume that the data domain has input patterns x_i *(i=1, 2,..,n)*, and a target pattern Y_i. The classifier produces the output y_i. The mean square error is mean of square of the error between the predicted and target output. It is given by:

$$E = \frac{\sum_i (y_i - Y_i)^2}{n} \quad (3.1)$$

3.5.2. Confusion Matrix

Assume that the cardiovascular classifier output set includes two typically risk prediction classes as: “High risk”, and “Low risk”. Note in the thesis, unless stated otherwise, alternative valued scales such as “Very High risk”, "High risk", and “Medium risk” are all referred to as “High risk”. Each pattern x_i *(i=1, 2..n)* is allocated into one element from the set {*P, N} (positive or negative*) of the risk prediction classes. Hence, each input pattern

might be mapped into one of four possible outcomes such as t*rue positive- true high risk (TP)*- when the outcome is correctly predicted as *High risk; true negative- true low risk (TN)*- when the outcome is correctly predicted as *Low risk; false negative-false Low risk (FN)*- when the outcome is incorrectly predicted as *Low risk*, in fact it is *High risk* (positive); or *false positive- false high risk (FP)* - when the outcome is incorrectly predicted as *High risk*, in fact it is *Low risk* (negative). The set of {*P, N*} and the predicted risk set can be built as a ***confusion matrix***.

		Predicted classes	
Expected /Actual Classes		High risk	Low risk
	High risk	*TP*	*FN*
	Low risk	*FP*	*TN*

Table 3.5: Confusion matrix.

From the confusion matrix in Table 3.5, the number of correct or incorrect (misclassification) patterns can be derived. The numbers along the major diagonal (from left to right) represent the correct while the rest represent the errors (confusion between the various classes).

3.5.3. Performance measures

Some related concepts according to Altman and Bland (1994a; 1994b); Dunham (2002); Ye(2003); Han and Kamber (2006); Kononenko and Kukar (2007); and Bramer (2007) such as the accuracy (ACC), sensitivity (*Sen*), specificity (*Spec*) rates, and the positive predictive value (PPV or precision), and the negative predictive value (NPV) can all be built from the confusion matrix. These rates are used to evaluate and discuss classification performance.

The accuracy (Duham, 2002; Ye, 2003; Han and Kamber, 2006; Kononenko and Kukar, 2007; Fielding, 2007; and Bramer, 2007) of a classifier is calculated by the total number of correctly predicted "High risk" (*true positive- true High risk*) and correctly predicted "Low risk" (*true negative- true Low risk)* over the total number of classifications. It is given by:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.2)$$

The error rate of performance, or misclassification rate, can be referred from this accuracy rate as: *1- ACC.*

However, the accuracy does not show how well the classifier can predict the positive ("High risk") and the negative ("Low risk") for the classification process. Therefore, the sensitivity, specificity, positive predictive value, and negative predictive value measurements are created for this purpose (see detail in Figure 3.8)

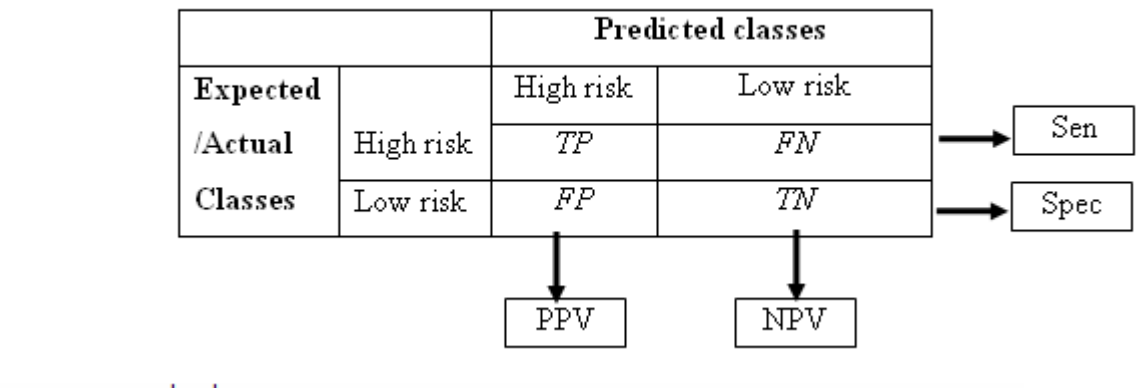

Figure 3.8: Classification performance rates.

The sensitivity (Duham, 2002; Ye, 2003; Han and Kamber, 2006; Kononenko and Kukar, 2007; Bramer, 2007; and Fielding, 2007) (given by Equation 3.3) is the rate of number correctly predicted "High risk" (*true positive- true high risk*) over the total number of correctly predicted "High risk" and incorrectly predicted "Low risk" (*false negative- false*

Low risk). This rate can be seen as the rate of correctly predicted "High risk" over the total of expected/actual "High risk".

$$Sen = \frac{TP}{TP + FN} \quad (3.3)$$

The specificity rate (Duham, 2002; Ye, 2003; Han and Kamber, 2006; Kononenko and Kukar, 2007; Fielding, 2007; and Bramer, 2007) is the rate of correctly predicted "Low risk" over the total number of expected/actual "Low risk". It is given by:

$$Spec = \frac{TN}{TN + FP} \quad (3.4)$$

The positive predictive value (Altman and Bland, 1994b; Fielding, 2007; and Bramer, 2007) is the proportion of correct "High risk" over the total number of predicted "High risk" (including correct "High risk" and incorrect "High risk" after classification process). It is given by:

$$PPV = \frac{TP}{TP + FP} \quad (3.5)$$

The negative predictive value (Altman and Bland, 1994b; Fielding, 2007; and Bramer, 2007) is the proportion of correct "Low risk" over the total number of predicted "Low risk" (including correct "Low risk" and incorrect "Low risk" after classification process). It is given by:

$$NPV = \frac{TN}{TN + FN} \quad (3.6)$$

Example 3.5

Assume that the confusion matrix and resulted evaluations is represented in Table 3.6.

		Predicted classes		Evaluations				
Expected/ Actual Classes		High risk	Low risk	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
	High risk	*80*	*5*	0.90	0.94	0.86	0.84	0.95
	Low risk	*15*	*95*					

Table 3.6: An example of confusion matrix and measurements.

From Table 3.6, the accuracy rate of 90% can be understood as a misclassification rate of 10%. The sensitivity (correct "High risk") rate of 0.94 is higher than the accuracy of the classification (0.90). Conversely, the specificity (correct "Low risk") rate of 0.86 is smaller than this accuracy. Therefore, the accuracy of 0.90 shows the trade-offs of performance between the correct "High risk" and "Low risk" predictions in the classification process.

The sensitivity of 0.94 in Table 3.6 shows that the rate of correctly predicted "High risk" is 80 over total of 85 expected/actual positive cases. This means the misclassification of 5 cases (5 per 85 positive cases) are predicted as the "Low risk" although they are the "High risk".

The positive predictive value of 0.84 shows the rate of expected correct "High risk" is 80 per total of 95 predicted positive cases. This means the misclassification of 15 cases (15 per 95 predicted positive cases) are the expected/actual "Low risk" although the classifier predicted them as the "High risk". Conversely, the same explanation can be done for the specificity rate and the negative predictive value for the correct "Low risk" predictions.

The sensitivity rate and the positive predictive value are of more interest in the thesis discussions, because they mirror the number of correct "High risk" patients in medical data domain.

3.6. Brief Literature Review of Pattern Recognition Techniques in Medicine

This section provides an overview of pattern recognition applications in medicine, particularly the cardiovascular area, using both supervised and unsupervised learning approaches. The supervised techniques are represented as neural network classifiers such as multilayer perceptrons, radial basis functions, and support vector machines. The unsupervised techniques are represented as the use of self organizing maps, and clustering algorithms. This review can be seen as the motivation to explore the potential value of the techniques outlined in this thesis.

3.6.1. Supervised Neural Networks in Medicine

Baxt (1991; 1992) designed the first application of a neural network to detect the presence of acute myocardial infarction in patient presented to the emergency department with anterior chest pain. The use of multilayer perceptron, with 351 hospitalized patients in a high likelihood of having myocardial infarction, gave good results with a sensitivity of 97.2% and a specificity of 96.2%. This technique was also used in the different researches of Harrison et al (1991); Baxt & Skora (1996; Heden et al (1996); Pedersen et al (1996); Jorgensen et al (1996); Ellenius et al (1997); Colombet et al (2000); Cacciafesta et al (2000); Cacciafesta et al (2001); and Gueli et al (2005) with good results in alternative areas of cardiac disease.

Radial basis function networks are less widely used than multilayer perceptron networks in the medical domain. Radial basis function networks were used by: Langdell and Mason (1998) for training and testing of spinal measurements in order to classify spines into exposed and unexposed classes; Maglaveras (1998) to deal with electrocardiogram (ECG) pattern recognition and classification in data sets from the MIT-BIH and the European ST-

T databases; and Luan et al (2005) to build quantitative structure–property relationship models to predict the pK^a values for new drugs.

A support vector machine (Boser et al, 1992; Cortes and Vapnik, 1995; Vapnik, 1995; Vapnik, 1998) is a technique used to reduce the dimension of feature space. For example, Wang et al (2004) used support vector machines to discriminate cardiovascular disease patients from non-cardiovascular disease controls. They reported that the specificity and sensitivity for clinical diagnosis of cardiovascular patients as 85% and 93% respectively. The support vector machine technique also can be seen in the research of Nurettin (2006) as a perturbation method to reduce feature space dimensions for an ECG recognition system. It discards the redundant data components from the training data set. The performance of the system resulted in 91.7% sensitivity and 87.3% specificity.

A comparison between alternative supervised neural network techniques can be seen in Serretti & Smeraldi (2004) where the multilayer perceptron out-performed the radial basis function with sensitivity measures of 85.61% and 35.21%, and specificity of 77.50% and 51.20% respectively. Kamruzzaman and Begg et al (2006) showed that the support vector machine demonstrates a higher performance than the multilayer perceptron with 3.21% and 1.93% in sensitivity and specificity measurement respectively. According to Lisboa (2004), the use of multilayer perceptron in medicine is greater than other neural network techniques. For example, Lisboa (2004) reviewed 24 journal papers in neural network areas of randomised controlled studies and controlled studies. However, only three publications used the alternative techniques of Bayesian evidence approximation (Matsui et al, 2003) and support vector machine (Lin et al, 2004; Chan et al, 2003). The remaining 21 studies relied on multilayer perceptron.

The comparison between neural network techniques such as multilayer feedforward and Bayesian networks is shown in Bigi et al (2005). 496 patients with acute myocardial infarction were used to predict the cumulative end-point of cardiac death, nonfatal reinfarction and unstable angina. They complained that multilayer feedforward did not improve the prognostic classification of patients with uncomplicated acute myocardial infarction as compared to a robust Bayesian classifier.

3.6.2. Unsupervised Pattern Recognition Techniques in Medicine

Clustering techniques have been reported by many researchers, for example Hebb (1949); Widrow and Hoff (1960); Rosenblatt (1958); Rosenblatt (1962); Carpenter *et al* (1990); Jain et al (1996); Kohonen (1990). This section focuses briefly on two main categories of clustering using in the cardiovascular data domain: self organizing maps (Kohonen, 1990); and partitional cluster algorithms such as the K-means (Forgey, 1965; MacQueen, 1967; Hartigan, 1975; Hartigan and Wong, 1979).

Although the use of self organizing maps for pattern recognition is wide spread, there are few applications in medical domains. The self organizing maps are often used when analysing medical images, such as registering multimodal retinal images (Matsopoulos et al, 2004), or in new drug research (Balakin et al, 2005). The self organizing maps are also used by Simelius et al (2003) for spatiotemporal analysis and classification of Body Surface Potential Mapping (BSPM) data in the cardiac domain, and by Mia et al (2003) for identifying clusters which are based on mammographic findings and patient age in the breast cancer domain.

According to Jain (1999); and Berkhin (2002), K-means algorithm is the most popular clustering tool used in scientific and industrial applications. Two classical K-means

methods are the techniques of Jancey (1966) and MacQueen (1967). Research into these methods continues to thrive with the introduction of new ideas and extensions to the original algorithms (Cheung, 2003; Lingras et al, 2002; and Lin et al, 2004). For example, Maschewsky-Scbneider and Greiser (1989) used cluster analysis (FAST CLUS, a method similar to K-means) in a group of 1372 men and women participating in the German cardiovascular prevention study to identify risk factor profiles. Another use of clustering algorithm is shown in Plaff et al (2004). By using "clusterbase/rulebase", the system predicts individual cardiovascular risks for 63 long-term hemodialysis patients. Syed et al (2007) used clustering algorithm for analyzing large amounts of cardiovascular signal data without any a priori knowledge about disease states. The use of max-min clustering and a fuzzy pre-clustering phase in this paper allowed the analysis of large amounts of data without excessive demands in terms of computational time or space.

Although the success of self organizing maps, K-means, and other clustering algorithms cannot be denied, they suffer from some inherent drawbacks. For example, these algorithms can only be used with numerical data whereas many medical domains require the use of alternative data types. The application and improvement of self organizing maps and K-means are issues that continue to challenge researchers throughout the field.

3.7. Summary

This chapter provides a general background on pattern recognition. The template matching, statistical, structural, and neural network approaches provide a general view for pattern recognition techniques. The combination of these approaches might lead to a set of techniques appropriate for any specific data domain. The decision tree approach is

mentioned in this chapter as an extension of the syntactic pattern recognition approach. The J48 algorithm is used for the comparison purpose in Chapter 8.

Standard measurements are introduced in detail as the tools to evaluate the performance in all thesis experiments. Mean square error and accuracy rates will show the overall performance for the experiments. The sensitivity rate and the positive predictive value show alternative way of looking at “High risk” patients after the classification process. Similarly, the specificity rate and the negative predictive value show alternative way of looking at “Low risk” patients after the classification process. The differences between them might provide for interesting discussions.

Linear models show many advantages in their implementation and interpretation. However, these models are of limited use, because of their linear functions used in the classification process. Nonlinear models such as neural network classifiers have the advantage of using alternative nonlinear functions in the classification process. Therefore, they can show the ability to deal with noisy data such as found in medical domains. Furthermore, the reviews of supervised neural network classifiers such as multilayer perceptron, radial basis function, and support vector machine; and unsupervised pattern recognition techniques such as self organizing maps and clustering algorithms in medicine are shown as the motivations to use them for this thesis.

Chapter 4

Supervised and Unsupervised Pattern Recognition Techniques

4.1. Introduction

Alternative pattern recognition approaches were discussed in chapter 3. Their techniques are applied broadly in many areas, particularly in the medical domain (Lisboa, 2002). This chapter is motivated by the need to determine the advantage of using supervised and unsupervised learning pattern recognition in the cardiovascular domain. The supervised learning techniques introduced in great detail are the multilayer perceptron, radial basis function, and support vector machine. They are neural network techniques reviewed in Chapter 3. Two unsupervised learning methods of pattern recognition and data mining are also discussed in this chapter. They are self organizing maps and the KMIX clustering algorithm.

4.2. Neural Network Pattern Recognition

Artificial neural networks have been proposed as a reasoning tool to support clinical decision making from the early days of computing (Lisboa, 2002). They are adaptive models for data analysis and particularly suitable for handling nonlinear functions. This section provides a detailed view about the concepts of the perceptron, multilayer perceptron, radial basis function, and support vector machine. The WEKA software tool

(WEKA, 2005) is also introduced in this section to show the applying of neural network in and their chosen parameters.

4.2.1. Perceptron and MultiLayer Perceptron

The **perceptron** is a one of the simplest neural networks (see Figure 4.1). With the input vector x, and a target vector Y_i (expected output), the network produces an output vector y_i (predicted output). The error between the predicted and the expected (target) output is calculated by mean square error (Equation 3.1 in Chapter 3). It can be rewritten as:

$$E = \frac{\sum_i (y_i - Y_i)^2}{n} \quad (4.1).$$

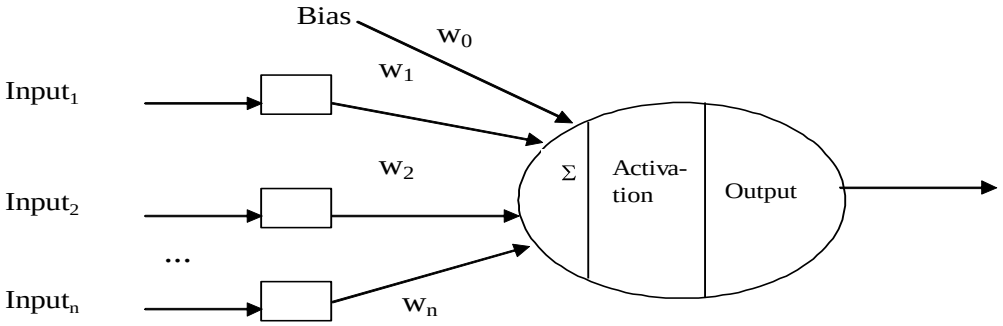

Figure 4.1: Node in full perceptron model.

Assume that the decision boundary for a perceptron is given by:

$$y_i = f(\sum_{j=1}^{n} w_j x_j + w_0) = f(net_i) \quad (4.2),$$

where w_j is the weights of the network, w_0 is bias, and f is an activation function, and $x = \{x_j\}$ is a pattern vector in the input space.

Hence, the outputs belong to appropriate classes depending on which side of a decision boundary they fall. For example, assume that a linear decision function defines the classes for the pattern x according to the following rule:

$$\text{If } \sum_{j=1}^{n} w_j x_j > 0 \text{ then the pattern } x \text{ belongs to the class } C_1$$

Otherwise, the pattern x belongs to the class C_2.

Therefore, the visualization of patterns classified by a perceptron via the linear decision boundary can be seen in Figure 4.2. In other words, the perceptron classification can be seen as a linear pattern recognition technique.

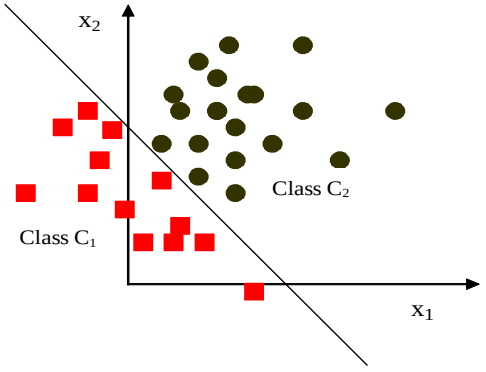

Figure 4.2: Example of linear hyper-plane.

A **multilayer perceptron** is formed through the combination of multiple perceptrons in separate layers. An example of a feed-forward multilayer perceptron can be seen in Figure 4.3

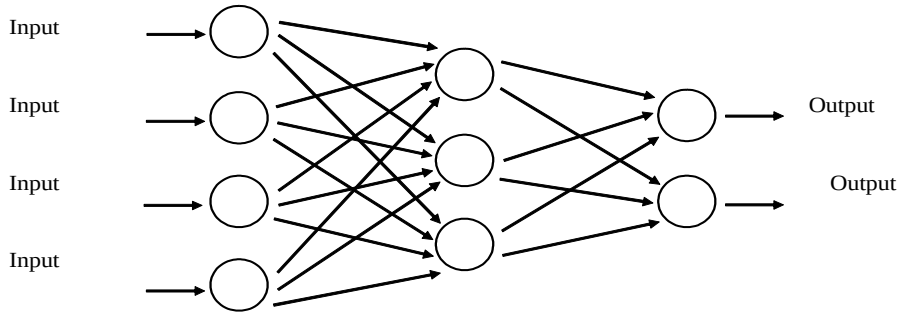

Figure 4.3: An example of a multilayer perceptron.

One popular algorithm for modifying weights in the multilayer perceptron learning process is the **back propagation algorithm.** According to Haykin (1999)**,** this algorithm consists of the following steps:

- **Step 1:** Pass the input nodes x_i forward through the network, and calculate the output as given by Equation (4.2).
- **Step 2**: For each output node, calculate the error as Equation (4.1).
- **Step 3**: In this step, the error for the hidden nodes are calculated (backward pass) as a calculation of the gradient descent: $$\delta^j = \begin{cases} E^i f'(y^i) & for\, output \quad i \\ f'(y^j)\sum \delta^{j+1} w^{j+1} & for\, hiddenlayer\ j \end{cases} \quad (4.3)$$

  where ***f'*** denotes the differentiation with respect to the arguments.

- **Step 4** (Learning updates): The weights are updated using the results of the forward and backward passes (using Widrow-Hoff learning or the Delta rule)

  $$w^j(t+1)=w^j(t) + \alpha w^j(t-1)+ \eta\delta^j(t)y^{j-1}(t) \quad (4.4)$$

  where t is number of iteration; η is a learning rate; and α is the momentum constant.

This momentum determines the influence of the old update upon the new one. It enables the learning process to persist in a direction of previous iterations, and to reduce the effect of small local optima. The iteration from step 1 to step 4 is continued until the necessary stopping criterion is satisfied.

The multilayer perceptron usually uses nonlinear activation functions in its neurons to define the outputs (Haykin, 1999); and produces a nonlinear relationship between inputs and outputs across the network. Therefore, the multilayer perceptron can be seen as a non-linear pattern recognition technique.

A commonly used form of nonlinear activation function (sigmoidal function) is given by:

$$y_i = f(net_i) = \frac{1}{1+\exp^{-net_i}} \quad (4.5)$$

where $net_i = \sum_{j=1}^{n} w_j x_j + w_0$

This activation function form is used in all thesis experiments in later chapters. Other forms such as hyperbolic tangent and so on can be seen in Schalkoff (1992), or Haykin (1999).

4.2.2. Radial Basis Function

The **radial basis function** has a structure similar to a multilayer perceptron except only one hidden layer is used in its topology. Each hidden unit acts as a local processor that computes a score for the match between input vectors and its connection weights or centres. The linear combination weights connecting hidden units to the outputs are used to produce the final classification (output).

Rather than using the sigmoidal activation function, the hidden units in a radial basis function use a Gaussian or some other kernel functions (see an example in Figure 4.4).

Consequently, radial basis function classification can also be seen as non-linear classification.

A popular form of Gaussian basic function (Haykin, 1999) used with parameter centre c and width δ (scalar value) of input vector x can be given by:

$$G(t) = e^{-\|x-c\|^2/(2\delta^2)} \quad (4.6)$$

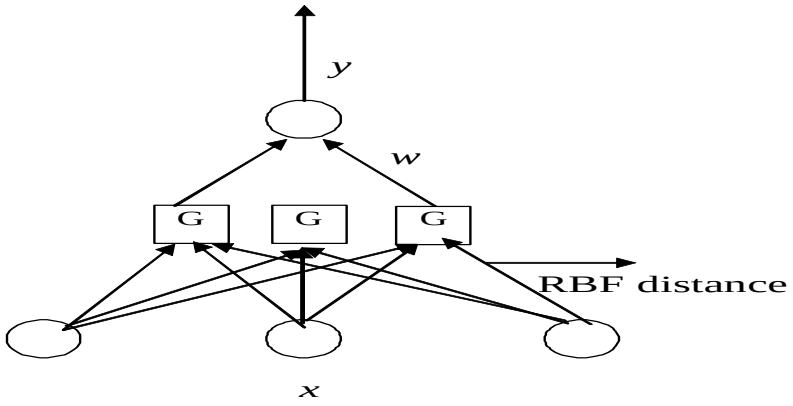

Figure 4.4: The example structure of radial basis function networks.

The main points in comparing radial basis functions with multilayer perceptrons are as follows:

- **One hidden layer**: Radial basis functions contain only one hidden layer whereas multilayer perceptrons might have more than one hidden layer. This enables multilayer perceptrons to, arguably, deal with more sophisticated classification problems.
- **Faster classification:** Multilayer perceptron inputs are weighted and summed before using the activation function whereas radial basis function pass its inputs to activation functions before weighting and summing. This means the multilayer perceptron uses global non-linear calculations between the inputs and the outputs

whereas the radial basis function only uses (locally) non-linear calculations in its hidden nodes but linear calculations in its output layer. Moreover, the multilayer perceptron requires a supervised training method in its algorithm whereas unsupervised techniques can be used in radial basis function to determine the basis functions (Bishop, 1995). Therefore, the radial basis function usually classifies the outputs faster than the multilayer perceptron. However, this might cause poorer classification because of the linear limitations as indicated in Chapter 3.

- **Disadvantage of localised minima**: According to Haykin (1999), radial basis functions transform the nonlinear data from input space into output space whereby the data becomes linear. This means the radial basis function uses the localised functions (local approximations) in attribute space whereas multilayer perceptron uses the long-range functions (global approximations) in its models. Hence, the radial basis function network might experience problems associated with local minima. This is not discussed further in this thesis. More detail can be seen in Haykin (1999).

4.2.3. Support Vector Machine

Support vector machine was developed by Vapnik and co-workers (Boser et al, 1992, Cortes and Vapnik, 1995), as a system for efficiently training linear learning machines in kernel-induced feature spaces. One of the key concepts in support vector machine is the definition of support vectors. These vectors help the network to classify clearly alternative output classes in high dimensional attribute space.

Support Vectors

Assume that a pattern data set can be described in an m-dimensional feature space. The idea of a support vector machine is to build a hyper-plane to separate the positive and the negative patterns in a given data set. This hyper-plane can be seen as a decision surface. The training points that are nearest to this hyper-plane can be seen as support vectors (see Figure 4.5).

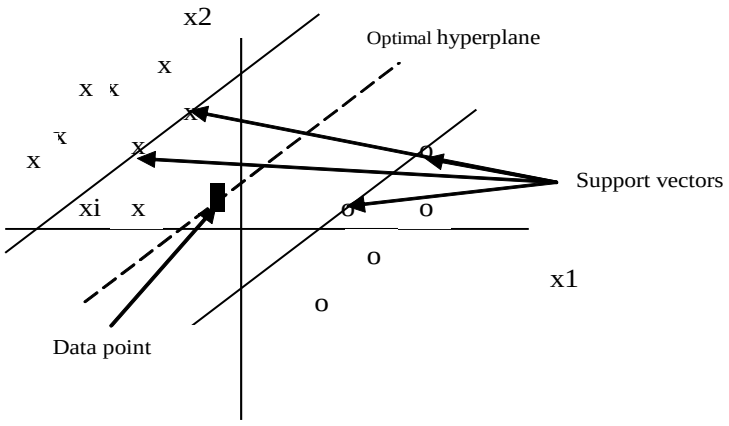

Figure 4.5: The description of support vectors.

The simplest model is a **maximal margin classifier,** which can be seen as the main building block for the later support vector machines (Cristianini and Shawe-Taylor, 2000).

Building a Support Vector Machine for Pattern Recognition

The key to understanding support vector machines is to see how it produces optimal hyper-planes to separate the patterns. According to Haykin (1999), two operations to build a support vector machine can be summarized as:

- **Map data to higher dimensional space**: It is a non-linear mapping based on Cover's theorem (Cover, 1965). This means the following two conditions need to be satisfied:
    - The transformation is non linear;

- And the dimensionality of the feature space is high enough.

- **Construct an optimal hyper-plane to separate the patterns:** This construction is based on the use of an inner-product kernel to define a linear function separating the vectors in feature space. Therefore, the hyper-plane can be formed as:

$$\sum_{j=1}^{m} w_j \varphi_j(x) + b = 0 \quad (4.7)$$

where x is a vector in input space, $\{\varphi_j(x)\}_{j=1}^{m}$ is a set of non-linear transformation vectors in feature space, w_j are the vector weights, and *b* is the bias.

Haykin (1999) introduced the inner-product kernel as $K(x, x_i) = \boldsymbol{w}^T \boldsymbol{x}_i$ in order to reformulate the hyper-plane. The Equation is given by:

$$\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b = 0 \quad (4.8),$$

where α_i are called Lagrange multipliers, *b* is the bias, $x = \{x_i\}$ is a pattern vector, and (x_i, y_i) are training vectors in the input space.

More detail on how to find the Lagrangian multipliers to define maximal margin hyper-plane can be seen in Haykin (1999). Examples of some inner-product kernels are given in Table 4.1. The expanding of inner-product kernels can be seen in Mercer (1909); and Courant & Hilbert (1970).

Type of support vector machine	**Inner-Product kernel $K(x,x_i)$, $i=1,2,..N$**
Polynomial learning machine	$(x^T x_i + 1)^p$
Radial-basis function network	$\exp(-\frac{1}{2\delta^2}\|x - x_i\|^2)$
Two layer perceptron	$\tanh(\beta_0 x^T x_i + \beta_1)$

Table 4.1: Some inner-product kernels (Haykin, 1999).

Figure 4.6 shows an architecture example of a support vector machine. The inner-production kernel functions are used in a hidden layer. The output can be calculated by using an activation function for the hidden nodes and the bias *b*.

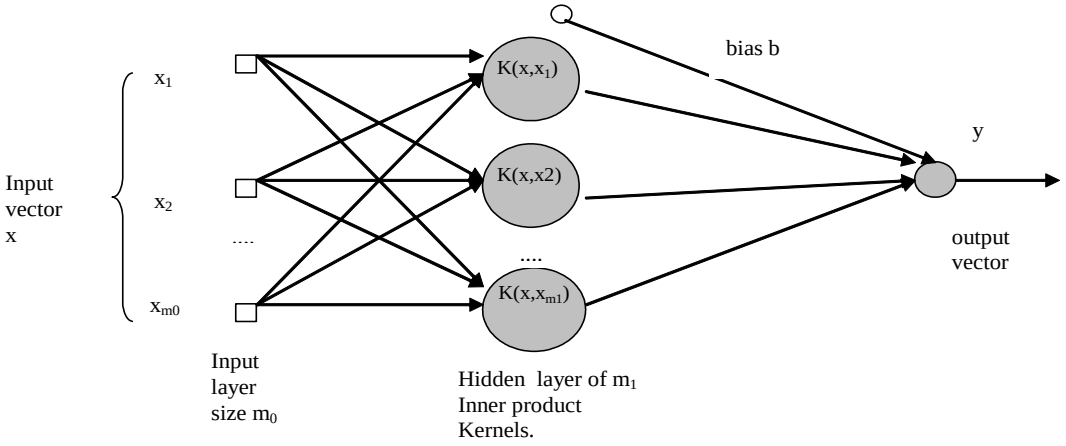

Figure 4.6: Architecture of support vector machine (Haykin, 1999).

4.2.4. WEKA Software Tool Introduction

WEKA software package (WEKA, 2005) is a tool for machine learning and data mining, which is implemented with object-oriented Java class hierarchy. The data set is represented in ARFF format file, which consists of a header describing the attribute types and the data as comma-separated list. However, data can be saved in MS EXCEL files as CSV data type (*.CSV). All the thesis data files are saved as this data type.

WEKA can demonstrate many aspects of data mining such as Regression, Association Rules, clustering algorithms, and so on. Neural network techniques such as multilayer perceptron, radial basis function and support vector machine are also implemented in WEKA as alternative classifiers. In general, WEKA implements the basic functions for alternative neural network techniques. For example, radial basic function is implemented in

WEKA using Gaussian basis function; multilayer perceptron is used with sigmoid function; and support vector machine is used with polynomial kernel in polynomial learning machine. All mathematic equations for these can be seen above in section 4.2. The parameters declared in WEKA for each classifier models are chosen upon the neural network techniques' use. For example, radial basis function is used with the declaration of parameter of c (number of centers); multilayer perceptron is used with the declaration of number of epochs, learn rate, and number of hidden nodes; and support vector machine is used with the chosen of alternative types of inner kernels (see Table 4.1).

4.3. Unsupervised Pattern Recognition.

In contrast to the theory of supervised learning introduced above, this section discusses two unsupervised learning methods of pattern recognition and data mining: self organizing maps and the KMIX clustering algorithm. The concept of self organizing maps is introduced with its characteristics of data clustering and projection. Their goal is to map data from a nonlinear space (attribute space) onto the lower (usually 2) dimensional output space. On the other hand, the KMIX algorithm is introduced as an instance of the popular K-means technique adapted for use with the mix of attribute types in the current data domain.

4.3.1. Self Organizing Map

The **self organizing map** (Kohonen, 1981; 1990a, 1990b) is a type of a neural network model that represents and clusters input data onto a lower dimension space (map). According to Haykin (1999), the input topological properties remain in the output space of the map.

One common example of the application of self organizing maps is the visualization of World Poverty Map (see in Kohonen, 1996; Kaski, 1997). Here a self organizing map has been used to classify statistical data describing various quality-of-life factors such as state of health, nutrition, educational services, and so on. The countries with similar quality-of-life factors end up clustered together. Further self organizing map applications can be seen in Kohonen (1996).

How It Works?

As indicated above, the self organizing map provides a topology preserving mapping from high dimensional input space onto a map of units (neurons) (see Figure 4.7). Note that the property of topology preserving means the mapping will preserve the relative distance between map points. These points, those near each other in the input space, are mapped to nearby map units in the self organzing map. Therefore, a self organizing map can serve as a cluster analyzing tool of high-dimensional data. Furthermore, it has the capability to recognize, generalize and characterize the input data. In the output space of a self organizing map, adjacent neurons are connected to each other by the neighborhood relation according to a pre-defined radius. Generally, two types of maps, either rectangular or hexagonal, are used to present its topology (or structure).

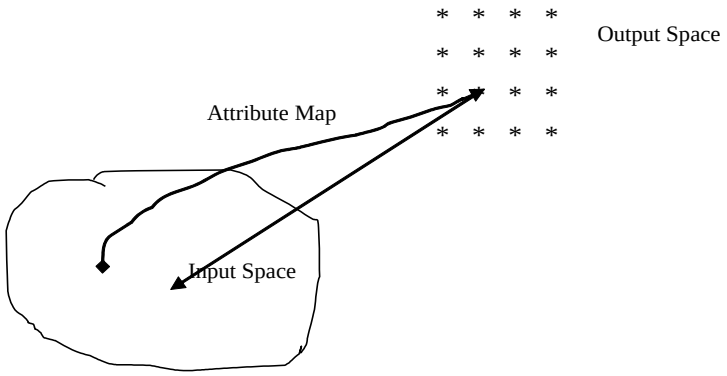

Figure 4.7: The description of a self organizing map technique.

The visualisation of a map can be seen in the U-matrix representation (unified distance matrix) (Ultsch et al, 1990; Ultsch, 1993). This matrix visualizes the distances between neurons. The distance between adjacent neurons is calculated and presented with different shades of colour. A light shade of colour between the neurons corresponds to a large distance, signifying a big gap between codebook values in the input space. Conversely, a dark shade of colour between the neurons signifies that the codebook vectors are close to each other in the input space. In other words, dark shading areas might be thought as clusters and the light shading areas as cluster separators (see an example of U-matrix in Figure 4.8).

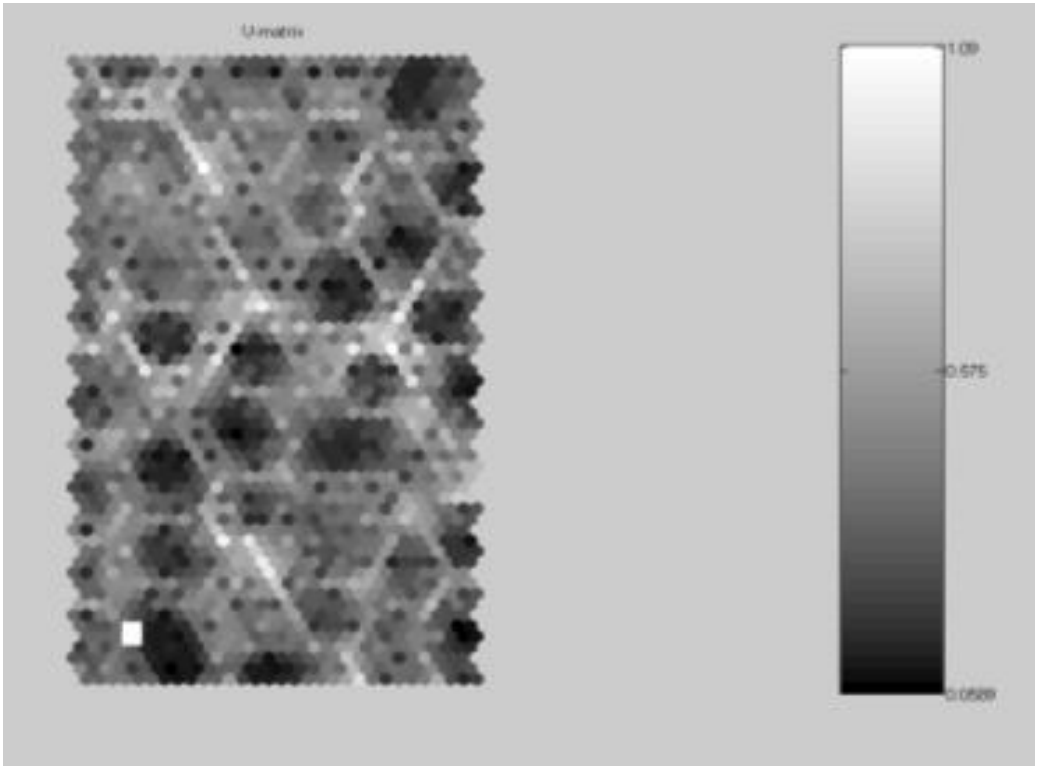

Figure 4.8: The example of the final U-matrix.

Algorithm of Self Organizing Map

Assume that $x = \{x(t)\}$, $t = 1, 2,..n$ is a vector in the input space, and the map is represented in k x d matrix (map size), and the weight can be seen as:

$$W = \{w_1, w_2,..., w_k\}.$$

Step 1: Initialize weight vectors w_i, $i = 1,..,k$.

Step 2: For each map unit i $(i = 1,..,k)$: Choose $x(t)$ from input space $(t = 1,2,..n)$; Calculate the distance from x(t) to each node i; Find the winer node c where the distance between the inputs and this node is minimized.

$$||x(t) - w_c(t)|| \leq ||x(t) - w_i(t)||, \quad i = 1,..,k \qquad (4.9)$$

Update the weight as follows:

$$w_i(t+1) = w_i(t) + \alpha * G_{ci}(t) * (x(t) - w_i(t)) \qquad (4.10)$$

where α is learning rate, $G_{ci}(t)$ is neighborhood function (Gausian function), and t = 1,2,...,n.

Step 3: Repeat step 2 until the algorithm time is terminated or all input vectors are tested.

Map Quality Measurement.

According to Kohonen (1995); and Kiviluoto (1996), the Davies-Bouldin index (Davies and Bouldin, 1979) is used to define the Average Quantization Error *(AQE*) and ToPographic Error (TPE) of the maps. The average quantization error measures the average error of distances between each pattern (in the data set) and its best matching unit on the map. It also can be seen as a measure of a map resolution and given by:

$$AQE = \frac{1}{N}\sum_{i=1}^{n}\left\|x_i - w_{ic(x)}\right\| \qquad (4.11)$$

where N is total of number of patterns, $w_{ic(x)}$ is the winner unit in the map of input vector x_i.

Therefore, the accuracy of the map can be calculated as:

$$Acc = \frac{1}{1 + AQE} \quad (4.12)$$

According to Kohonen (2001), topographic error is a proportion of all data vectors for which the first and second best matching units are not adjacent units. It is given by:

$$TPE = \frac{1}{N}\sum_{i=1}^{n} u(x_i) \quad (4.13)$$

where

$$u(x_i) = \begin{cases} 1 & \text{if the first and the second best matching units of } x_i \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases}$$

4.3.2. KMIX Algorithm

This section introduces the KMIX algorithm, which can be seen as an instance of the K-means technique. This algorithm uses both Euclidean and the Hamming dissimilarity measurements instead of just the Euclidian measurements of the K-means. These measures will be adequate for the mixture of attribute types in the thesis data.

4.3.2.1. Introduction and Notations

Clustering analysis is a machine learning area of particular interest to pattern recognition data mining. The resulting data partition improves the data understanding, and reveals its internal structure. Clustering has been used in many application domains including biology, medicine, and so on, as indicated in Chapter 3. Further clustering applications can be seen in Dunham (2002); Berkhin (2002); Mirkin (2005).

One well-known partition clustering algorithm is the K-means algorithm (Forgey, 1965; Jancey, 1966; MacQueen, 1967; Hartigan, 1975; Hartigan and Wong, 1979) which, according to Berkhin (2002), is the most popular clustering tool used in scientific and industrial applications. The main duty of K-means is to partition n patterns in input space into k clusters such that similar patterns belong to the same clusters; and dissimilar patterns will belong to alternative clusters.

Assume that X is a pattern (observation, case, or patient record in data set). X typically consists of m components: $X = (x_1, x_2, \ldots, x_m) = (x_j)_{j=1,..m}$. Note that each component in multidimensional space is an attribute (continuous or discrete) in the data domain. Therefore, we have a $n \otimes m$ pattern matrix, where n patterns $\{X_i\}_{i=1..n}$; $X_i = (x_{i,1}, x_{i,2}, .., x_{i,m})$, and m attributes.

Similarity Measurements

A similarity measurement is the strength of relationship between two patterns in the same multidimensional space. It can be represented as $sim_{ij} = sim(x_i, x_j)$, $i,j = 1,2,.. n$. According to Gower (1985), a similarity is regarded as a symmetric relationship. This means $sim(x_i, x_j) = sim(x_j, x_i)$. Contrastingly, dissimilarity measures of patterns have been introduced as the complement of similarity measures. A list of dissimilarity measures can be seen in Gower (1985). The dissimilarity measures used in this thesis are as follows:

- **Continuous attributes**: the most common measure used is the Euclidean distance between two patterns. It is given by:

$$dissim(x_i, x_j) = [D(x_i, x_j)]^2 = \sum_{k=1}^{m_1} (x_{ik} - x_{jk})^2, i, j = 1,2..n \quad (4.14),$$

where D is Euclidean distance, $m_1 <= m$ (m_1 is number of "continuous" attributes).

- **Discrete (categorical) attributes**: the similarity measure between two patterns depends on the number of similar values in the categorical attribute (Kaufman & Rousseeuw, 1990). This means the dissimilarity is a number of different values in this categorical attribute. It is given by:

$$dissim(x_i, x_j) = d(x_i, x_j) = \sum_{k=1}^{m_2} \theta(x_{ik}, x_{jk}) \quad i, j = 1,2..n \quad (4.15);$$

where $\theta(x_{ik}, x_{jk}) = \begin{cases} 0 & if\ x_{ik} = x_{jk} \\ 1 & if\ x_{ik} \neq x_{jk} \end{cases}$, $k = 1,2,..m_2; i, j = 1,2.., n$, and m_2 is number of categorical attributes.

- **Boolean attributes**: The dissimilarity measures are calculated as in the categorical or continuous attributes according to the interpretation of the attribute.

Centre Vectors

Assume that the data attribute set includes continuous and discrete attributes. Note that Boolean attributes are treated as continuous or discrete as indicated above. Therefore, there are two types of centre vectors. Assume that m attributes contains the p first continuous attributes; and m-p remaining discrete attributes. This means each pattern X in the input space can be seen as:

$$X=(x_{i1}, x_{i2,}\ldots x_{ip}, x_{ip+1,} x_{ip+2,}\ldots x_{im}))$$

If Q is a centre vector for the sub data set C, Q can be represented as:

$$Q= (q_{j1,} q_{j2,}\ldots q_{jp}, q_{jp+1,} q_{jp+2,}\ldots, q_{jm})$$

The task now is to find p continuous attribute values, and m-p discrete attribute values for centre vector Q. According to Han (1981), these centre attribute values can be calculated as follows:

- **Continuous attribute**: The centre values $\{q_{jk}\}_{k=1,..p}$,=$\{\text{mean}_k\}$, where mean_k is the average of k^{th} attribute.

- **Discrete attribute**: The centre values $\{q_{jk}\}_{k=p+1,.., m}$ =$\{\text{mode}_k\}$, where $mode_k$ is the "mode" of k^{th} attribute.

Definition 1: A vector Q is a "mode vector" of a data set $C = (X_1, X_2, \ldots X_c)$, $c<=n$ if the distance from each vector $X_{i, i=1,..,c}$ to this vector is minimized.

This means

$$d(C,Q) = \sum_{i=1}^{c} d(X_i, Q) \quad (4.16)$$

is minimized.

According to the theory in Huang (1998), the Equation (4.16) is minimized if the frequency of value q_k in data set C, for k^{th} attribute, is equal or greater than the frequency of all different x_{ik} such that $x_{ik} \neq q_k$. Therefore, we can choose the mode vectors of m-p attributes as the highest frequency values in these attributes. Their forms can be seen as follows:

$$\{q_{jk}\} = mode_{k=}\{\text{"max freq" } Val_{Ck}\},\ k=p+1,.., m. \quad (4.17)$$

Accuracy Measure

The accuracy (Acc) for measuring the quality of clustering algorithm is given by:

$$Acc = \frac{\sum_{i=1}^{K} a_i}{n} \quad (4.18)$$

where n is the number of samples in the dataset, a_i is the number of data samples occurring in both cluster i and its corresponding class, and K is number of clusters.

Consequently, the clustering error *(err)* is defined as:

$$err = 1 - Acc \quad (4.19)\ .$$

4.3.2.2 KMIX Algorithm

Step 1: Initialise K clusters according to K partitions of data set.

Step 2: Update K centre vectors in the new data set (in the first time, the centre vectors are calculated)

$$Q_j = (q^N_{j1}, q^N_{j2}, \ldots, q^N_{jp}, q^C_{jp+1}, \ldots, q^C_{jm}), \; j = 1, 2, \ldots, k$$

where $\{q^N_{ji}\}_{i=1,2..p} = \{mean^N_{ji}\}$ (mean of i^{th} attribute in cluster j),

and $\{q^C_{ji}\}_{i=p+1,..m} = \{mode^C_{ji}\}$ (max freq value in attribute i^{th} in cluster j).

Step 3: Update clusters as the following tasks:

- Calculate the distance between X_i in i^{th} cluster to K centre vectors:

$$d(X_i,Q_j) = d^N(X_i,Q_j) + d^C(X_i,Q_j); \; j=1,2,..k$$

where $d^N(X_i,Q_j)$ is calculated according to Equation (4.14),

and $d^C(X_i,Q_j)$ is calculated according to Equation (4.15)

- Allocate X_i into the nearest cluster such that $d(X_i,Q_j)$ is minimised.
- Repeat for whole data set, and save them to the new data partition with K new centre vectors.

Step 4: Repeat step 2 and 3 until no more change in the distance between X_i and new K centre vectors.

4.3.2.3. Standard Dataset Comparisons

To verify the KMIX algorithm, alternative data sets from the UCI repository (the empirical analysis of machine learning algorithms- Merz & Merphy, 1996) are used. KMIX is applied to alternative data types such as discrete numerical (Small Soybean; Michalski and Chilausky, 1980); mixture of numerical and categorical (Zoo small; Richard, 1990; Merz & Merphy, 1996); and all categorical (Votes data set; Jeff, 1987). The experiments are

performed using randomised data sets. The discussions are based on the error rates (or accuracy rates) in these experiments. Furthermore, a comparison between the KMIX results, publicised results, and standard K-means is shown to discuss the use of this algorithm for the thesis data.

Small Soybean

This data set contains 47 samples and 35 attributes. The experimental results can be seen in Table 4.2. Overall, the error rates are negligible (in average of 0.04).

Experimental types	***Actual classes/Clusters***	**C1**	**C2**	**C3**	**C4**	**Error**
Original	*D1*	10	0	0	0	0 (0.00%)
	D2	0	10	0	0	
	D3	0	0	10	0	
	D4	0	0	0	17	
Rand 1	*D1*	0	10	0	0	0.02 (2.13%)
	D2	10	0	0	0	
	D3	0	0	0	10	
	D4	0	0	16	1	
Rand 2	*D1*	0	0	10	0	0.04 (4.26%)
	D2	0	10	0	0	
	D3	10	0	0	0	
	D4	2	0	0	15	
Rand 3	*D1*	0	10	0	0	0.08 (8.51%)
	D2	10	0	0	0	
	D3	0	0	10	0	
	D4	0	0	4	13	
Rand 4	*D1*	10	0	0	0	0.06 (6.38%)
	D2	0	0	0	10	
	D3	0	2	8	0	
	D4	0	16	1	0	

Table 4.2: The experiment results with Small Soybean Data.

Zoo Small

This data contains 101 cases distributed in 7 categories and 18 attributes (15 Boolean, 2 numerical, and 1 unique attributes). The results in Table 4.3 show the accuracy and error for 10 experiments with alternative randomised data sets for the KMIX algorithm. The average error is about 0.15 (0.84 of accuracy). This error result is similar to the results of Shehroz and Shri (2007) (0.166).

Random experiments	**Accuracy**	**Error**
Rand0	0.90	0.10
Rand1	0.75	0.25
Rand2	0.90	0.10
Rand3	0.82	0.18
Rand4	0.90	0.10
Rand5	0.88	0.12
Rand6	0.81	0.19
Rand7	0.84	0.16
Rand8	0.87	0.128
Rand9	0.80	0.198
Average	**0.84**	**0.15**
Std Deviation	**0.05**	**0.05**

Table 4.3: 10 test results of Zoo data set in randomisation.

Vote data

The data set contains 435 records with 2 output classes labelled as 168 “republicans” and 267 “democrats”. Table 4.4 shows the experimental results from 10 randomised data sets. Overall, the accuracy is about 86%. The sensitivity rates are quite consistent (an average of 0.95) over the experiments except in Rand0 (0.83). Figure 4.9 shows that although the

experiments fluctuate in sensitivity and specificity rates their accuracy and the error rates remain consistent.

		C1	C2	Sen	Spec	Acc	Err
Rand0	*Democrat*	222	45	0.83	0.92	0.86	0.13
	Republican	14	154				
Rand1	*Democrat*	261	6	0.98	0.69	0.87	0.13
	Republican	52	116				
Rand2	*Democrat*	257	10	0.96	0.67	0.85	0.14
	Republican	55	113				
Rand3	*Democrat*	257	10	0.96	0.67	0.85	0.14
	Republican	55	113				
Rand4	*Democrat*	253	14	0.95	0.73	0.86	0.13
	Republican	45	123				
Rand5	*Democrat*	261	6	0.98	0.69	0.87	0.13
	Republican	52	116				
Rand6	*Democrat*	257	10	0.96	0.67	0.85	0.14
	Republican	55	113				
Rand7	*Democrat*	252	15	0.94	0.74	0.86	0.13
	Republican	44	124				
Rand8	*Democrat*	257	10	0.96	0.67	0.85	0.14
	Republican	55	113				
Rand9	*Democrat*	257	10	0.96	0.67	0.85	0.14
	Republican	55	113				
Average				**0.95**	**0.71**	**0.86**	**0.14**

Table 4.4: Results of 10 executions of Votes recording data set.

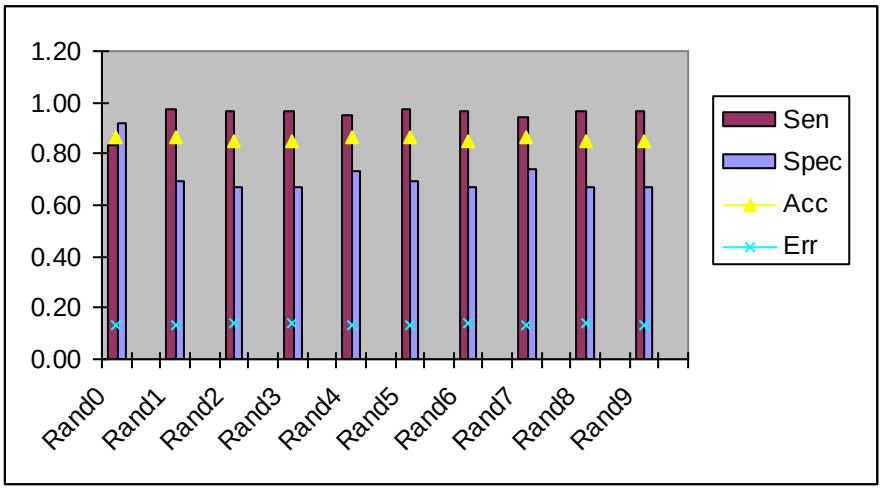

Figure 4.9: Vote data results in sensitivity, specificity, accuracy and error rates.

Discussions

The comparison results based on the error rates can be seen in Table 4.5 for the KMIX algorithm with the others of recent publications and K-means. A WEKA software package is used here for the K-means algorithm. Note that the data sets used in K-means are the original data (no random ordering of the data); and all data is transformed into numerical data type.

Data set	Data Type	Publication results	KMean	KMIX
Soy Bean	Integer	0.11 ~	0.23	0.07
Votes	Categorical	0.132	0.14	0.14
Zoo small	Mixed	0.166	0.22	0.15

Table 4.5: Publication comparisons.

From Table 4.5 the KMIX performs as well as other published results for the Soy Bean (Ohn et al, 2004); Votes (Shehroz and Shri, 2007; Zengyou, 2005), Zoo small (Jeff,1987). Furthermore, the KMIX achieve better results compared to the standard K-means algorithm except for the use of the Votes data set. For example, the K-means clustering achieves the performance at 77% (error rate of 0.23) whereas the KMIX performance is at 93% for the

small soybean data (with the use of original data set). Therefore, the KMIX algorithm seems suitable for use with the thesis data.

4.4. Summary

This chapter focuses on three neural network techniques: Multilayer perceptron, radial basis function and support vector machine, as examples of non-linear classifiers. The multilayer perceptron produces the outputs via the global non-linear calculations between the inputs and the outputs. The radial basis function uses locally non-linear calculations in its hidden nodes, and linear calculations in its output layer to produce the outputs. The support vector machine classifies alternative output classes in high dimensional attribute space (optimal hyper-planes). Multilayer perceptron usually use non linear (typically sigmoidal) activation functions while radial basis function usually use Gausian functions and support vector machine can use both activation function types for its inner-product kernel (see Table 4.1).

Two methods of clustering data, self organizing maps and the KMIX algorithm, are described as examples of unsupervised learning pattern recognition techniques. The self organizing maps clustered and mapped the patterns onto the final map (U-matrix) where the dark grey units show the close distance of patterns in the input space. In other words, these patterns can be seen as in the same clusters. KMIX, as an example of the K-means algorithm, is compared to publicised results from standard machine learning data to show its ability to deal with alternative data types.

These supervised and unsupervised learning techniques will be discussed in greater detail in the case studies in Chapter 6 which make use of cardiovascular data.

Chapter 5

Data Mining Methodology and Cardiovascular Data

5.1. Introduction

This chapter describes the general process of "knowledge discovery from data" (Hand et al, 2001), gives the definition of data mining, and discusses data mining methodologies. Two popular data mining methodologies, CRISP_DM (Shearer, 2000) and SEMMA (SAS, 2008), are introduced as examples of data mining methodologies. An adopted methodology, based on Davis (2007), is described and used in this thesis.

The thesis data mining framework steps are described in detail. The strategy for preparing experimental data is also presented in this chapter. Three types of thesis experiments are explained in a systematic way to demonstrate the use of the thesis methodology in detail. As part of the data mining methodology, the structure, and nature of the data used in this thesis is analysed. The thesis data contains cardiovascular patient information collected between 1982 and 1999 from clinical sites in Hull and Dundee. Partial data from each of these sites is used in the thesis case studies. The combined data from both these sites is used in the main thesis experiments.

5.2. Data Mining and Thesis Methodology

5.2.1. What is Data Mining?

There are many definitions of data mining. Hand et al (2001) produced a general definition as follows:

> "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."
>
> Hand et al (2001); MIT Press. Cambridge. Chapter 1; Page: 1.

On the other hand, Tan et al (2006) defined that data mining provides a way to get the information buried in the data. This means it finds patterns hidden in large and complex collections of data, where these patterns elude traditional statistical approaches to analysis. In medical domains, data mining can be seen as a pattern recognition system to predict patient risks from patient records.

Hand et al (2001) and Tan et al (2006) list the various data mining tasks as follows:

- **Exploratory data analysis**: Data is represented in some graphical ways such as graphs, pie charts, plots, and so on.
- **Descriptive modelling**: Data is described using alternative models such as density estimation models, or cluster analysis. This explores the nature of the data and summarizes underlying relationships between patterns in the data.
- **Predictive modelling**: This task is to build a classification and regression model for the target variable as a function of the explanatory variables. The POSSUM and PPOSSUM systems described in Chapter 2 are examples of such a model. They produce patient risks based on logistic regression functions of the physiological scores and operative severity scores.
- **Discovering patterns and rules**: This task is to detect patterns in the data. For example, this task is to find the unknown patterns in a data domain; or to find the rules between values for data attributes.

- **Content retrieval**: This task means finding the patterns hidden in the data domain.

5.2.2. Data Mining Methodology and Criteria

One perspective on the process of the "knowledge discovery from data" (Hand et al, 2001) can be seen in Figure 5.1 below, where the process is described as the following steps:

(1) **Selection step**: To obtain the raw data from various sources, and then identify the target data of use in the following data mining steps.

(2) **Pre-processing step**: Erroneous data may be identified then corrected, or removed. For example, missing data values can be supplied. This step can be seen as the data cleaning step. Alternatively, filter methods might be used in this step to produce an adaptable data set for the user's requested purposes.

(3) **Transformation step**: The pre-processed data is transformed into a more useable format in order to be easily used with techniques of the later steps.

(4) **Data mining step**: Based on user's purpose and the tasks being requested, appropriate data mining techniques are used on the transformed data set. This step involves the use of pattern recognition techniques to produce classifications or clusters or whatever alternative outcomes are required.

(5) **Evaluation/Interpretation step**: The patterns produced from the data mining step are evaluated by standard measurements such as mean square error, confusion matrix, and so on. An interpretation method is then applied to produce the meaningful and clarified knowledge for the end users.

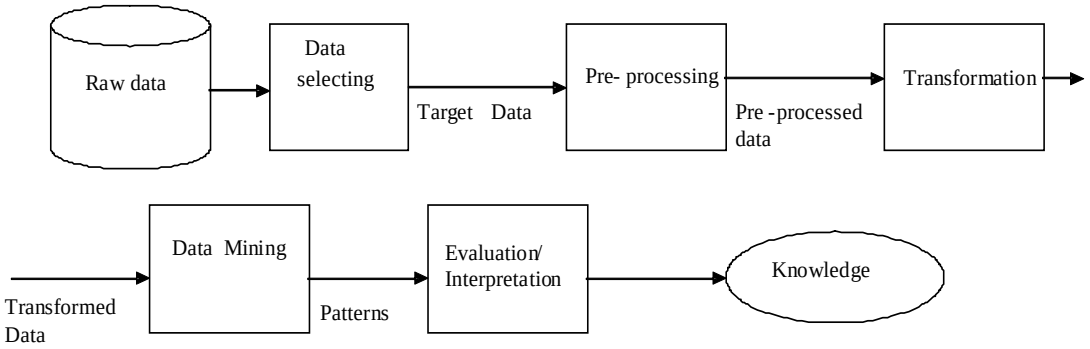

Figure 5.1: Overview of a "knowledge discovery from data" process (Hand et al, 2001).

It is clear from Figure 5.1 that, data mining is a particular step in the "knowledge discovery from data" process whereby a set of specific techniques are used to extract patterns from the transformed data.

What Is A Data Mining Methodology?

A data mining methodology is a system or a strategy for using alternative techniques to take raw data to a transformed data set in order to produce knowledge for users.

What Are Criteria?

How is the right methodology for a "knowledge discovery from data" process chosen? It has to be satisfied the following two main conditions:

- **Adequate:** The methodology has to be adequate to the data domain. This ensures the techniques can effectively mine the data.
- **Accuracy extractions:** The extracting patterns should be accurate and satisfactory with user's requirements.

Therefore, the criteria for a data mining methodology can be summarized as:

- **Right techniques choices:** This means the chosen techniques are appropriate for the selected data set. The criterion might be represented by questions such as: Are the

techniques suitable for the data domain, or easy to use? For example, in CRISP-DM (Shearer, 2000), appropriate models are chosen according to the relationships defined by association rules; or in this thesis framework, the supervised techniques are used whenever there is a request for risk prediction.

- **Correctness of measurement methods**: The use of correct measurement methods is another criterion in determining data mining performance. These measures are to ensure the classification results are both accurate and trustworthy. Standard measures such as *mean square error, confusion matrix*, s*ensitivity and specificity rates, the positive predictive value, and negative predictive value* are used in this thesis.
- **Usefulness of end results:** The results of data mining process should be meaningful for the user's purposes. For example, the unsupervised classifiers' results in this thesis are shown the internal structure of data in the data domain.

5.2.3. Examples of Data Mining Methodologies

There are two popular existing data mining methodologies for the "knowledge discovery from data" process (KDNuggets, 2007): CRISP_DM (Shearer, 2000), and SEMMA (SAS, 2008). CRISP-DM is being developed by an industry led consortium as the CRoss-Industry Standard Process Model for Data Mining (see Figure 5.2). It consists of a set of tasks described at four levels from general to specific (Chapman et al, 1999). At the top level, the data mining process is organized into a number of phases where each phase consists of several generic tasks at the second level. The second level includes generic tasks which can cover all possible data mining situations such as the process tasks, possible data mining applications, and techniques. In the third level, the specialized task shows detailed actions

in generic tasks for certain specific situations. For example, if the generic task is a "*dealing with missing data"*, the more detailed tasks in the third level will be a category of specialized missing data tasks namely "*dealing with missing numeric values"*; "*dealing with missing categorical values"*; and so on. The fourth level, as the process instance, is a record of the actions, decisions, and results of an actual data mining engagement.

An example of the use of CRISP-DM methodology is the predictive modelling, association rule use, and sequence detection to predict the onset and successful diagnosis of thrombosis (Jensen, 2001).

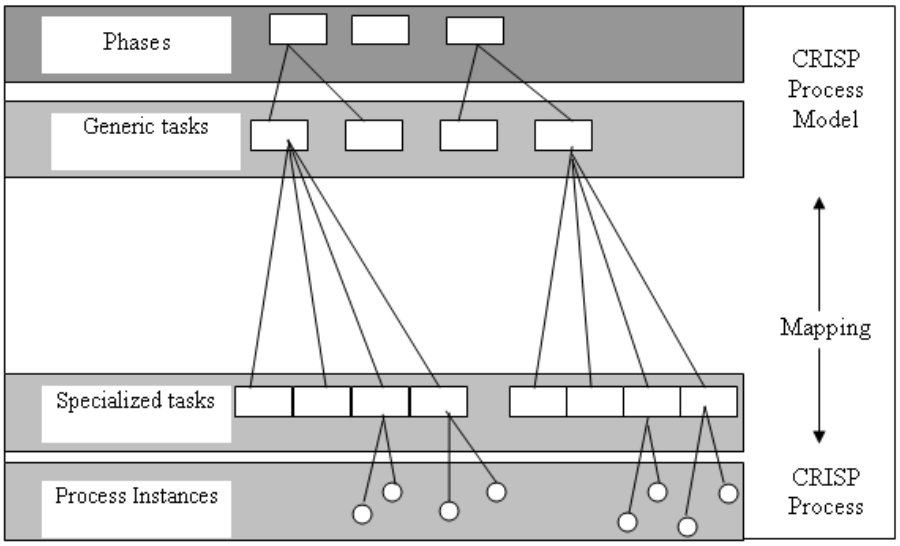

Figure 5.2: The methodology of CRISP-DM (Shearer, 2000).

SEMMA is a data mining methodology derived from the Statistical Analysis Software Institute (SAS, 2008) consisting of the five steps: Sample, Explore, Modify, Model, and Assess (SEMMA). All cases from data set are taken and partitioned into the training, validation and test sets in the Sample step. The Explore step allows data sets to be visualised statistically and graphically. The Modify step allows the transformation of the data or deals with missing values in the data set. The Model step requires the fitting of the

data mining and machine learning techniques such as decision trees and neural networks. Lastly, the Assess step means using alternative partitions of test sets to validate the derived model in order to estimate how well the data mining process performs. One of the uses of SEMMA methodology is the estimating of a nationwide statistic for hernia operations using the claims database of the Korea Health Insurance Cooperation (Kang et al, 2006). The claims database was divided into the electronic interchange database (EDI_DB) and the sheet database (Paper_DB). SEMMA is used here to produce a predictive model for the sheet database. In this database, the operation and management codes were not shown for the "facts" and the "kinds" of operations whereas they are shown in the EDI_DB. The model predicts potential hernia surgery cases extracted by matching management code from the Paper_DB to appropriate records in the EDI_DB. More detail about the model can be seen in (Kang et al, 2006).

5.2.4. Thesis Methodology

The data mining methodology adopted for this thesis can be seen in Figure 5.3. This methodology is derived from one developed for the teaching of data mining and decision systems (Davis, 2007). The following are reasons for using this specific methodology:

- **The existing methodologies are not suitable for the thesis data:** As described above the CRISP-DM and SEMMA methodologies are too big and too complicated for use with the thesis data domain. For example, CRISP-DM contains the "Business Understanding" and "Deployment" phases whereas the thesis methodology does not. SEMMA includes the task of representing data sets statistically and graphically, again not required for the purposes of this thesis.

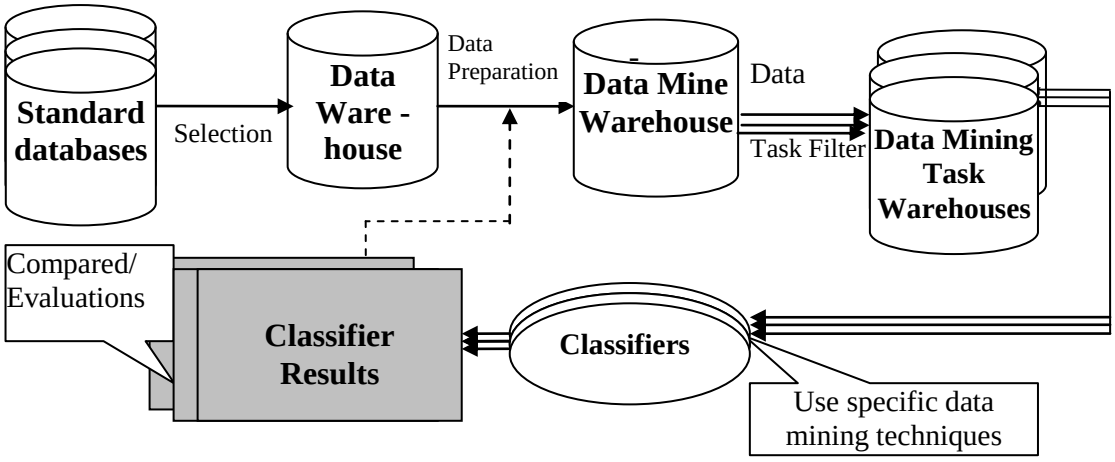

Figure 5.3: A general thesis frame work; based on (Davis, 2007).

- **Demonstrating thesis objectives**: This thesis focuses on investigations in the use of alternative pattern recognition and data mining techniques in the cardiovascular data. Therefore, its objectives concentrates on the demonstrations and evaluations of supervised versus unsupervised techniques with the thesis data.

The thesis methodology can be seen in the following steps:

- **Step 1 (Selection):** The data set relevant to the thesis experiments is chosen from various sites in the "Standard Databases".
- **Step 2 (Data Preparation):** Data is analysed by using data mining methods in order to define how the data is to be made more meaningful and useable for the classification techniques used in later steps. For example, data is cleaned by supplying missing values; or data is transformed to more appropriate value types such as numerical for the use of neural network techniques.
- **Step 3 (Data Task Filter):** Whenever a specific data mining task is requested, the data set in the "Data Mine Warehouse" is selected. Heuristic decision rules are

applied in this step to define expected outcomes for the prediction in later steps. The selected data set is then stored in the "Data Mining Task Warehouses".

- **Step 4 (Data Mining Techniques):** In this step, a suitable classifier is chosen with an appropriate data set for the task requested in step 3. For example, the clustering algorithm KMIX is chosen to fulfil the clustering task requested from the "Data Task Filter" step.
- **Step 5 (Comparison/ Evaluation):** The classified results are compared or evaluated based on standard measures such as *mean square error, confusion matrix, sensitivity and specificity rates, the positive predictive value, and negative predictive value.*
- **Step 6 (Building New Models):** In some specific circumstances, unsupervised clustering results might be used for the next prediction task. This means clustering outcomes can be seen as the expected outcomes for the next classification process. In this step, clustering results are combined with an appropriate data set to create a new model (clustering model). The data set is then stored in the "Data Mine Warehouse" for further prediction processes. Other tasks are then repeated from step 3 to step 5.

5.3. Application of Data Mining Methodology

This section introduces the thesis data collected from the Hull and Dundee sites. The analysis of this data, and the detailed preparation data steps for the thesis experiments are also shown in this section.

5.3.1. Cardiovascular Data

Hull Site Data

The data from this site includes 98 attributes and 498 cases of cardiovascular patients. The detailed structure of the original data can be seen in Appendix A. The main characteristics of the given data are as follows:

- **Redundant attributes**: These are the date and time attributes with mostly null values; or explanatory attributes; and so on. These attributes bear little relevance to the thesis experiments, or the a-priori outcome models. These attributes will be eliminated in the selection stage (see detail in “Data Selection Strategy” section).
- **Missing values**: The data has 7018 out of 42914 cells (16%) with missing values after removing the redundant attributes indicated above. The method of dealing with missing values will be shown in “Data Selection Strategy” section.
- **Noisy and inconsistent data**: These are abbreviations in categorical attributes and outlier values in some numerical attributes. For example, the attribute “CAROTID_DISEASE” includes a mixture of abbreviated and fully specified values such as “asymptomatic carotid disease”, “Asx”, and so on. In fact, both these values have the same meaning. Therefore, these inconsistent entries are harmonised as single values.
- **Scoring values**: The original data included the physiological score and operative severity score values taken from the POSSUM and PPOSSUM systems. Furthermore, the data in this site includes enough information for use separately in the POSSUM and PPOSSUM calculations.

The valid or missing value frequencies of some significant attributes can be seen in Table 5.1. These attributes are labelled as "PATIENT_STATUS"; "Heart Disease”; "Diabetes”;

and “Stroke”. These attributes are highlighted in the research of collaborative clinicians (Kuhan et al., 2001), as some of the main factors expected to contribute to the outcomes for patient risks.

		Diabetes	Heart Disease	Stroke	PATIENT_STATUS
Number	*Valid*	497	497	497	498
	Missing	1	1	1	0

Table 5.1: The frequencies of significant attributes in the Hull site data.

It is clear from Table 5.1 that there is one case (1 out of 498) that includes missing values in all the significant attributes except the “PATIENT_STATUS” attribute. However, the “PATIENT_STATUS” attribute is the most significant, and this attribute will be the main factor for outcome calculations in later chapters. Therefore, this case with some missing values will not be eliminated. Its missing values will be filled by the use of data mining method (see detail in “Data Selection Strategy” section).

Dundee Site Data

This data includes 57 attributes, and 341 cases from cardiovascular patients at the Dundee site. The detailed structure of the original data can be seen in Appendix A. This data site has similar characteristics to the Hull site such as redundant attributes, missing values, and noisy values. The method of data treatments such as elimination of redundant attributes, filling the missing data, and so on is based on the strategy indicated in “Data Selection Strategy” section below. The main characteristics can be seen as follows:

- **Redundant attributes**: For example, the attribute “*ADMISSION_DATE*” shows patient’s operation date; or the two attributes “*Surgeon name1*” and “*Surgeon*

name2" represents names of operating doctors. Their values might be helpful in a general evaluation, but offer little relevance to the specific purposes of this thesis.

- **Missing values**: The data includes 1912 output of 12311 cells with missing values (16%) after deletion of the above redundant attributes.
- **Noisy and inconsistent data**: As an example of numerical outlier values, the attribute "*PACK YRS*" has a big gap between the maximum value of 160, and the minimum value of 2. This affects the transformation process as it unduly changes the mean of the attribute values.
- **Scoring values**: The site does not include the scored values (physiological score, and operative severity score) from the POSSUM and PPOSSUM systems. Furthermore, the data in this site is insufficient to use with the scoring systems of POSSUM and PPOSSUM, as it lacks information for these systems' variables.

To complete a similar analysis, as with the Hull site, the valid or missing value frequencies for some significant attributes can be seen in Tables 5.2 below.

		30 D stroke/death	**Heart Disease**	**Diabetes**	**Stroke**
N	*Valid*	341	334	341	340
	Missing	0	7	0	1

Table 5.2: The frequencies of the significant attributes for the Dundee site.

Table 5.2 shows that the attribute "Heart Disease" has 7 missing values whereas there is only one missing value for the "Stroke" attribute. The method of dealing with missing values is identical to the method indicated for the Hull site.

5.3.2. Thesis Experimental Steps

The detailed steps of the thesis methodology can be redrawn as shown in Figure 5.4. The process flow and individual steps in Figure 5.4 can be illustrated in detail as follows:

- **Step 1 (Selection):** A data set is selected from one or both of the Hull and Dundee sites, and stored in the "Data Warehouse". Note that data from both sites were collected and stored in various (Excel) computer files in earlier studies. The data here is understood as "Raw data" in the "knowledge discovery from data" process. Therefore, the other process steps such as pre-processing and transformation steps are needed.

Alternatively, a selection of the data set derived from the Hull site is stored in the "POSSUM, PPOSSUM Data Warehouse" for use with the POSSUM and PPOSSUM systems. The risk results are selected and combined with an appropriate data set in the Hull site. This combined data set is then used to produce the scoring risk models for use in the later steps.

- **Step 2 (Clean/Transform/Filter):** Data is cleaned and transformed by using the appropriate data mining methods as detailed in the "Data Preparation Strategy" section below. Alternative models are then built. Note that the term model here means the filtering data set (from the filtering task) derived from the use of heuristic clinical decision rules. The three types of models introduced in Chapter 6 are:
    - **Clinical Models**: These are based on the decision rules of significant attributes in the data domain. For example, the clinical models *CM3a, CM3b, CM4a,* and *CM4b,* are based on significant attributes derived from clinical advice (Kuhan et al, 2001).

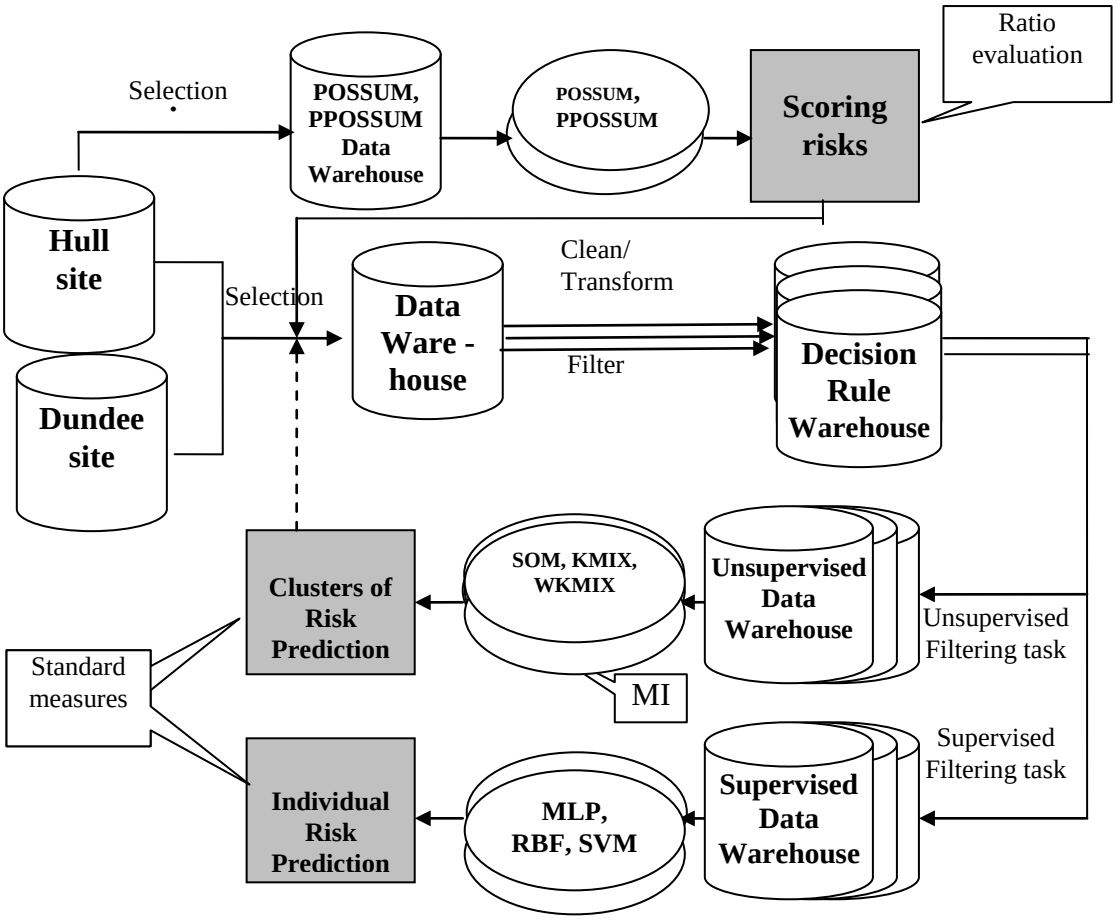

Figure 5.4: The detailed steps for the thesis experiments.

- **Scoring Risk Models**: These are built from the POSSUM and PPOSSUM systems results, to produce the *Mortality, Morbidity, or Death rate* models.
- **Clustering Models**: These are derived from the results of applying the clustering algorithms (if step 5 is required).

- **Step 3 (Data Mining Techniques):** Depending on the purpose of the classification or clustering task, this step is to choose appropriate pattern recognition and data mining techniques. They can be divided into supervised and the unsupervised methods. From this point, the appropriately formatted data set is then selected and stored in the "Unsupervised Data Warehouse" or "Supervised Data Warehouse" according to the technique chosen. The unsupervised techniques contain self

organizing maps, KMIX, and WKMIX (see detail in Chapter 4 and Chapter 8) whereas the supervised techniques include multilayer perceptron, radial basic function, and support vector machine (see detail in Chapter 4). Note that the mutual information feature selection method (described in Chapter 8) might be used here to measure the significance of data attributes to the outcome classes.

- **Step 4 (Compared/ Evaluation):** All results of the unsupervised or supervised techniques are evaluated by one or more of the standard measures as indicated above.
- **Step 5 (Building New Models):** This step might be used if the unsupervised classifiers' outcomes become the expected outcomes for new risk prediction models. A new classification process starting from step 3 is then created.

5.3.3. Data Preparation Strategy

In data mining methodologies, the preparation of data is an important task. The careful preparation of the data contributes heavily to the success in applying data mining classifiers. From the specific steps for all experiments, as indicated in Figure 5.4, a strategy for preparing data in each stage can be seen as follows:

Data Selection Stage

In this stage, the attribute set is selected by eliminating redundant and irrelevant attributes. For example, the attribute "*Theatre session date"* in the Hull site, reflecting the operation date for patients, is not relevant to any of data mining tasks here. Therefore, it can be eliminated. The "empty attributes", such as "*LOWEST_BP*", which represents the lowest blood pressure measurements during an operation, contain mostly null values. As almost of its values are null except for four entries (see Table 5.3), such attributes are best removed.

By the use of this method, the attributes in the Hull site are reduced from 98 (original) attributes to 86 "meaningful" attributes. Similarly, the original 57 attributes in the Dundee site are reduced to 36 "meaningful" attributes.

		LOWEST_BP
N	*Valid*	4
	Missing	494

Table 5.3: The frequencies of LOWEST_BP attribute in the Hull site.

Data Cleaning Stage

The most significant work in this stage is dealing with missing data values. There are many methods to deal with missing data values such as linear regression, decision tree computation, standard deviation, mean-mode method, and so on. The detail for each method except the mean-mode is not described here. Detail of these methods can be seen in Pyle (1999) and Han and Kamber (2001). The mean-mode method, for each type of attribute, can be seen as follows:

- **Numerical attributes**: Fill missing values by the mean of the “non-missing” values (Pyle, 1999).
- **Categorical attributes**: Fill missing values by the mode (Han and Kamber, 2001). This is the maximum of frequency of the “non-missing” categorical values for the attribute.
- **Boolean attributes**: The missing values here can be treated as for categorical attributes. This means missing values are filled by the mode of the attribute.

For example, Table 5.4 shows the rates of missing values in both the Hull and Dundee sites.

Attribute Name	Descriptions	Type	% Missing Values	
			Hull Site	Dundee Site
Heart disease	Any heart disease	Boolean	1/498 (0.002%)	7/341 (2.05%)
ECG	Electrocardiogram	Categorical	0	16/341 (4.69%)
Blood loss	Blood loss in operation	Continuous	8/498 (0.016%)	243/341 (71.26%)

Table 5.4: The missing values rates of some attributes for both the Hull and Dundee sites.

The missing values for the "Heart disease" attribute in both sites will be filled by the mode valued "N" (see Table 5.5). The missing values for "ECG" in the Dundee site (16 out of 341 cases - see in Table 5.4) will be replaced by the mode valued "Normal", and the missing values for the continuous attribute of "Blood loss" will be filled by the mean value of 317 (for the Hull site) and 213.17 (for the Dundee site).

Attribute Name	Mean/Mode	
	Hull Site	Dundee Site
Heart disease	N (295/498)	N(255/341)
ECG	Normal (338/498)	Normal (233/341)
Blood loss	317	213.17

Table 5.5: The mean/mode values of some attributes for both the Hull and Dundee sites.

Data Transformation Stage

This step concentrates on the transformation from various attribute types to appropriate formatted types for the use with the selected data mining techniques. Three transformation types are:

- **Numerical attributes:** Continuous attribute values are rescaled into the range of [0, 1] with a linear normalisation formula. It is given by:

New value = (original value - minimum value)/(maximum value - minimum value).

For example, Table 5.6 shows the descriptive statistics for the continuous "Age" attribute in the Hull site.

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	498	38	93	67.96	7.958

Table 5.6: The descriptive statistics of Age attribute for the Hull site.

The new values in this attribute are rescaled in to the range of [0,1] by applying the above formula as follows:

New value = (original value - 38)/55.

- **Boolean attributes:** Boolean values are transformed (from T or F, Yes or No, etc.) to values of 0 or 1.
- **Categorical attributes**: The categorical values are transformed whenever there is a request for numerical transformation for the data mining techniques. Two phases of this transformation are: Firstly, categorical values are transformed to special discrete values of "*Normal*" and "*Abnormal*" in term of the medical signal. Then, these are transformed in to numerical Boolean values of 0 or 1. Note that the value of "*Normal*" here can stand for the values of "*No", "None",* or *"Normal*" etc. in the attribute. Conversely, the value of "*Abnormal*" stands for all the other medical symptoms.

This transformation is based on the specific characteristics of the data domain, and derived from the advice of medical experts (Kuhan et al, 2003). For example, the transformation

method is applied to the categorical attribute of ECG in both Hull and Dundee site data. This attribute in the Dundee site contains 3 values of "*Normal"; "A-fib<90*"; and "*other*". However, it includes 8 separate values in the Hull site. They are: *">=5 etopics/min"; "Afib 60-90"; "A-fib <90"; "Normal"; "other"; "other abnormal rhythm"; "Qwaves";* and *"ST/T wave changes*". Obviously, the value of "*other"* has a different meaning in the Dundee and the Hull sites. Therefore, to be consistent across the data domain for both sites, the attribute *ECG* will be transformed into two discrete categorical values of "*Normal*" and "*Abnormal*". The value of "*Normal*" stands for all values as "*Normal*" in both sites; the value of "*Abnormal*" stands for the rest. These values of "*Normal*" and "*Abnormal*" are then transformed to numerical Boolean values of 0 or 1 respectively.

Additionally, there is another reason for the use of the above categorical transformation for this data domain. This transformation might help to reduce the nodes in the input layer in a neural network topology. This helps to reduce the complexity of a neural network process. For example, the clinical model CM1 (detailed structure can be seen in section C.5.1 in Appendix C), contains a set of 24 input attributes and 839 cases derived from both the Hull and Dundee sites. Assume that all the data needs to be presented as numerical data. Therefore, the categorical values have to be transformed into numerical ones. Alternatively assume that the categorical attribute is transformed into binary sub-attributes, where each sub-attribute represents an individual value in the original one. Therefore, as indicated above the ECG attribute needs 9 binary sub-attributes (3 binary sub-attributes for the Dundee site and 8 binary sub-attributes for the Hull site). Similarly, the *CAROTID_DISEASE* attribute needs 15 binary sub-attributes; the *ARRHYTHMIA* attribute needs 4 binary sub-attributes; the *CCF* attribute needs 4 binary sub-attributes; the *PATCH* attribute needs 9 binary sub-attributes; and the *RESPIRATORY* attribute needs 4 binary sub-

attributes. Obviously, CM1 now has 63 input-attributes instead of 24. Therefore, the network topology is more complicated and needs a greater sized data population to run any experiment than the one with only 24 input nodes and 839 cases.

Data Filtering Stage

In this step, the use of alternative decision rules is applied to identify expected risks for individual patients before using the classification techniques. Note that these rules are based on clinician's advice (Kuhan et al, 2001), or the result from other classification processes (e.g the POSSUM, the PPOSSUM, or unsupervised clustering algorithms).

- **Supervised Filtering task**: The data set in the "Supervised Data Warehouse" is taken from the "Decision Rule Warehouse", but with a labelled a-priori outcome, for use with supervised techniques such as multilayer perceptron, radial basis function, and support vector machine..

For example, the clinical model CM3aD in the Case Study II in Chapter 6 contained 16 inputs and 1 output attribute, and has two levels of risk derived from heuristic decision rules. These rules are based on clinician's advice on the attributes *"PATIENT STATUS"* and *"COMBINE"* as follows:

$$\Sigma(\text{PATIENT STATUS, COMBINE}) = 0 \rightarrow \textit{"Low risk"}$$

$$\Sigma(\text{PATIENT STATUS, COMBINE}) \geq 1 \rightarrow \textit{"High risk"}.$$

Therefore, the model outcome set contains 284 "Low risk*",* and 57 "High risk*"* patterns (see Table 5.7).

		Frequency	**%**	**Valid Percent**
Valid	*High risk*	57	16.7	16.7
	Low risk	284	83.3	83.3
	Total	341	100.0	100.0

Table 5.7: The frequency of outcome for the clinical model CM3aD.

- **Unsupervised Filtering task**: Similarly, the data set is taken from the "Decision Rule Warehouse" without the expected outcome attribute. This set is then stored in the "Unsupervised Data Warehouse". Appropriate unsupervised techniques such as the self organizing maps and clustering algorithms of KMIX and WKMIX are applied to give outcome labels.

For example, model CM3aD in the Case Study IV in Chapter 6, taken from "Unsupervised Data Warehouse", contains 341 cases and 16 inputs after eliminating the outcome for applying KMIX clustering algorithm (see detail in the Case Study IV in Chapter 6). The same decision rules as in the Case Study II in Chapter 6 are applied for evaluating the clustering outcome results.

5.3.4. Explanatory Case Studies

This section describes the use of the general thesis methodology for three typical thesis case study experiments. The detail of data preparation and experimental steps for each case study can be seen in Appendix C.

POSSUM and PPOSSUM Classifiers

This section describes how the Case Study I in Chapter 6 fits with the thesis methodology.

- **Step 1 (Selection)**: A selection of data set derived from the Hull site contains 3 attributes and 498 cases. The data is formatted to use with the POSSUM and PPOSSUM systems.
- **Step 2 (Clean/Transform/Filter)**: This step is ignored, because the data is already clean and ready to be used with the POSSUM/ PPOSSUM formulas.
- **Step 3 (POSSUM and PPOSSUM calculations)**: Data is used with the POSSUM and PPOSSUM formulas (Equations 2.1, 2.2, and 2.3 in Chapter 2) to calculate

Mortality, *Morbidity*, and *Death rate* risk scores. The patient risk scores are divided into various groups ranging from 0% to 100%. For example, a group of the range from 0% to 10% shows the patients, whose risk scores are in this range. The averaged risk score for each group is also calculated. The predicted numbers of *Mortality*, *Morbidity*, and *Death rate* for each group are then produced.

- **Step 4 (Compared/ Evaluation)**: The performances of the POSSUM and PPOSSUM classifiers are evaluated based on comparisons between predicted numbers of M*ortality*, or *Morbidity*, or D*eath rate* and the actual outcome in each group and overall across all groups.
- **Step 5 (Building New Models)**: Individual categorical risk is generated based on the average value of the overall risk scores. The risk category contains two levels of risk ("*High risk*" and "*Low risk*") depended on the higher or smaller of threshold value (average value of overall risk scores). For example, assume that the average of the "*Death rate*" risk scores is 25.18%; this then defines a threshold. The individual outcome risk is "*High risk"* if the individual risk score is higher than this threshold. Otherwise, the individual outcome risk is "Low *risk".* These categorical risk results are then selected and combined with an appropriate data set derived from the Hull site. This data set is then stored in the "Data Warehouse" to produce scoring risk models used in later experiment.

Supervised Classifiers

This section describes the use of the thesis methodology for model CM3aD of the Case Study II in Chapter 6.

- **Step 1 (Selection):** The data set is selected from the Dundee site with 18 attributes, and 341 cases.
- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning task**: The missing values are filled using the following methods: Numerical missing values are replaced by the mean of the "non-missing" numerical values; categorical missing values are replaced by the mode of "non-missing" categorical values; and Boolean missing values are filled by the mode of "non-missing" Boolean values.
    - **Transformation task**: The experiment requires all numerical data. Therefore, three transformation methods as indicated in the "Data Preparation Strategy" section are used.
    - **Filtering task**: The heuristic decision rules are based on two attributes of *"PATIENT STATUS"* and *"COMBINE"* as follows:

        Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$ *"Low risk"*

        Σ(PATIENT STATUS, COMBINE) $\geq$ 1 $\rightarrow$"*High risk"*
- **Step 3 (Data Mining Techniques):** The techniques used in this step are: multilayer perceptron; radial basis function; and support vector machine. The WEKA software package (WEKA, 2005) is used for this. For example, the multilayer perceptron technique, described in the Case Study II in Chapter 6, is used with a 16-0-1 topology (16 input nodes, 0 hidden nodes, and 1 output -2 class nodes). Alternative parameters are used with this topology such as changing the number of cycles, or learning rates. These changes are used to compare performance. The data is split into a training set of 90% of population, and the rest is for a test set (10%). A 10-fold cross-validation method is used.

- **Step 4 (Comparison/ Evaluation)**: Standard measurements such mean square error, confusion matrix, sensitivity and specificity rates, positive predictive value, and negative predictive value are used to evaluate the classification results.

Unsupervised Classifiers

The last example here is an analysis of how thesis methodology fits to the unsupervised experiments with two clinical models of CM3aD, and CM3bD (Case Study IV in Chapter 6).

- **Step 1 (Selection):** The data set is selected from the Dundee site with 18 attributes, and 341 cases.
- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning task**: Missing values are filled using the same method as in the previous section.
    - **Transformation task**: This task requires only numerical attributes. Hence, the continuous data is rescaled into the range of [0, 1] using the linear equation as indicated above. The other (categorical and Boolean) attributes are ignored in this step.
    - **Filtering task**: The following heuristic decision rules are applied based on the two attributes *"PATIENT STATUS"* and *"COMBINE"*. The model CM3aD has two levels of risks ("*High risk", "Low risk")* as given by:

$$\Sigma(\text{PATIENT STATUS, COMBINE}) = 0 \rightarrow \textit{"Low risk"}$$

$$\Sigma(\text{PATIENT STATUS, COMBINE}) \geq 1 \rightarrow \text{"}\textit{High risk"}$$

The model CM3bD has three levels of risks ("*High risk*", "*Medium risk"*, "*Low Risk")* as given by:

Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$ *"Low risk"*

Σ(PATIENT STATUS, COMBINE) = 1 $\rightarrow$ *"Medium risk"*

Σ(PATIENT STATUS, COMBINE) = 2 $\rightarrow$ *"High risk"*

- **Step 3 (Data Mining Techniques)**: The clustering algorithm KMIX is used in this step with both models without the expected outputs indicated above. The number of required clusters is 2 and 3 according to the model CM3aD and CM3bD respectively.
- **Step 4 (Compared/ Evaluation):** The clustering results are compared to the expected outcomes defined in step 2 by using standard measures such as *confusion matrix,* s*ensitivity*, s*pecificity* rates, and so on.
- **Step 5 (Building Clustering Models)**: The new clustering models of CM3aDC and CM3bDC are built based on the KMIX results. This means the input set is the same as in the models of CM3aD and CM3bD. However, these new models' outcomes are derived from the KMIX results. Both new models, CM3aDC and CM3bDC, are then applied again from step 3 in the thesis framework. A new process is created for a request for the use of supervised neural network techniques. The results are then measured and evaluated with standard measures.

5.4. Summary

Data mining is a particular set of steps in the framework of the "knowledge discovery from data" process. Alternative views of the definition of data mining are shown in order to produce the definition of a data mining methodology. The data mining methodology selection criteria are discussed in order to provide a general view about the thesis methodology. Additionally, the popular existing data mining methodologies of CRISP-DM and SEMMA are discussed to show the motivation for producing a thesis specific

methodology. A data mining methodology is derived from Davis (2007). The illustration of the thesis methodology shows the experimental framework for the thesis. The analysis of data from both the Hull and Dundee sites and some detailed examples of data preparation strategy is shown as an application of applying data mining methodology for this thesis. Furthermore, three typical thesis case studies demonstrate the fitting of the thesis methodology for the data domain used in this thesis. The data from both sites will be used partially or fully in different case studies and the thesis experiments in later chapters.

Chapter 6

Experimental Models and Case Studies

6.1. Introduction

This chapter defines the set of main variables selected from the data domain. These variables are significant in defining the inputs for alternative experimental models in the thesis. Note that the term model here means the filtering data set that is ready formatted for use with alternative pattern recognition and data mining techniques. Two types of experimental models are introduced in this chapter: clinical models and scoring risk models. The input set of the a-priori (clinical) models are based on common attributes of both the Hull and Dundee sites and significant attributes derived from Kuhan et al (2001). The input set of scoring risk models are based on main variables derived from Copeland et al (1991) and the POSSUM and PPOSSUM systems. The expected (and alternative) outcomes for these models are inferred from heuristic rules based on two attributes "*PATIENT STATUS*" and "*30D stroke/death*". These outcomes are divided into categorical levels of risks, such as "High risk", "Medium risk", "Low risk", and so on.

This chapter also discusses the thesis case studies. The POSSUM and PPOSSUM, and pattern recognition (supervised and unsupervised) techniques are applied with alternative models and the thesis data in these case studies.

6.2. Main Variables for Risk Prediction Models

According to Kuhan et al (2001), the significant cardiovascular factors identified by logistic regression analysis can be seen in Table 6.1. It is clear that the results in Table 6.1 show the good relationship between three risk factors to the cardiovascular outcomes as all the regression coefficients (parameter estimate) are positive and the significant levels (P values) are less than 0.05 (95 % of confidence interval). For example, the regression coefficient of "Heart disease" factor of 0.992 and its standard deviation error of 0.402 with the significant level of 98.6% (P value of 0.014) means this factor strongly influences to the outcome for cardiovascular models. These attributes (Heart disease, Diabetes, and Stroke) are the main factors for the outcome calculations for the thesis models.

Risk Factor	Parameter Estimate	Standard error	P
Heart disease	0.992	0.402	0.014
Diabetes	0.996	0.450	0.027
Stroke	0.827	0.394	0.036

Table 6.1: Significant risk factors in cardiovascular models (Kuhan et al, 2001).

The significant variables taken from the given data from both the Hull and Dundee sites can be seen in Table 6.2. They are selected using the advice of a clinician expert in the cardiovascular area (Kuhan et al, 2001). Some of these variables are in the 12 physiological and 6 operative factors described in Chapter 2 (for example, "*Age", "ECG", "Blood loss",* and *"Duration*"). Additionally, some of the variables in Table 6.2 are signal symptoms in the POSSUM and PPOSSUM systems. For example, the variable "*RESP-problem*" (*Respiratory*) here is derived from "*Respiratory signs*" in the POSSUM system; and the variable "*Hypertension*" is derived from "*Systolic blood pressure*".

It is clear from Table 6.2 that overall the number of missing values for the Dundee site is greater than for the Hull site. For example, the rate of missing values of "*Blood loss*"

attribute in the Hull site is negligible (8 per 498) whereas the rate of missing values in this attribute in the Dundee site is over two-thirds (243 per 341 - 71.26%).

Attribute Name	Descriptions	Type	% Missing Values	
			Hull Site	Dundee Site
Age	Patient Age	Continuous	0	0
Sex	Patient gender	Boolean	0	0
Heart disease	Any heart disease	Boolean	1/498 (0.002%)	7/341 (2.05%)
Diabetes	Diabetes disease	Boolean	1/498 (0.002%)	1/341 (0.29%)
Stroke	Signal of stroke	Boolean	1/498 (0.002%)	0
Side	Side of operation	Boolean	0	0
RESP-problems	Respiratory disease	Categorical-discrete	0	14/341 (4.11%)
Renal failure	Renal failure	Boolean	0	6/341 (1.76%)
ASA grade	Grade of Acetyl Salicylic Acid	Discrete	0	37/341 (10.85%)
Hypertension	Hypertension symptom	Boolean	0	6/341 (1.76%)
ECG	Electrocardiogram	Categorical	0	16/341 (4.69%)
Duration	Duration of operation	Continuous	0	0
Blood loss	Blood loss in operation	Continuous	8/498 (0.016%)	243/341 (71.26%)
Shunt	Shunt	Boolean	0	5/341 (1.47%)
Patch	Patch	Categorical	0	10/341 (2.93%)
CABG	Coronary Artery Bypass Graft surgery	Boolean	0	8/341 (2.35%)

Table 6.2: Statistical analysis of main variables.

6.3. Clinical Risk Prediction Models

This section presents alternative risk prediction models for the experiments in this thesis. CM1 is the base model. Model CM2 is derived from the model CM1 (same input attributes) with a different outcome. The models CM3a, CM3b and CM4a, CM4b are derived from the models of CM1, and CM2 with a smaller set of attributes and alternative outcomes.

6.3.1. Clinical Model 1 (CM1)

The data structure here includes 25 attributes (24 inputs and 1 attributes used for the outcome calculation) and 839 cases. Its inputs are common attributes derived from the Hull and Dundee sites. The data is a combination of both these sites (498 cases of the Hull site and 341 cases of the Dundee site). The data structure and its summary can be seen in Table 6.3.

Note that the column of “Missing values” shows the number of missing values for each attributes. The column of “Max Freq/Mean” shows the maximum frequency values of either categorical or Boolean attributes, or the mean value of numerical attributes respectively. These values will be used as the replacements for missing values as indicated the “Data Preparation Strategy” section in Chapter 5. For example, the attribute “ANGINA” has 12 missing values. The mode value of this attribute is “N” (570 from 839). Hence, missing values will be replaced by the value of “N”. The number of missing values in the numerical attribute of “DURATION” (72) is filled by the mean of “non-missing” values (1.57). Note that attribute “COMP_GROUP” in Table 6.3 contains a large number of missing values (605/839). Therefore, this attribute is eliminated.

The expected risks of this model are calculated based on "*PATIENT_STATUS*" as follows:

IF PATIENT_STATUS = "Dead" ➔ *"High risk"*

Otherwise, ➔ *"Low risk".*

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
AGE	Continuous	0	[38,93]	67.99
ANGINA	Boolean	12	Y/N/Null	570 (N)
ARRHYTHMIA	Categorical	8	none/A-Fib<90/min/A-Fib <90/Null/Other	792 (None)
ASPIRIN	Boolean	166	Y/N/Null	648 (Y)
ASA_GRADE	Continuous	38	[1,4]	2.24
BLOOD_LOSS	Continuous	252	[0,2000]	300.45
CABG_PLASTY	Boolean	9	Y/N/Null	778 (N)
CAROTID_DISEASE	Categorical	2	N/A	303 (TIA)
CCF	Categorical	9	<1/12/ >1/12/None/Null/Yes	803 (None)
COMP_GROUP	Categorical	605	N/A (removed)	N/A
D	Boolean	1	Y/N/Null	748 (N)
DURATION	Continuous	72	[0.7-5]	1.57
ECG	Categorical	33	Normal/Null/other abnormal/Q wave/ST/A-Fib<<90/and so on	571 (Normal)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449 (N)
Smoking	Boolean	0	Y/N	787 (Y)
PATCH	Categorical	253	PTFE/Dacron/Vein/Other Vein/Stent	171 (PTFE- 170/341-Dundee site); 185 (Dacron -185/499 - Hull site)
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/Mild COAD/Mod COAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
WARFARIN	Boolean	5	Y/N/Null	809 (N)
R1-A SIDE	Boolean	0	Left/Right	458 (left)

Table 6.3: The CM1 and CM2 data structure and summary.

6.3.2. Clinical Model 2 (CM2)

This model has the same input attribute set as the model CM1 (as given in Table 6.3). However, the expected risk is calculated from the two attributes of "*PATIENT_STATUS"* and "*30D stroke/death"*. Assume that the attribute "*PATIENT_STATUS"* is assigned as "Attr1", and the attribute of "*30D stroke/death"* is assigned as "Attr2". The expected risks are given by:

IF Attr1 = "Dead" Or Attr2 = "Y" ➔ *"High risk"*
Otherwise, ➔ *"Low risk"*

6.3.3. Clinical Model 3a (CM3a)

This model includes 18 attributes; containing 16 attributes used for inputs and the rest used for the expected outcome calculation. The number of model cases is 839 derived from the Hull and Dundee sites. The data structure and its summary can be seen in Table 6.4. The expected outcome is based on the attributes of "*PATIENT_STATUS"* (Attr1), and "*30D stroke/death"* (Attr2); and is given by:

IF Attr1 = "Dead" Or Attr2 = "Y" ➔ *"High risk"*
Otherwise, ➔ *"Low risk"*

6.3.4. Clinical Model 3b (CM3b)

This model has the same structure as in model CM3a (see in Table 6.4). However, the expected outcomes contain alternative categorical values labelled as "V*ery High risk*"; "*High risk"*, "*Medium risk*", and "*Low risk*". These values are based on attributes *"PATIENT STATUS" (*Attr1), and "*30D stroke/death"* (Attr2).

IF Attr1 = "Dead" AND Attr2 = "Y" ➔ *"Very High risk"*
Else IF Attr1 = "Dead" ➔ *"High risk"*
Else IF Attr2 = "Y" ➔ *"Medium risk"*
Other wise, ➔ *"Low risk"*

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
30D stroke/death	Boolean	0	Y/N	806 (N)
AGE	Continuous	0	[38,93]	67.99
ASA_GRADE	Continuous	38	[1,4]	2.24/0.46
BLOOD_LOSS	Continuous	252	[0,2000]	300.45
CABG_PLASTY	Boolean	9	Y/N/Null	778(N)
D	Boolean	1	Y/N/Null	748(N)
DURATION	Continuous	72	[0.7-5]	1.57
ECG	Categorical	33	Normal/Null/other abnormal/Q-wave/ ST/A-Fib<<90/; so on	571 (Normal)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449(N)
PATCH	Categorical	253	PTFE/Dacron/Vein/ OtherVein /Stent	171/341 PTFE-Dundee; 185/499-Dacron -Hull
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/MildCOAD/ ModCOAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
R1-A SIDE	Boolean	0	Left/Right	458 (left)

Table 6.4: CM3a and CM3b data structure and summary.

6.3.5. Clinical Model 4a (CM4a)

This model includes 16 attributes (14 attributes used for inputs and 2 attributes used for the expected outcome calculation) and 839 patient records derived from the Hull and Dundee sites. The data structure and its summary can be seen in Table 6.5.

The expected outcomes are the same as model CM3a outcomes. They are given by:

IF Attr1 =”Dead” Or Attr2= “Y” ➔ “High risk”

Otherwise, ➔ *"Low risk"*

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean/Stdev
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
30D stroke/death	Boolean	0	Y/N	806 (N)
AGE	Continuous	0	[38,93]	67.99
ASA_GRADE	Continuous	38	[1,4]	2.24/0.46
D	Boolean	1	Y/N/Null	748(N)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449(N)
PATCH	Categorical	253	PTFE/Dacron/Vein/Other Vein/Stent	171/341-PTFE -Dundee; 185/499 -Dacron - Hull site
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/Mild COAD/Mod COAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
R1-A SIDE	Boolean	0	Left/Right	
CONS	Categorical	0	1;2;3;4;5	383 (4)
Vascular Unit	Categorical	0	1;2	498 (2)

Table 6.5: CM4a and CM4b data structure and summary.

6.3.6. Clinical Model 4b (CM4b)

The data structure for this model is the same as in the model CM4a (see Table 6.5). However, the expected outcomes are the same as model CM3b outcomes.

They are given by:

IF Attr1 = "Dead" AND Attr2 = "Y" ➔ *"Very High risk"*
Else IF Attr1 = "Dead" ➔ *"High risk"*
Else IF Attr2 = "Y" ➔ *"Medium risk"*
Other wise, ➔ *"Low risk"*

6.4. Scoring Risk Models

The data for these models is selected from the Hull site and the POSSUM and PPOSSUM classified results. Three scoring risk models are introduced as: Mortality, Morbidity, and Death rate. These models share the same structure (in Table 6.6) with 498 and 22 input attributes.

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PhysiolScore	Continuous	0	[12,41]	20.37
OpSevScore	Continuous	0	[13,23]	14.29
AGE	Continuous	0	[38,93]	67.99
RESPIRATORY	Categorical	1	Normal/MildCOAD/ModCOAD/ Severe COAD/Null	431(Normal)
WARFARIN	Boolean	2	Y/N/Null	474(N)
RESP_SYSTEM	Categorical	1	Limiting SOB/No SOB/ Null/SOB at rest/ SOB in exertion	468 (No SOB)
BP	Continuous	21	[90,220]	151.9
PULSE	Continuous	23	[42,110]	74
JVP	Boolean	2	N	495(N)
WCC	Continuous	10	[4, 24.3]	7.67
HAEMOGLOBIN(Hb)	Continuous	10	[7.7,18.2]	13.9
UREA	Continuous	8	[2.1, 17.2]	6.34
SODIUM(Na)	Continuous	11	[122, 146]	138.5
POTASSIUM(Ka)	Continuous	9	[3, 5.6]	4.3
ECG	Categorical	16	≥5 ectopics/min; Afib 60-90; Normal; Null; Other abnormal; Q waves; ST/T Wave change	338 (Normal)
GCS(Coma Score)	Continuous	1	[15]	15
URGENCY	Categorical	1	Elective; Scheduled urgent	497(Elective)
BLOOD_LOSS	Continuous	8	[100, 1800]	318
NO_PROCS	Discrete number	59	[1; 2; 3]	418 (1)
OP_SEVERITY	Categorical	0	Major Plus	497 (Major Plus)
MALIGNANCY	Categorical	0	None	497 (None)
PERI_SOILING	Categorical	0	None	497 (None)

Table 6.6: Scoring risk models' input structure and summary.

There is a special pattern in this model with almost null values except the scoring values. This case will be eliminated in other models. The expected outcomes for these models can be seen in Table 6.7. These outcomes are calculated upon the comparison between the appropriate POSSUM and PPOSSUM outcome values (Mortality, Morbidity, and Death rate) and their threshold values. Note that the threshold value here means the average value (mean) of overall risk scores in the appropriate POSSUM and PPOSSUM outcomes.

Models	Outcome
Mortality model	The risk is based on the mean value of "Mortality outcome values" (MortV) *IF MortV >= mean* ➔ *"High risk"* *Otherwise,* ➔ *"Low risk"*
Morbidity model	The risk is based on the mean value of "Morbidity outcome values" (MorbV). *IF MorbV >= mean* ➔ *"High risk"* *Otherwise,* ➔ *"Low risk*
Death Rate model	The risk is based on the mean value of "Death rate outcome values" (DR). *IF DR >= mean* ➔ *"High risk"* *Otherwise,* ➔ *"Low risk*

Table 6.7: Outcome calculations for the scoring risk models.

6.5. Thesis Case Studies

6.5.1. Case Study I

This section discusses the use of POSSUM and PPOSSUM with cardiovascular data derived from the Hull clinical site. The mortality and morbidity risks are calculated for each patient. The individual risk predictions are grouped into alternative bands from 0-100%.

The ratios between predicted deaths and actual deaths for each group and overall are discussed. The aim of this study is to analyse and discuss the POSSUM and the PPOSSUM classifications; the detailed steps for this experiment can be seen in section C.1 in Appendix C.

Data

The experimental data set contains 3 attributes and 498 cases. Note that two of the three attributes are the clinician generated physiological score (PS) and the operative severity score (OS). A statistical analysis in terms of these scores can be seen in Table 6.8. The minimum scores of the PS and the OS are nearly the same (12). The maximum of the PS is 41 while the maximum OS score is approximately half at 23.

		PS	**OS**
N	*Valid*	498	498
	Missing	0	0
Mean		20.36	14.30
Std. Error of Mean		0.247	0.064
Std. Deviation		5.507	1.421
Minimum		12	13
Maximum		41	23

Table 6.8: Statistical analysis of the PS and the OS scores.

Method

Mortality, and Morbidity rates are calculated with the equations in Chapter 2 (equations (2.1), (2.2) for the POSSUM, and (2.3) for the PPOSSUM). The linear analysis method for predicted deaths is used here. The patients were divided into alternative groups according to their predicted mortality rates as bands: 0-10%, 10-20%, 20-30%, 30-40%, 40-50%, and greater than 50%. The mean of predicted mortality risk represents the average risk for

patients in each range. For example, the average mortality risk for patients in the first group (band of 0-10%) is 6.75% (see in Table 6.9). The number of operations is the number of patients in each group. Predicted death (E) is the number of dead patients as predicted by POSSUM. The Reported death (O) is the number of actual dead patients in each group. The performance of the system is measured by the ratio of observed to predicted mortality (O/E).

Results and Discussions

Table 6.9 shows mortality results of the POSSUM scoring risk system in 7 patient groups. The number of predicted and observed mortality for each group is also shown. The prediction performance is indicated by the ratios between the observed and predicted deaths in each group as well as for the patients in overall.

Range of predicted rate	Mean predicted risk of Mortality (%)	No of operations	Predicted death	Reported death	Ratio
0-10%	6.75%	274	18	29	1.61
10-20%	14.85%	148	22	28	1.27
20-30%	24.97%	44	11	8	0.73
30-40%	34.90%	11	4	3	0.75
40-50%	43.10%	16	7	7	1.00
>50%	60.85%	5	3	3	1.00
0-100%	13%	498	65	78	1.20

Table 6.9: Comparison of observed and predicted death of the POSSUM logistic equations.

The risk groups of 20-30%; and 30-40% achieved ratios of 0.73 and 0.75 respectively. This means the POSSUM system over-predicts the mortality risk for patients in these bands. Two other risk groups (40-50%; and >50%) have the same ratio being equal at exactly 1. This means the POSSUM system predict the number of mortality risk exactly the same as

the actual death of patients. Overall, POSSUM achieved results better than a ratio of 1.2 (band of 0-100%) except for the first and the second groups (bands of 0-10% and 10-20%) with the ratios of 1.61 and 1.27 respectively.

The performances of the PPOSSUM system for predicting mortality rates can be seen in Table 6.10. There is only one risk group which has the same rate of predicted deaths and reported deaths (the ratio of 1). This means the PPOSSUM predicted exactly as reported from the actual data for this group. The overall ratio is 3.12, meaning that the PPOSSUM under-predicts deaths for patients. This means the PPOSSUM performance is worse than expected.

Range of predicted rate	Mean predicted risk of Mortality (%)	No of operations	Predicted deaths	Reported deaths	The ratio
0-10%	3.00%	438	13	60	4.62
10-20%	13.48%	39	5	9	1.80
20-30%	23.25%	12	3	3	1.00
30-40%	32.27%	5	2	4	2.00
40-50%	44.86%	3	1	2	2.00
>50%	58.37%	1	1	-	0.00
0-100%	5%	498	25	78	3.12

Table 6.10: Comparison of observed and predicted death from PPOSSUM logistic equations.

The comparison between Table 6.9 and Table 6.10 shows that the POSSUM performance is better than PPOSSUM, in general. Overall, both POSSUM and PPOSSUM underestimate the risk for patients (*O/E* ratios are 1.20 and 3.12 respectively). Both systems estimate the O/E ratios according to the linear analysis method. It seems that the POSSUM uses an appropriate analysis method, because its ratio is quite close to 1.0. However, this linear analysis method seems to be inappropriate for the PPOSSUM results, because the ratio of 3.12 reflects the number of predicted deaths is too far short to the actual deaths. This result

is against the discussion of Wijesinghe et al (1998) where the linear analysis method is found to be appropriate for the PPOSSUM results. Therefore, there is an inconsistency between the linear and exponential analysis methods for POSSUM and PPOSSUM. The ambiguous use of linear and exponential analysis methods are also shown in (Yii and Ng, 2002). As discussed in Chapter 2, POSSUM and PPOSSUM also have some disadvantages. For example, these systems might have ambiguous interpretations for the categorical risks in the risk scale such as "*High*", "*Medium*", or "*Low"*. From this point, a system to improve on the above disadvantages is needed. Pattern recognition and data mining classifiers might provide suitable candidates capable of producing better results.

6.5.2. Case Study II

This section demonstrates the use of the alternative neural network techniques with the thesis data. Alternative network topologies and parameters are applied to discuss their relative performances. The detailed steps of data preparation and process explanations can be seen in section C.2 in Appendix C

Data

Two clinical models, CM3aD and scoring risk model Hull_POSS, are used. Note that experimental data is prepared by supplying missing values and transforming into appropriate numerical values using the methods explained in Chapter 5.

Model CM3aD

The CM3aD data is taken from the Dundee site with a selection of 16 input attributes and 341 patients. The two expected outcome levels are calculated in the following heuristic

formula. This calculation is based on two attributes of *"PATIENT STATUS"* and *"COMBINE"* derived from (*Heart Disease, Diabetes, and Stroke).*

$$\Sigma(\text{PATIENT STATUS, COMBINE}) = 0 \rightarrow \textit{"Low risk"}$$

$$\Sigma(\text{PATIENT STATUS, COMBINE}) \geq 1 \rightarrow \text{"}\textit{High risk}\textit{"}$$

Therefore, the model CM3aD contains 284 "Low risk*",* and 57 "High risk*"* patterns.

Model Hull_POSS

The data is taken from the Hull site with a selection of 22 input attributes based on Copeland et al (1991) and 497 cases. Basically, these attributes are the main factors for the POSSUM and PPOSSUM systems. Through analysing the data, 6 "empty" attributes are eliminated by applying data mining methodology for preparing experimental data (see detail in "Data Preparation Strategy" section in Chapter 5). For example, the attribute "PERI_SOILING" is eliminated, because it contains 497/497 values of "None".

The expected outcome is calculated based on the attribute "PATIENT_STATUS" as follows:

IF PATIENT_STATUS = "Dead" ➔ *"High risk"*

Otherwise, ➔ *"Low risk"*

Therefore, data set includes 16 inputs, and the expected outcome set contains 78 values of "High risk" and 419 values of "Low risk".

Method

Alternative neural network techniques are applied using the WEKA software package (WEKA, 2005). A method of 10 cross-validation folds is used. This means the data is divided into 10 partitions. A random partition (10% of population) is selected as a test set whereas the rest (90%) is used for the training set; and this process is repeated 10 times. This is to avoid the over-fitting problem in the training process.

Alternative topologies are used with alternative parameters such as number of hidden nodes; number of epochs; and so on. The outcome results are represented in a confusion

matrix with standard measurements of sensitivity, specificity, positive predictive value, negative predictive value, and mean square error.

Results and Discussions

Table 6.11 and Table 6.12 show the results for the two models CM3aD and Hull_POSS. Two models CM3aD and Hull_POSS are applied with alternative parameters such as the epochs of 100 or 500, and the learn rates of 0.01 or 0.3 (for multilayer perceptron - see Table 6.11); or the number of center c of 1 or 2 (for radial basic function - Tables 6.11 and 6.12); or alternative types of kernel functions applied in support vector machines (polynomial or radial basic function - Table 6.11). The representative parameters in the experimental models in the thesis are chosen based on the lowest mean square errors (MSE) for the models.

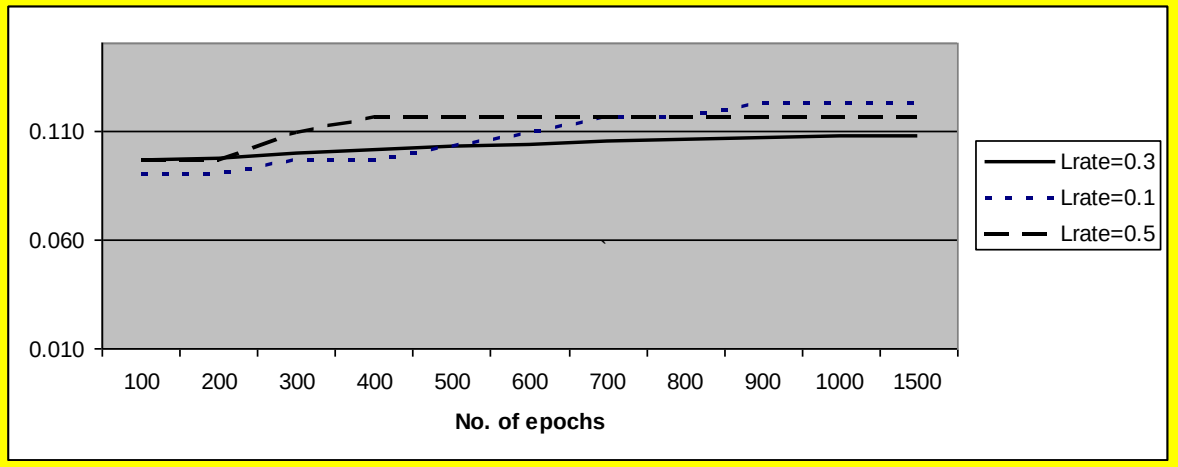

Figure 6.1: Alternative epochs and learn rates applied to CM3aD model.

For example, Figure 6.1 show alternative epochs and learn rates applied to the CM3aD multilayer perceptron model with the topology of 16 inputs; 0 hidden node; 2 nodes for the output. It is clear that the MSE of each experiment is a negligible change (lowest MSE is 0.09 and highest MSE is 0.12). However, according to Haykin (1999), the lower learn rate helps in smoother learning, and reducing epochs can help to reduce the over-fitting in

leaning process. Moreover, in general the model produced quite stable MSE (about 0.09 on average - see Figure 6.1) with the learn rate of 0.3 with alternative epochs. Therefore, the representative epoch (smallest) and learn rate can be used in this model as 100, and 0.3 respectively.

Topologies and Parameters	**Risk**	**Confusion Matrix**		**ACC**	**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
		High risk	*Low risk*						
MLP_TP1 (2H; η=0.3; 500 epochs)	*High risk*	27	30	0.88	0.47	0.96	0.73	0.90	0.09
	Low risk	10	274						
MLP_TP2 (0H; η=0.01;100 epochs)	*High risk*	0	57	0.83	0.00	1.00	N/A	0.83	0.11
	Low risk	0	284						
MLP_TP3 (0H; η=0.3; 100 epochs)	*High risk*	28	29	0.90	0.49	0.98	0.85	0.91	0.09
	Low risk	5	279						
MLP_TP4 (0H; η=0.3; 500 epochs)	*High risk*	27	30	0.90	0.47	0.98	0.84	0.90	0.09
	Low risk	5	279						
RBF_TP6 (c=1)	*High risk*	27	30	0.85	0.47	0.93	0.56	0.90	0.1
	Low risk	21	263						
SVM_TP8 (poly kernel, w=1, p=1)	*High risk*	17	40	0.88	0.30	1.00	0.94	0.88	0.11
	Low risk	1	283						
SVM_TP9 (poly kernel, w=2, p=2)	*High risk*	27	30	0.89	0.47	0.98	0.82	0.90	0.1
	Low risk	6	278						
SVM_TP10 (rad-kernel w=1;δ=0.01)	*High risk*	0	57	0.83	0.00	1.00	N/A	0.83	0.16
	Low risk	0	284						

Table 6.11: Alternative topologies and techniques for the CM3aD model.

The assigned symbols in both Tables 6.11 and 6.12 for alternative neural network parameters are as follows:

- Multilayer perceptron: The parameter set contains number of hidden nodes; learning rate (η); and number of training epochs.
- Radian basis function: The parameter is number of centres (c).

- Support vector machine: The parameter contains the type of kernel functions; margin *w;* and exponent p (if applicable).

For example, the classifier *MLP_TP1 (2H; η =0.3; 500 epochs)* in Table 6.11 means the multilayer perceptron classifier used with a hidden layer of 2 nodes, the learning rate (η) of 0.3 and 500 of training epochs.

Classifier	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV	MSE
		High risk	*Low risk*						
Hull_POSS_TP1 (MLP_2H_0.3_500)	*High risk*	9	69	0.82	0.12	0.96	0.33	0.85	0.14
	Low risk	18	401						
Hull_POSS_TP2 (MLP_0H_0.3_500)	*High risk*	6	72	0.84	0.08	0.98	0.46	0.85	0.14
	Low risk	7	412						
Hull_POSS_TP3 (RBF_c_2)	*High risk*	1	77	0.84	0.01	0.99	0.20	0.84	0.13
	Low risk	4	415						
Hull_POSS_TP4 (SVM_Poly_p_2)	*High risk*	0	78	0.84	0.00	1.00	0.00	0.84	0.16
	Low risk	2	417						

Table 6.12: Hull_POSS model results with alternative techniques and parameters.

Overall, all classifiers from Tables 6.11 and 6.12 achieve small mean square error rates (average of about 0.12) for both CM3aD and Hull_POSS models. In Table 6.11, the classifier MLP_*TP2 (none hidden node; η=0.01; 100 epochs)* predicted all expected "High risk" patterns as "Low risk", despite this topology seemingly appropriate for the data domain population of 341 cases (about 10 cases per weight per class).

This poorest result might be due to inappropriate network parameters (e.g the learning rate or number of training epochs), as can be seen from a later experiment (classifier MLP_TP3), where the learning rate (η) is increased to 0.3. Consequently, the sensitivity rate achieved from this topology is 0.49. From this point, the classifier MLP_*TP2 (0H;*

η=0.01; 100 epochs) had the worst sensitivity rate (0.00), because the learning rate (η) is inappropriate (too small). The model MLP_TP3 (0H; 0.3; 100 epochs) and the model MLP_TP4 (0H; 0.3; 500 epochs) have the same topology and learn rate, and resulting MSE (0.09). However, the model MLP_TP3 shows faster convergence than MLP_TP4 (100 and 500 epochs respectively). Moreover, the MLP_TP3 predictive rates such as sensitivity, positive predictive value and negative predictive value are little higher than these in MLP_TP4 (0.49, 0.85, 0.91 and 0.47, 0.84, 0.90 respectively). Therefore, the parameters chosen in the MLP_TP3 model can be seen as more appropriate for the multilayer perceptron model with the given data.

From Table 6.12, all classifiers achieve the same accuracy rate (0.84) except the classifier Hull_POSS_TP1 (0.82). The sensitivity rates as well as the positive predictive values are very poor (less than 0.12 and 0.46 respectively). Furthermore, there are big difference between the sensitivity rates and the positive predictive values except for the very poor classifier Hull_POSS_TP4. For example, in the classifier Hull_POSS_TP1, its sensitivity is 0.12 whereas its positive predictive value is 0.33. This sensitivity shows that 9 "High risk" patterns are predicted correctly in "High risk" class (horizontal comparison in the confusion matrix) whereas 69 "High risk" patterns are mis-predicted as "Low risk" class. These mis-predictions might be explained by the natural structure of the data, with multiple forms of patterns in "High risk" class, some of which are similar to some patterns in the "Low risk" class. In a medical context, the mis-classification of patients can be explained that their medical symptoms are similar to "Low risk" patients although they are the expected "High risk" patients (as indicated in the heuristic rule above in this section). By contrast (vertical comparison), 18 "Low risk" patterns are mis-predicted into the "High risk" class possibly because their data forms are similar to the "High risk" patterns. Therefore, the poor "High

risk" predictions might imply confusion in the data patterns or the nature of problem or the difficulty of measuring the influential parameters. Further investigation, in particular into the "High risk" borderline patterns, is necessary.

Tables 6.11 and 6.12 also show that the multilayer perceptron seems to be the "most appropriate network" for risk predictions applied with both CM3aD and Hull_POSS models in general. Furthermore, these results indicate that radial basis functions show the "poorest results". These results might be explained by the disadvantages of radial basis function networks as indicated in Chapter 4.

The appropriate choice of neural network topology and its parameters help to produce the appropriate results and avoid over-parameterisation. According to Haykin (1999), the smallest number of hidden neuron is chosen so that it can produce a performance better or close to a Bayesian classifier's results. The chosen appropriate number of hidden nodes can be seen in the comparison between the use of neural network and Bayes classifier for models CM3aD and Hull_POSS (see in Table 6.13).

Classifiers	**Risk**	**Confusion matrix**		**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
		High risk	*Low risk*					
CM3aD_MLP(0H;0.3; 100 epochs)	*High risk*	28	29	0.49	0.98	0.85	0.91	0.09
	Low risk	5	279					
CM3aD_Bayes	*High risk*	28	29	0.49	0.94	0.62	0.90	0.09
	Low risk	17	267					
Hull_POSS(MLP_0H_0.3_500)	*High risk*	6	72	0.08	0.98	0.46	0.85	0.14
	Low risk	7	412					
Hull_POSS_Bayes	*High risk*	8	70	0.10	0.94	0.25	0.85	0.14
	Low risk	24	395					

Table 6.13: The comparison between multilayer perceptron and Bayes classifiers.

The results in Table 6.13 show that almost ratios of the mean square error, sensitivity, specificity, and negative predictive values are similar in both classifiers of neural network

and Bayes. Only the positive predictive values of the multilayer perceptron are clearly better than that of the Bayes classifier.

The chosen neural topology and number of hidden nodes depend on the data domain size. This thesis uses a heuristic calculation of 10 cases per class per network weight to determine the adequate number of hidden layers and hidden nodes. This is to avoid over-parameterisation problems, in which the data will not support the use of too many layers and nodes. Therefore, for example, the multilayer perceptron with the topology of 16-0-1 (16 input nodes; 0 hidden nodes; and 1 output -2 class nodes) is suitable for the CM3aD data set with 341 patterns; i.e. 16 weights with 10.66 examples per class per weight. However, it means there are no hidden layer in the network, as the data can only support for the network output class (2 classes) and its weights (16). The requested examples for the network are about 341.12 cases (16*10.66*2).

According to Haykin (1999), the learning rate is chosen to achieve a realistic training period, and the number of epochs is kept as small as possible. For example, Table 6.11 results show that the best performance for model CM3aD is from the classifier MLP_TP3 (0H; 0.3; and 100 epochs). This is marginally better than the more complex network MLP_TP1 (2H; η=0.3; 500 epochs) which makes use of a hidden layer, with 2 nodes.

The cross-validation method used in the neural network techniques can help to avoid the over-fitting of topology, parameters and so on in the training process. In this thesis, the folding cross-validation method is used. The final mean square errors are achieved over the average of all the mean square error for each fold.

Alternative number of cross-validation folds (k) is investigated to find the appropriate number of folds, which can then be used for all thesis experiments. The two best multilayer

perceptron topologies from Table 6.11 and Table 6.12 are used for two models CM3aD and Hull_POSS. The results can be seen in Table 6.14.

It is clear that the mean square error rates are the same with three choices of *k* (*k*=5, *k*=10, and *k*=15). However, the other evaluations (sensitivity; specificity; positive predictive value; and negative predictive value) show that the choice of k of 10 achieves the better results for model CM3aD; and not much difference for model Hull_POSS. Therefore, from now on the number of cross-validation folds used in all thesis experiments is 10.

Classifiers	**Risk**	**Confusion Matrix**		**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
		High risk	*Low risk*					
CM3aD k=10	*High risk*	28	29	0.49	0.98	0.85	0.91	0.09
	Low risk	5	279					
CM3aD k=15	*High risk*	27	30	0.47	0.96	0.71	0.90	0.09
	Low risk	11	273					
CM3aD k=5	*High risk*	26	31	0.46	0.95	0.63	0.90	0.09
	Low risk	15	269					
Hull_POSS k=10	*High risk*	6	72	0.08	0.98	0.46	0.85	0.14
	Low risk	7	412					
Hull_POSS k=15	*High risk*	9	69	0.12	0.97	0.45	0.86	0.14
	Low risk	11	408					
Hull_POSS k=5	*High risk*	9	69	0.12	0.96	0.38	0.85	0.14
	Low risk	15	404					

Table 6.14: Alternative number of cross-validation experiments.

As discussed above, the correct prediction of "High risk" patterns is of concern in the thesis. All "High risk" prediction results (from Tables 6.11 and 6.12) show a big gap between the sensitivity rates and the positive predictive values. These distances either mirror the poor performance of classifiers or the internal structure of the data, which the

classifier could not recognise. Therefore, a more in depth investigation of the structure of data is necessary. Unsupervised pattern recognition and data mining techniques might provide the abilities for this.

6.5.3. Case Study III

This section demonstrates the use of a self organizing map for the data derived from the Dundee site (model CM3bD). The use of self organizing map here is to show the abilities of unsupervised pattern recognition techniques when applied into thesis data domain. The detailed steps of data preparation and experimental process can be seen in section C.3 in Appendix C.

Data

The data is taken from the Dundee site with a selection of 16 input attributes and 341 patients. The expected risks are calculated based on two attributes, "PATIENT STATUS" and "COMBINE", in the following rules:

If Σ(PATIENT STATUS, COMBINE) = 0 ➔ *"Low risk"*
If Σ(PATIENT STATUS, COMBINE) = 1 ➔"*Medium risk"*
If Σ(PATIENT STATUS, COMBINE) = 2 ➔*"High risk"*

Hence, the CM3bD model contains 48 values of "*High risk*"; 73 values of "*Medium risk*"; and 220 values of "*Low risk"*. Note that these risks are only used for the visualization in the final map, they are not involved into the clustering process.

Method

A self organizing map tool is used with the SOM Toolbox (SOM toolbox, 2000) of the Matlab software package (Mathworks, 1994). A map is created of size [30, 16]. Note that the map's size is calculated based on a heuristic formula derived from Alhoniemi et al (2005). The self organizing map produces the U-matrix to visualise the data on the map.

Alternative component planes are also represented as individual maps for each attribute. A clustering map is shown via applying this model data with a clustering algorithm. The expected risks are also shown in this map.

Results and Discussions

The final map has an average quantization error of 0.43, and topographic error of 0.00. Therefore, the accuracy of the map calculated according to Equation (4.12) in Chapter 4 is about 0.7 (70%). The visualization results for the U-matrix can be seen in Figure 6.2.

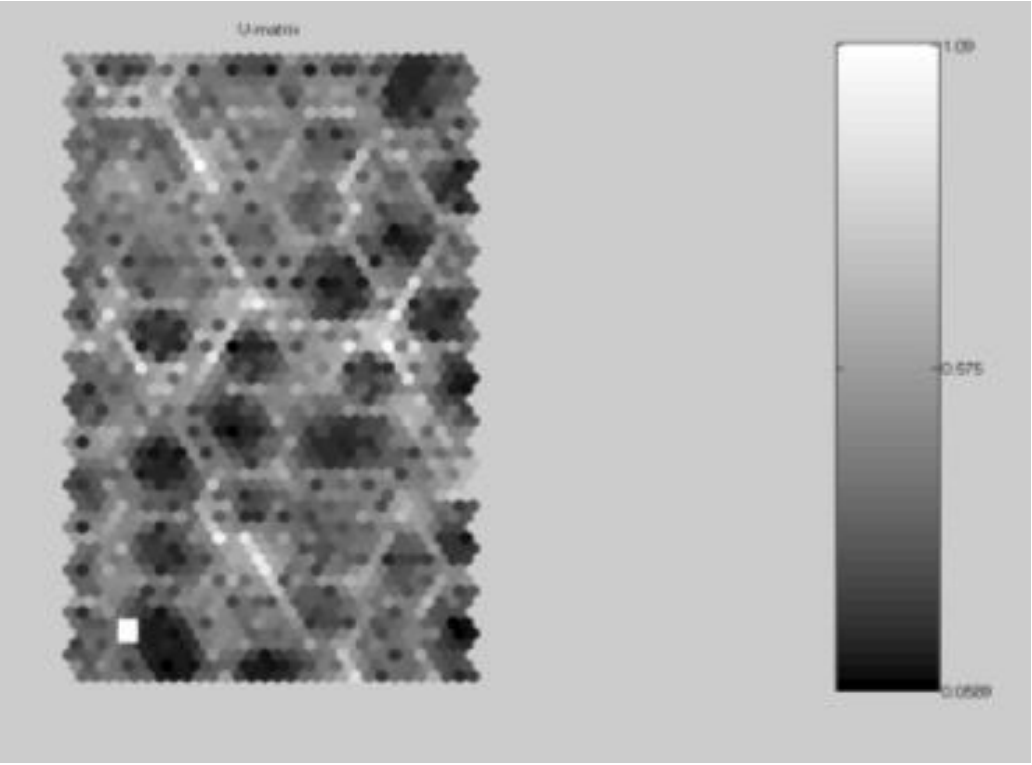

Figure 6.2: The final U-matrix of data set for model CM3bD.

From Figure 6.2, the colour scale shows that the distance between clusters is very small in most areas in the map, because they are almost the same colour (middle dark in the black and white scale). Hence, we can suggest that these input patterns might have similar pattern forms.

When the correlation between input attributes is of interest, it is convenient to look at the "component planes" where each plane presents an attribute in data domain. Figure 6.3 shows the component plane visualisations for the map in Figure 6.2.

In Figure 6.3, the component plane of "PATCH" shows the data for the "PATCH" attribute. It is clear that the high values are allocated in the upper right corner of the map whereas the low values are distributed in the middle to bottom left.

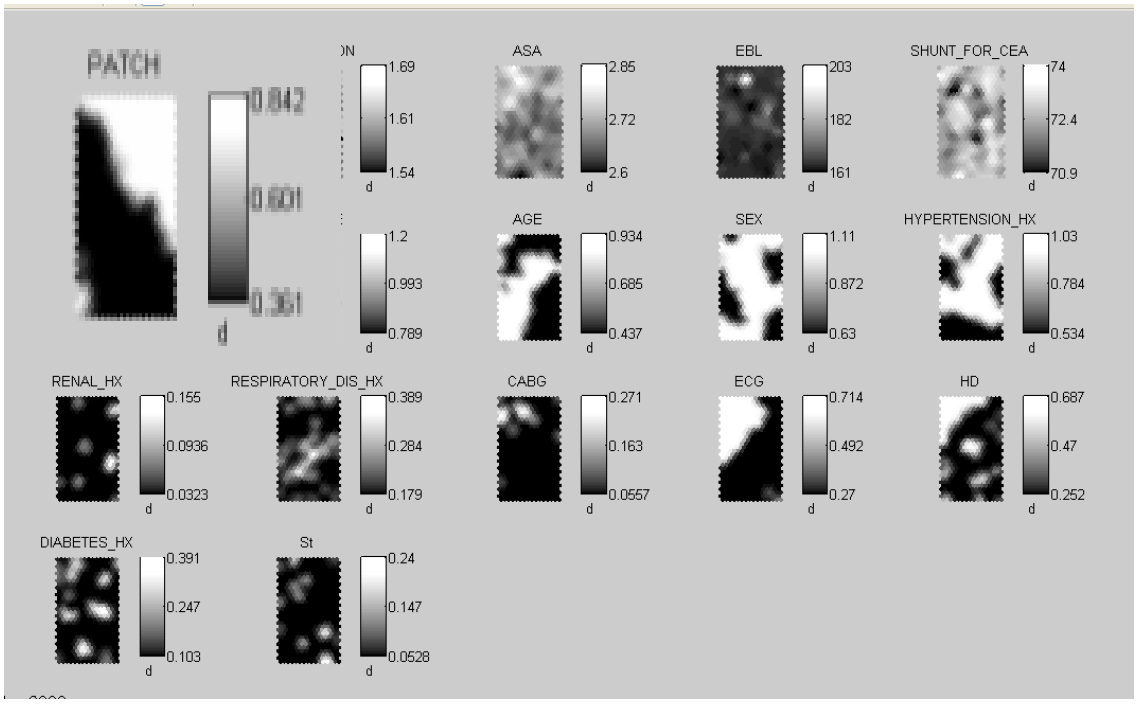

Figure 6.3: The component planes for the attributes in CM3bD model.

Also from Figure 6.3, attributes "EBL"; "RENAL_HX"; "CABG"; "RESPIRATORY_DIS_HX"; "DIABETES_HX"; and "St" seem to be similar, because they have low values distributed in a similar area of the map.

The SOM algorithm is applied to the input map to cluster the input data (see the results in Table 6.15). Assume that clusters "C1"; "C2"; "C3" correspond to the classes "High"; "Medium"; and "Low" respectively. According to Table 6.15, the accuracy rates are poor (less than 0.50). Figure 6.4 shows the clustering map where each generated cluster is coded

by the black and white colour scale. The expected labelled risks can be seen in the map as in Figure 6.5 below. From Figure 6.4 and Figure 6.5, it is suggested that the cluster on the top left of clustering map (cluster 2) can be seen as the "Medium" cluster. This is because many "Medium" labels are distributed in this area (Figure 6.5).

	C1 (High)	**C2 (Medium)**	**C3 (Low)**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
High	13	13	22	0.45	0.55	0.40	0.33	0.62
Medium	17	23	33					
Low	42	90	88					

Table 6.15: SOM-Clustering results for CM3bD model.

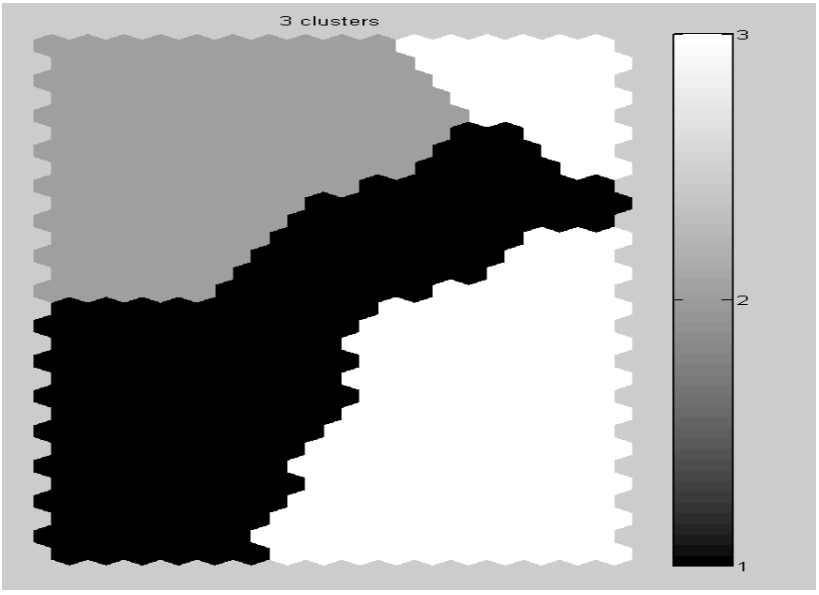

Figure 6.4: The clustering result for the map in Figure 6.2.

However, it seems to be difficult to identify "High risk" and "Low risk" clusters, because the labelled clustering map (Figure 6.5) does not show these clearly. This might demonstrate noise and the diffuse delineation of risk in the data. This needs further investigation.

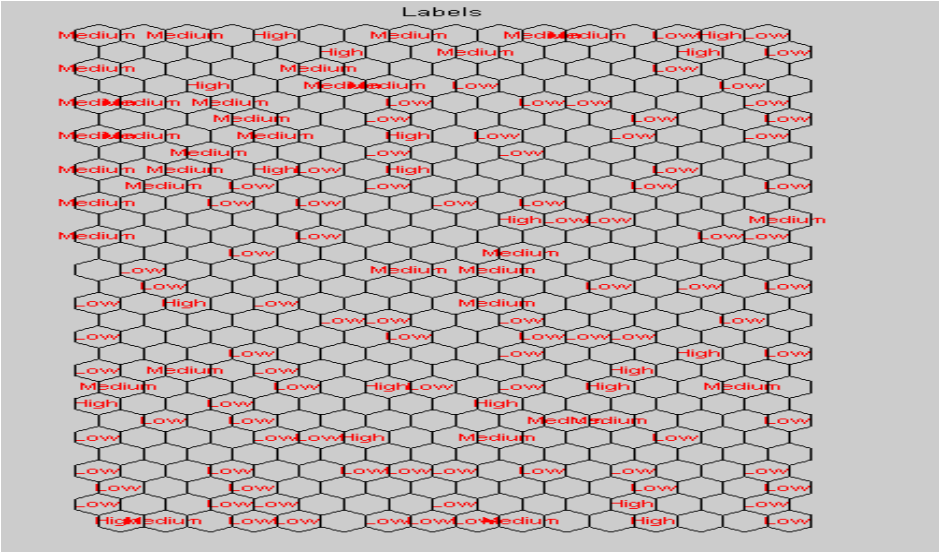

Figure 6.5: The labelled risks for the map in Figure 6.4.

6.5.4. Case Study IV

This section demonstrates the use of KMIX algorithm applied to two of the cardiovascular models (CM3aD and CM3bD) mentioned in the Case Study II and Case Study III. The detail of experimental steps can be seen in Appendix C (section C.4). The objective in this section is to discuss what the clustering reflects on the data structure as well as the differences between sensitivity versus positive predictive value.

Data

The given data sets contain 16 inputs, two attributes for the expected outcome calculations, and 341 patient cases. The preparation data tasks are shown in Case Study II and Case Study III. The expected outputs are calculated according to the following rules. Note that these outcomes are not involved to the clustering process. They are used just for comparison on the clustering results.

- **Model 1 (CM3aD)**: The two outcomes are calculated based on two attributes (*"PATIENT STATUS"* and *"COMBINE"), using*:

Σ(PATIENT STATUS, COMBINE) = 0 ➔ *"Low risk"*

Σ(PATIENT STATUS, COMBINE) ≥ 1 ➔"*High risk"*

- **Model 2 (CM3bD):** Three outcomes are given, using:

Σ(PATIENT STATUS, COMBINE) = 0 ➔ *"Low risk"*

Σ(PATIENT STATUS, COMBINE) = 1 ➔"*Medium risk"*

Σ(PATIENT STATUS, COMBINE) = 2 ➔*"High risk"*

Therefore, the model CM3aD has 57 values of "*High risk*"; and 284 values of "*Low risk*". The model CM3bD contains 48 values of "*High risk*"; 73 values of "*Medium risk*"; and 220 values of "*Low risk"*.

Method

Two models CM3aD and CM3bD are used with KMIX algorithm with alternative number of clusters (k=2, and k=3 respectively). Clustering results are then assigned as expected outputs for the new clustering models CM3aDC and CM3bDC. These models are then used with neural network techniques. The classification results are then discussed using standard measures.

Results and Discussions

The clustering results can be seen in Table 6.16 and Table 6.17 for models CM3aD and CM3bD respectively. Clusters of "C2" and "C1" correspond to the classes "High risk" and "Low risk" in Table 6.16; and clusters "C3"; "C2"; and "C1" correspond to the classes "High risk"; "Medium risk"; and "Low risk" in Table 6.17. These assumptions are based on the highest number of "High risk"; "Medium risk"; and "Low risk" patterns belong to the output clusters.

	C2 (High risk)	C1 (Low risk)	ACC	Sen	Spec	PPV	NPV
High risk	39	18	0.48	0.68	0.44	0.20	0.88
Low risk	158	126					

Table 6.16: Clustering results for CM3aD model.

	C3 (High risk)	C2 (Medium risk)	C1 (Low risk)	ACC	Sen	Spec	PPV	NPV
High risk	18	17	13	0.42	0.70	0.26	0.34	0.62
Medium risk	28	22	23					
Low risk	103	59	58					

Table 6.17: Clustering results for CM3bD model.

According to Table 6.16 and Table 6.17, the accuracy rates are poor (less than 0.50). For both the experiments, the sensitivity rates are more than double the positive predictive values. The sensitivity rates here mirror the correct clustered "High risk" patterns over total actual "High risk" expectations whereas the positive predictive values mirror the correct clustered "High risk" patterns over total clustered "High risk" outcomes. For example, 39 over 57 (expected) "High risk" patterns (horizontal comparison in Table 6.16) show the correct distributions (with a 0.68 of sensitivity). Contrastingly, 39 over 197 (clustering) "High risk" patterns (vertical comparison) show the correct distributions (with a 0.20 positive predictive value). Therefore, 39 "High risk" patterns here can be seen as the true "High risk" patterns. The other patterns (see in Table 6.16) can be explained as follows: 18 expected "High risk" patterns are clustered into "Low risk" class, because their forms are similar to this class form; 158 "Low risk" patterns are clustered into "High risk" class, because their pattern forms are similar to the "High risk" patterns. The other mis-clustering patterns in the confusion matrix in Table 6.16 and Table 6.17 are similarly explained. From this point, the poor performance of the clustering results might arise because of the poor

quality of clustering algorithm or the nature of the given data or the difficulty of measuring influential parameters. Therefore, further investigation is needed.

Two new models CM3aDC and CM3bDC, based on the clustering results, are built and then used with alternative neural network techniques (see results in Tables 6.18 and 6.19). For example, the multilayer perceptron is used with a topology of 16-0-1 (16 input nodes; 0 hidden nodes; and 1 output-2 class nodes for model CM3aDC) or 16-0-3 (16 input nodes; 0 hidden nodes; and 3 output nodes for model CM3bDC); with a learning rate of 0.3; and 100 training cycle epochs. Note that ten-fold cross-validation is used for all experiments.

Classifiers		C2H	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3aDC-MLP (MLP16-0-1; 0.3; 100 epochs)	*C2H*	188	9	0.95	0.95	0.94	0.96	0.94	0.04
	C1L	8	136						
CM3aDC- RBF (RBF_c=1)	*C2H*	189	8	0.95	0.96	0.93	0.95	0.94	0.03
	C1L	10	134						
CM3aDC-SVM (SVM_poly_p=1)	*C2H*	185	12	0.91	0.94	0.86	0.90	0.91	0.09
	C1L	20	124						

Table 6.18: Neural network results for CM3aDC model.

Classifier		C3H	C2M	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3bDC-MLP (MLP16-0-3; 0.3; 100 epochs)	*C3H*	97	0	1	0.96	0.98	0.89	0.96	0.98	0.02
	C2M	0	148	1						
	C1L	8	2	84						
CM3bDC-RBF (RBF_c=1)	*C3H*	81	7	10	0.92	0.94	0.87	0.95	0.85	0.06
	C2M	9	135	5						
	C1L	5	7	82						
CM3bDC-SVM (SVM_poly_p=1)	*C3H*	97	0	1	0.97	0.99	0.90	0.96	0.99	0.08
	C2M	0	149	0						
	C1L	7	2	85						

Table 6.19: Neural network results for CM3bDC model.

Overall, Tables 6.18 and 6.19 shows that the achieved accuracy rates are quite high (over 0.90). This is also true for the sensitivity rates and positive predictive values (over 0.90). There are very negligible differences between sensitivity rates and positive predictive

values. For example, the classifier CM3aDC-MLP (in Table 6.18) produces the positive predictive value 0.96 whereas its sensitivity of 0.95; and the classifier CM3bDC-RBF (in Table 6.19) produce the positive predictive value of 0.95 with its sensitivity of 0.94. Similar comparisons are found for the differences in the specificity rates and negative predictive values. These results demonstrate that the neural networks can replicate the clustering results, resulting from the KMIX algorithm. Therefore, the nature of the problem and the difficulty of measuring influential parameters might be the main reason to cause the poor clustering performance in Tables 6.16 and 6.17.

6.6. Discussion

To compare the use of neural network techniques and POSSUM and PPOSSUM systems, a best performance (in sensitivity rates) in Table 6.3 is chosen. The compared sensitivity rates results can be seen in Table 6.20. Note that the "Mortality" values in the POSSUM and PPOSSUM systems can be seen as the "High risk" values in the Hull_POSS model.

Classifier	No of cases	Predicted deaths/High risk	Reported deaths/High risk	Sensitivity rates	Ratios
Hull_POSS_TP1 (MLP_2H_0.3_500)	498	9	78	0.12	8.67
POSSUM	498	65	78	0.83	1.20
PPOSSUM	498	25	78	0.32	3.12

Table 6.20: Neural network and POSSUM and PPOSSUM sensitivities comparison.

It is clear from Table 6.20 that POSSUM and PPOSSUM have a higher performance than the neural network classifier. The POSSUM ratio O/E of 1.20 reflects that the system classified the "High risk" (Death) patients quite close to the reported "High risk" (Death).

Contrastingly, the neural network classifier (Hull_POSS_TP1) has a ratio of 8.67. It means neural network under-predicts deaths for “High risk” patients in data domain.

The POSSUM sensitivity is highest whereas the neural network classifier (Hull_POSS_TP1) produced the poorest sensitivity rate (0.12). Therefore, it seems that the use of linear method with POSSUM system is adequate for this data set. However, as indicated above and in Chapter 2, there is an ambiguity in the use of the (linear) evaluation method in the estimation of POSSUM and PPOSSUM. The predicted “High risk” or “Low risk” results depend upon the chosen threshold value in the risk scale. For example, in Table 6.9 and Table 6.10, threshold value for “High risk” (Death) and “Low risk” is the mean of the “Mortality” calculations. Therefore, there might be ambiguous interpretations for the categorical risks in the risk scale. This will be highlighted by dividing the categorical risk into a smaller scale such as “High risk”, “Medium risk”, “Low risk”.

Although the pattern recognition techniques (neural network) produced poor sensitivity results compared to POSSUM and PPOSSUM, they provided another in depth view into the data domain such as the observations and the evaluation of internal pattern forms via confusion matrix. Moreover, the neural network classifiers produced their results via the validation of independent test sets during the prediction process whereas POSSUM and PPOSSUM did not. By building alternative models for use with neural network techniques, the classifiers can avoid the ignorance of significant contributed factors into the patient risks as indicated in Kuhan et al (2003). For all of these reasons, neural network classifiers are used in this thesis

Case Study II shows the big gap between the sensitivity rates and positive predictive values. For example, the classifier MLP_TP3 (0H; η=0.3; 100 epochs) in Table 6.11 has sensitivity rate of 0.49 whereas its positive predictive value is 0.85.

The gap between the sensitivity rates and the positive predictive values is investigated by the use of unsupervised pattern recognition techniques (KMIX). These classifiers show the actual forms in the internal structure of the data which the supervised learning techniques could not recognise. The poor accuracy rates (less than 0.50) from the use of clustering techniques in the models of CM3aD (KMIX) and CM3bD (KMIX and SOM) suggest there is a problem in these models, such as the nature of the problem and the difficulty of measuring influential parameters. The reuse of supervised pattern recognition techniques for clustering models CM3aDC and CM3bDC derived from models CM3aD and CM3bD provided higher performance in the sensitivity estimations (in about 0.95 in average - see Tables 6.18 and 6.19). It is beyond the bounds of this thesis to give a clinical description for the pattern forms found to be significant in clustering as clinical trials would be required. As expected supervised results on clustering classes show high performance, and their results might be used as trained classifiers in a decision tool. This high performance for clustering models is in sharp contrast to the poor quality performance in the clinical models.

6.7. Summary

By using common attributes and the main factors in the Hull and Dundee sites, the thesis data can be divided into 6 clinical models (CM1; CM2; CM3a; CM3b; CM4a; and CM4b). Additionally, three scoring risk models (Mortality; Morbidity; and Death rate) are generated by a combination of data from the Hull site and the POSSUM and PPOSSUM

results. CM1 and CM2 are generated as the main models. The remaining models (CM3a; CM3b; CM4a; and CM4b) are based on these models with alternative inputs and output attributes. All these models will be used in Chapter 7 in the main thesis experiments.

The data derived from the thesis data domain is used in alternative models as the case studies before running experiments for the main thesis data models. The POSSUM and PPOSSUM produced better sensitivity rates compared to supervised neural network classifiers (Table 6.20). The poor sensitivity rates and the gap between the sensitivities and positive predictive values can be explained by the results from the use of unsupervised pattern recognition classifiers (KMIX). It is suggested that many "High risk" pattern forms are alike to the "Low risk" ones in the data domain. Hence, many of the "High risk" patients have similar medical symptoms to some of "Low risk" patients. Therefore, this ill-defined classification border might cause the poor sensitivity performance in the neural network classifiers.

The unsupervised clustering results also disagree with the outcomes for clinical models via the poor accuracy rates in KMIX and SOM. This is supported by the high performance of the pattern recognition classifiers on the clustering models. This might be due to the nature of the problem and the difficulty of measuring influential parameters. The next chapter will analyse in great detail pattern recognition classifiers with all the above models. This might provide some clarity on the performance of the classifiers in the clinical models.

Chapter 7

Results and Analysis

7.1. Introduction

This chapter analyzes the results from applying the classifier techniques introduced in Chapters 4 to the data models discussed in Chapter 6. Two categories of pattern recognition and data mining techniques are applied here: supervised neural networks (multilayer perceptron, radial basis function, and support vector machine); and the KMIX-unsupervised clustering algorithm. The classification results are measured and evaluated by using the standard measurements indicated in Chapter 3 such as confusion matrix, sensitivity, specificity, positive predictive value, and negative predictive value. This gives rise to a discussion on the performance of the classifiers, and the nature of the data.

7.2. Experiment Results

This section shows the results of all experiments with clinical models, scoring risk models, and clustering models. The detailed structure of all models can be seen in Chapter 6. The detailed data preparation and processing steps for each model can be seen in Appendix C (section C.5). Note that the experimental data is derived from a combination of the Hull and the Dundee data sites.

7.2.1. Clinical Models CM1 and CM2

Table 7.1 shows the results of neural network techniques applied to the CM1 and CM2 models. The techniques chosen in these experiments are multilayer perceptron, radial basis function, and support vector machine. The classifier labels can be understood as the name of models plus the classifier techniques applied. For example, *CM1-MLP* means the CM1 model is applied with a multilayer perceptron.

Classifiers	**Risk**	**Confusion Matrix**		**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
		High risk	**Low risk**					
	High risk	9	117					
CM1-MLP	*Low risk*	34	679	0.82	0.07	0.95	0.21	0.85
	High risk	0	126					
CM1-RBF	*Low risk*	0	713	0.85	0.00	1.00	N/A	0.85
	High risk	30	96					
CM1-SVM	*Low risk*	112	601	0.75	0.24	0.84	0.21	0.86
	High risk	6	133					
CM2-MLP	*Low risk*	27	673	0.81	0.04	0.96	0.18	0.83
	High risk	0	139					
CM2-RBF	*Low risk*	0	700	0.83	0.00	1.00	N/A	0.83
	High risk	24	115					
CM2-SVM	*Low risk*	125	575	0.71	0.17	0.82	0.16	0.83

Table 7.1: Experimental results of CM1 and CM2 models.

The detail of the techniques and their parameters are as follows: the multilayer perceptron technique is used here with a 25-2-1 topology (25 input nodes; 2 hidden nodes; 1 output - 2 class nodes), a learning rate η of 0.3, and 500 training epochs; the radial basis function classifier has centre parameter c of 2; and the support vector machine uses a poly kernel

function with the exponent parameter p of *2*. The 10-fold cross-validation method is used for these experiments.

Overall, the accuracy rates of all classifiers are over 0.70. The correct predicted "Low risk" rates (specificity) and the "Low risk" predictive rates (negative predictive value) are high whereas the equivalent rates for "High risk" are very poor (0.93; 0.84 vs 0.09; 0.19 on average). The two radial basis function classifiers, CM1-RBF and CM2-RBF, predicted all expected "High risk" patients as "Low risk". This again shows the technique's disadvantages for use with the thesis data.

From Table 7.1, the correct predicted "High risk" rates (sensitivity as well as the positive predictive value) are very poor (0.09 and 0.19 on average except CM1-RBF and CM2-RBF). The nature of the problem and the difficulty of measuring influential parameters might be the cause for these poor performances.

7.2.2. Clinical Models CM3a and CM4a

Table 7.2 shows the results for supervised neural networks on the clinical risk prediction models CM3a and CM4a. These models share the same expected outputs but their input sets are different (as indicated in Chapter 6). The techniques and their parameters used in these experiments are the same as the description in section 7.2.1 except for the reduced input set; for example, the topology of the multilayer perceptron is now 16-2-1 (16 input nodes; 2 hidden nodes; and 1 output - 2 class nodes).

Overall, the accuracy rates are over 0.77 for all classifiers. The sensitivity and positive predictive values ("High risk" predictions) are still very poor (average about 0.07 and 0.25 respectively) whereas the specificity and negative predictive value ("Low risk" predictions)

are high (about 0.95 and 0.84 in average). Also from Table 7.2, the radial basis function classifiers again show poorest results in the correct "High risk" predictions (none of "High risk" patients is correctly predicted).

Classifiers	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV
		High risk	Low risk					
CM3a-MLP	*High risk*	13	126	0.81	0.09	0.95	0.28	0.84
	Low risk	34	666					
CM3a-RBF	*High risk*	0	139	0.83	0.00	1.00	N/A	0.83
	Low risk	0	700					
CM3a-SVM	*High risk*	16	123	0.77	0.12	0.90	0.19	0.84
	Low risk	67	633					
CM4a-MLP	*High risk*	14	125	0.81	0.10	0.95	0.30	0.84
	Low risk	32	668					
CM4a-RBF	*High risk*	0	139	0.83	0.00	1.00	N/A	0.83
	Low risk	0	700					
CM4a-SVM	*High risk*	18	121	0.79	0.13	0.92	0.24	0.84
	Low risk	58	642					

Table 7.2: Experimental results of CM3a and CM4a models.

The results in Table 7.2 show the considerable distances between the sensitivity versus the positive predictive value as well as the specificity versus the negative predictive value. As indicated in Chapter 6, models CM3a and CM4a are derived from model CM2 with a smaller attribute set (reduction from 25 input attributes to 16 input attributes). Therefore, the selection of what are thought to be significant attributes does not improve classification performances. Furthermore, the poor gap between correct predicted "High risk" (sensitivity) and correct predictive "High risk" (positive predictive value) persists. Again,

the nature of the problem, the difficulty of measuring influential parameters, and the resultant poor mapping between input attributes and outcomes, is suspected.

7.2.3. Clinical Models CM3b and CM4b

The results of two models CM3b and CM4b can be seen in Table 7.3. These models share the same input attribute sets as in models CM3a and CM4a respectively. However, the expected output sets are expanded using alternative risk categories such as “Very High risk”; “High risk”; “Medium risk”; and “Low risk”. The hope is that the expansion of categorical risks will show an improvement in the classification results.

The evaluation measures here are based on confusion matrix with an assumption that the number of “Very High risk”, “High risk”, and “Medium risk” are referred to as the number of positive outcomes, and the number of “Low risk” is referred to as the number of negative outcomes. The network topologies and their parameters are the same as in the CM3a and CM4a experiments except the increased number of output nodes (4). This is required by the binary representation for the categorical outcomes of “Very High risk”; “High risk”; “Medium risk”; and “Low risk”. For example, the multilayer perceptron used with models CM3b and CM4b has topologies of 16-2-4 (16 input nodes; 2 hidden nodes; and 4 output nodes) and 14-2-4 (14 input nodes; 2 hidden nodes; and 4 output nodes).

Surprisingly from Table 7.3, all expected “Medium risk” patients are predicted into “Low risk” class except for the classifiers CM3b-SVM (only one pattern correctly falls into “Medium risk” class) and CM4b-MLP (one pattern falls into “High risk” class). This suggests that the combinations of data models and classifiers support just three levels of risks.

Classi-fiers	Risk	Confusion Matrix				ACC	Sen	Spec	PPV	NPV
		Very High risk	**High risk**	**Medi -um risk**	**Low risk**					
CM3b-MLP	*Very High risk*	2	1	0	16	0.84	0.04	0.97	0.25	0.84
	High risk	0	3	0	104					
	Medium risk	0	0	0	13					
	Low risk	5	13	0	682					
CM3b-RBF	*Very High risk*	0	0	0	19	0.85	0.00	1.00	N/A	0.83
	High risk	0	0	0	107					
	Medium risk	0	0	0	13					
	Low risk	0	0	0	700					
CM3b-SVM	*Very High risk*	0	3	0	16	0.79	0.08	0.90	0.14	0.83
	High risk	1	6	0	100					
	Medium risk	0	0	1	12					
	Low risk	13	46	8	633					
CM4b-MLP	*Very High risk*	0	1	0	18	0.83	0.07	0.96	0.27	0.84
	High risk	0	8	0	99					
	Medium risk	0	1	0	12					
	Low risk	0	27	0	673					
CM4b-RBF	*Very High risk*	0	0	0	19	0.85	0.00	1.00	N/A	0.83
	High risk	0	0	0	107					
	Medium risk	0	0	0	13					
	Low risk	0	0	0	700					
CM4b-SVM	*Very High risk*	2	4	0	13	0.8	0.14	0.90	0.21	0.84
	High risk	2	11	0	94					
	Medium risk	0	0	0	13					
	Low risk	17	45	9	629					

Table 7.3: Experimental results of CM3b and CM4b models.

The sensitivity rates and positive predictive values of all classifiers are very poor (an average of 0.06 and 0.22 respectively except the classifiers CM3b-RBF and CM4b-RBF).

Furthermore, the distances between these rates are not reduced. Therefore, the expansion of outcome risk labelling does not help to clarify and improve the classification results in particular with regard to "High risk" predictions.

7.2.4. Scoring Risk Models

This section demonstrates the results of the POSSUM and PPOSSUM systems via the scoring risk models as *Mortality*, *Morbidity*, and *Death rate*. Note that in the confusion matrix, the expected outcomes ("High risk" and "Low risk") are derived from the actual number of "Dead" and "Alive" patients. Table 7.4 shows the POSSUM and PPOSSUM results via the use of standard measurements.

Overall, all classifiers (Mortality, Morbidity, and Death rate) produce the similar standard measurement rates. For example, all accuracy rates are at about 0.83. From this point, the results seem to show that POSSUM and PPOSSUM classifiers have a stable performance in their classification process for the thesis data.

Classifiers	**Risk**	**Confusion Matrix**		**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
		High risk	**Low risk**					
	High risk	10	69					
Mortality	*Low risk*	15	405	0.83	0.13	0.96	0.40	0.85
	High risk	15	64					
Morbidity	*Low risk*	24	396	0.82	0.19	0.94	0.38	0.86
	High risk	10	69					
Death rate	*Low risk*	16	404	0.83	0.13	0.96	0.38	0.85

Table 7.4: Confusion matrix for scoring risk models.

Nevertheless, there still is a big gap between numbers of the correct "High risk" and the correct predictive "High risk" via the sensitivity and the positive predictive value. For

example, the Mortality classifier predicts 13% of correct predicted “High risk” (sensitivity of 0.13) whereas it predicts 40% of correct predictive “High risk” (positive predictive value of 0.40). Also according to this classifier’s confusion matrix, 10 over 79 (13%) “Death” patients (“High risk”) are correctly predicted whereas 69 over 79 “Death” patients are mis-predicted into “Alive” (“Low risk”) class. This might mean that these (69) mis-classified patients might have pattern forms (original data) similar to the “Alive” patients. In contrast, the vertical comparison in the confusion matrix shows 15 “Alive” (“Low risk”) patients are mis-predicted into “Dead” (“High risk”) class, perhaps because their pattern forms are similar to “High risk” patients. Note that these results are taken from the existing POSSUM and PPOSSUM risk assessment systems. Therefore, the performance in “High risk” prediction (sensitivity and positive predictive value) again shows the nature of the problem and the difficulty of measuring influential parameters. The use of unsupervised techniques to investigate data structure may help to resolve this.

7.2.5. KMIX Clustering Results

This section demonstrates the use of the KMIX algorithm for models CM3a and CM3b as the representatives for all the above models. Tables 7.5 and 7.6 show the results for two of these models. Clusters of “C2H” and “C1L” correspond to the classes “High risk” and “Low risk” in Table 7.5; and clusters “C4VH”; “C3H”; “C2”; and “C1” correspond to classes “Very High risk”; “High risk”; “Medium risk”; and “Low risk” in Table 7.6. These assumptions are based on the highest number of “Very High risk”; “High risk”; “Medium risk”; and “Low risk” patterns in the output clusters.

Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV
High risk	48	91	0.60	0.35	0.65	0.16	0.83
Low risk	248	452					

Table 7.5: The clustering results for model CM3a.

From Table 7.6, no pattern falls into cluster "C3H". This seems to indicate as there is no "High risk" class in the outcome set. In other words, there seems to be just three data clusters in the domain.

Risk	C4VH	C3H	C2M	C1L	ACC	Sen	Spec	PPV	NPV
Very High risk	7	0	6	6	0.45	0.89	0.38	0.18	0.96
High risk	43	0	33	3					
Medium risk	5	0	5	3					
Low risk	249	0	199	280					

Table 7.6: The clustering results for model CM3b.

There are considerable differences in the sensitivity versus positive predictive value as well as the specificity versus negative predictive value over the two Tables 7.5 and 7.6. For example, the clustering for model CM3a (in Table 7.5) achieves a sensitivity rate of 0.35 whereas its positive predictive value is 0.16. Note that the CM3a model is derived from the CM3aD model in section 6.5.4 in Chapter 6 (containing data just from the Dundee site) plus an expansion of patterns. Therefore, the increasing number of patterns improves a little the clustering performance in accuracy rate (0.48 in Table 6.16 in Chapter 6; 0.60 in Table 7.5 in Chapter 7).

Also by looking at the confusion matrix in Table 7.5, the gap between sensitivity versus positive predictive values might be explained in terms of the way the KMIX algorithm

works. 91 patterns, which are mis-clustered into “C1L” class, because their distances to the “Low risk” centre vector are smaller than their distances to the “High risk” centre. In other words, their forms are similar to the “Low risk” patterns in data space. Contrastingly, 248 “Low risk” patterns are mis-clustered into “C2H”, because they are nearer to the “High risk” centre vector than the “Low risk” centre. This means their pattern forms are similar to the “High risk” class patterns. The same explanation can be used for the distances between the sensitivity versus positive predictive value in Table 7.6. All the above misclassifications show that the data seems not to support the expected outcomes. To verify the output from the clustering algorithm, as repeatable as a classification, the use of neural network techniques for the KMIX outcomes is necessary.

7.2.6. KMIX Clustering Models Results

Here the KMIX results are used as expected outcomes for the new clustering models (CM3aC and CM3bC). Alternative neural network techniques are applied to these new models to discuss classification results. The CM3aC model has 296 expected values of “C2H” and 543 values of “C2L”. The CM3bC model contains 304 values of “C4VH”, no values of “C3H”, 243 values of “C2M”, and 292 values of “C1L”. The parameters for the neural network techniques are: multilayer perceptron is used with topologies of 16-2-1 (16 input nodes; 2 hidden nodes; 1 output - 2 class nodes for CM3aC); and 16-2-4 (16 input nodes; 2 hidden nodes; 1 output - 2 class nodes for CM3bC), learning rate η of 0.3, and 500 training epochs; radial basis function is used with number of centre c of 2; and support vector machine is used with the poly kernel function and exponent parameter p of 2. The number of cross-validation folds is 10. Tables 7.7 and 7.8 show the experimental results.

Overall, all classifiers produce very high accuracy rates (over 0.98) except classifier CM3aC-RBF (0.77) in Table 7.7.

The multilayer perceptron classifiers have highest performance in both Tables 7.7 and 7.8 according to all the standard measurements of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and mean square error (about 0.99 in all measured rates except 0.97 of negative predictive value in Table 7.8).

Classifier	**Risk**	**C2H**	**C1L**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
CM3aC-MLP (MLP_2H_0.3_500)	*C2H*	296	0	1	1	1.00	0.99	1	0
	C1L	2	541						
CM3aC-RBF (RBF_c=2)	*C2H*	230	66	0.77	0.78	0.77	0.65	0.86	0.14
	C1L	124	419						
CM3aC-SVM (SVM_poly_p=2)	*C2H*	293	3	0.99	0.99	0.99	0.99	0.99	0.01
	C1L	3	540						

Table 7.7: The CM3aC model results.

Classifiers	**Risk**	**C4VH**	**C2M**	**C1L**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
CM3bC-MLP (MLP_2H_0.3_ 500)	*C4VH*	302	1	1	0.99	0.99	0.99	0.99	0.97	0.01
	C2M	3	233	7						
	C1L	1	3	288						
CM3bC-RBF (RBF_c=2)	*C4VH*	300	2	2	0.98	0.99	0.97	0.98	0.98	0.01
	C2M	3	235	5						
	C1L	4	5	283						
CM3bC-SVM (SVM_Poly_p= 2)	*C4VH*	304	0	0	0.98	0.99	0.96	0.98	0.99	0.07
	C2M	0	240	3						
	C1L	0	12	280						

Table 7.8: The CM3bC model results.

Contrastingly, overall the radial basis classifiers (CM3aC-RBF and CM3bC-RBF) have the poorest performance according to all the evaluation measures (for example, 0.77; 0.78; 0.77; 0.65; 0.86; and 0.14 respectively in Table 7.7).

The classifiers CM3aC-MLP (Table 7.7) and CM3bC-SVM (Table 7.8) predict correctly all "C2H" and "C4VH" patterns into the right classes. Furthermore, there is very negligible difference between pair-rates of sensitivity versus positive predictive value as well as specificity versus negative predictive value. These rates mirror the correct predictions for all patterns as the vertical or horizontal comparisons in confusion matrix in both Tables 7.7 and 7.8. For example, there is no mis-classification in "C1L" class, and just 2 patterns are mis-classified into "C2H" class (see CM3aC-MLP confusion matrix in Table 7.7). As indicated above, these outcomes are derived from KMIX outcomes for models CM3a and CM3b. Therefore, these neural network classifiers replicate the KMIX clustering results. In other words, there are negligible overlap predictions for patient risks ("High risk" and "Low risk") in the confusion matrix. This suggests that the nature of the problem and the difficulty of measuring influential parameters might cause the poor clustering performance in the above sections in particular with regard to "High risk" patterns. A deeper investigation of the KMIX results is necessary.

7.3. Discussion

For the discussion in this section, two groups of classifiers are used. The first group contains the classifiers derived from clinical models and scoring models (see the results rewritten in Table 7.9). The second group contains classifiers derived from clustering

models (CM3aC-MLP; CM3aC-RBF; CM3aC-SVM; CM3bC-MLP; CM3bC-RBF; and CM3bC-SVM). In the first group, the radial basis function classifiers produced the poorest results of "High risk" patterns (sensitivity of 0.00 and positive predictive value of "N/A"). They again show their limited use for the risk prediction process in this data domain. The comparison of the accuracy rates of all classifiers can be seen in Figure 7.1. Overall, there is little fluctuation in the accuracy rates of all classifiers (an average of about 0.78). Although radial basis function classifiers produced the poorest performance for "High risk" patients (in the sensitivity and positive predictive value measurements), their accuracy rates (see in Figure 7.1) are the highest compared to other techniques (0.84 in the average). This is due to the accuracy rates representing the trade-off between correct "High risk" and correct "Low risk" predictions (their specificity rates and negative prediction values are the highest compared to other classifiers).

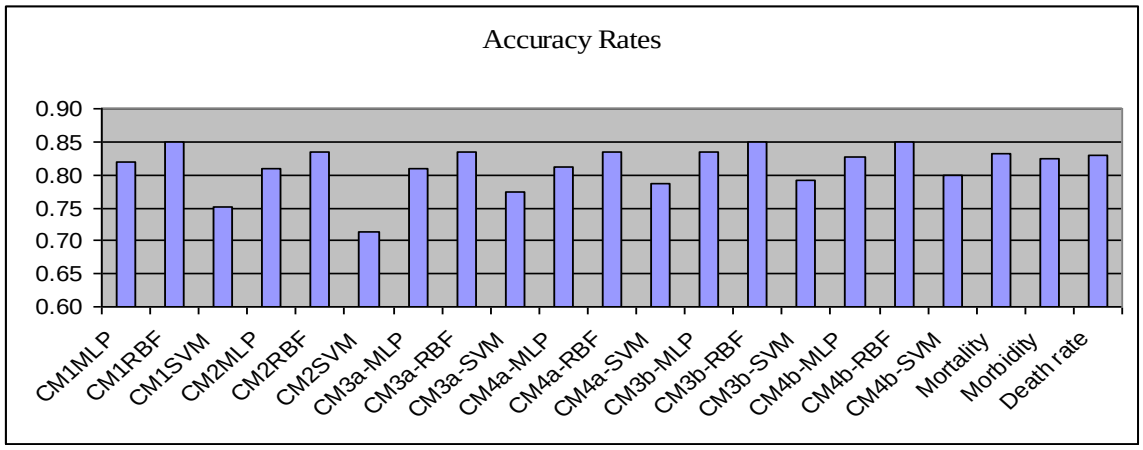

Figure 7.1: The comparisons of accuracy rates over all classifiers.

Classifiers	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
CM1-MLP	0.82	0.07	0.95	0.21	0.85
CM1-RBF	0.85	0	1	N/A	0.85
CM1-SVM	0.75	0.24	0.84	0.21	0.86
CM2-MLP	0.81	0.04	0.96	0.18	0.83
CM2-RBF	0.83	0	1	N/A	0.83
CM2-SVM	0.71	0.17	0.82	0.16	0.83
CM3a-MLP	0.81	0.09	0.95	0.28	0.84
CM3a-RBF	0.83	0	1	N/A	0.83
CM3a-SVM	0.77	0.12	0.9	0.19	0.84
CM4a-MLP	0.81	0.1	0.95	0.3	0.84
CM4a-RBF	0.83	0	1	N/A	0.83
CM4a-SVM	0.79	0.13	0.92	0.24	0.84
CM3b-MLP	0.84	0.04	0.97	0.25	0.84
CM3b-RBF	0.85	0	1	N/A	0.85
CM3b-SVM	0.79	0.08	0.92	0.14	0.85
CM4b-MLP	0.83	0.07	0.96	0.24	0.85
CM4b-RBF	0.85	0	1	N/A	0.85
CM4b-SVM	0.8	0.15	0.91	0.23	0.86
Mortality	0.83	0.13	0.96	0.4	0.85
Morbidity	0.82	0.19	0.94	0.38	0.86
Death rate	0.83	0.13	0.96	0.38	0.85

Table 7.9: Results of first group's classifiers.

To compare all the first group classifiers' performance with the random classification performance, three random representative classifiers are built. These random results are based on three types of heuristic outcomes labelling ("*PATIENT STATUS*" attribute; "*PATIENT STATUS*" and "*30D stroke/death"*; and POSSUM and PPOSSUM threshold) indicated in Chapter 6. The prediction rates (in Table 7.10) for the random classifiers are calculated as follows:

SenRand = p(High risk) - number of "High risk" over total number of data set.

SpecRand = q(Low risk) - number of "Low risk" over total number of data set

ACCRand = PRand(true positive $\cup$ true negative), where PRand is random probability of total true positive and true negative in risk predictions.

$$ACCRand = PRand(truepositive) + PRand(truenegative) - PRand(truepositive \cap truenegative)$$

$$ACCRand = p(High\ risk)*p(High\ risk) + q(Low\ risk)*q(Low\ risk).$$

For example, model CM1 contains 126 "High risk" and 713 "Low risk". Therefore, the random sensitivity is the probability of 126 "High risk", from a total of 839 patterns, classified as "true positive" (0.15). The random specificity rates are given via a similar explanation.

The random accuracy is the probability of total number of "High risk" and "Low risk" classified to correct "High risk" and "Low risk" classes respectively. Therefore, it is calculated as the total of probability of true "High risk" (true positive), and probability of true "Low risk" (true negative), minus the probability of intersection of true positive and true negative (= $\varnothing$). The random classification results can be seen in Table 7.10.

Random Classifiers	**ACC (Rand)**	**Sen (Rand)**	**Spec (Rand)**
Random1 (CM1)	0.74	0.15	0.85
Random2 (CM2;CM3a,b;CM4a,b)	0.72	0.17	0.83
Random3 (POSSUM and PPOSSUM)	0.73	0.16	0.84

Table 7.10: Results of random classifiers for first group's models.

Tables 7.9 and 7.10 show that all classifiers have higher accuracy compared to the random classifiers (Random1 and Random2) in overall (0.84 and 0.73 in average respectively). The neural network classifiers have poorer sensitivity and higher specificity rates than the random classifiers except the radial basis function classifiers. The POSSUM and PPOSSUM classifiers (Table 7.9) have nearly the same in the average of the sensitivity rate (0.15) compared to the random predictions (Table 7.10). These results suggest that the POSSUM and PPOSSUM classified "High risk" patients to equivalent level as the random and better than the neural network classifiers.

Importantly, the random specificity rates in Table 7.10 are five times higher than the sensitivity rate. This reflects the much higher density of "Low risk" patterns in data space. This might cause the poor sensitivity performance for all classifiers.

The average classification rates for the four subgroups (multilayer perceptron, support vector machine, radial basis function and the POSSUM/PPOSSUM calculations) can be seen in the Table 7.11. Here, the POSSUM and PPOSSUM achieved the highest overall results compared to the other techniques. The accuracy rates show negligible differences between the four techniques. However, there are considerable differences between the correct "High risk" (sensitivity) and correct predictive "High risk" (positive predictive value) except for the radial basis function classifier. All classifiers demonstrate poor performance in "High risk" predictions.

The average classification performances for the second group (clustering models) are shown in Table 7.12.

Techniques	ACC	Sen	Spec	PPV	NPV
MLP	0.82	0.07	0.96	0.24	0.84
RBF	0.84	0	1	N/A	0.84
SVM	0.77	0.15	0.89	0.2	0.85
POSSUM/ PPOSSUM	0.83	0.15	0.96	0.39	0.86

Table 7.11: The average classification rates of subgroups models.

Techniques	ACC	Sen	Spec	PPV	NPV
MLP	0.99	0.99	0.99	0.99	1
RBF	0.88	0.88	0.87	0.82	0.93
SVM	0.99	0.99	0.98	1	1

Table 7.12: The average classification rates of clustering models (second group).

The multilayer perceptron and support vector machine classifiers achieve the highest measurement rates (sensitivity, specificity, accuracy, positive predictive value, and negative predictive value) whereas the radial basis function classifiers achieve the poorest. Specially, there is negligible difference between the sensitivity and positive predictive value rates over these techniques. This is also true for the differences of the specificity rates and negative predictive values. This means these results demonstrate that the neural networks can replicate the clustering results from the KMIX algorithm. Therefore, the poor clustering performance in Tables 7.5 and 7.6 above might be influenced by the nature of the problem and the difficulty of measuring influential parameters or the KMIX clustering performance failing to find centre vectors for patterns for the different outcomes.

To further investigate the KMIX performance, the distance in data space from expected classes (“High risk” and “Low risk”) to the clustering outcomes (“High risk” and “Low risk” clusters) are calculated. Furthermore, the gaps between different groups in the

classifier's confusion matrix are also calculated to investigate the distribution of the data in the attribute space. Note that these distances are calculated based on the distances between centre vectors, which are representative for the groups.

Mathematically, the centre vector contains m components, with p first continuous components and m-p categorical components. Note that the Boolean components are treated as if categorical. Therefore, the centre vectors can be rewritten in the form:

$$Q= (q_{j1,}\ q_{j2,}\ldots\ q_{jp},\ q_{jp+1,}q_{jp+2,}\ldots,\ q_{jm}),$$

where $\{q_{jk}\}_{k=1,..p}$,={mean_k}, and mean_k is the average of k^{th} continuous attribute; $\{q_{jk}\}_{k=p+1,..,\ m}$ ={mode_k}, and $mode_k$ is the *"max frequency of* Val_{Ck} " in the k^{th} categorical attribute.

The distance between two centre vectors is calculated as follows:

$$d(Qi,Qj) = dN(Qi,Qj) + dC(Qi,Qj);\ j=1,2,..k$$

where $d^N(Q_i,Q_j)$, and $d^C(Q_{i,}Q_j)$ are calculated according to Equation (4.14), and Equation (4.15) in Chapter 4.

In summary, this distance is calculated based on Euclidean distances for the continuous attributes and Hamming distances for the categorical attributes. More detail about the centre vector as well as distance calculations can be seen in section 4.3.2 of Chapter 4.

The CM3a clustering results are used here as the representative for this investigation (see confusion matrix result in Table 7.5). The resultant distances can be seen in Table 7.13 and Table 7.14 below. Note that "Expected High" means the expected "High risk" ("High risk" class), and "Cluster High" means "High risk" cluster. The same explanation is used for "Expected Low" and "Cluster Low".

Groups	Distances
Expected High – Cluster High	2.00
Expected High – Cluster Low	0.01
Expected Low – Cluster Low	0.00
Expected Low – Cluster High	2.01

Table 7.13: The distances from expected classes to alternative clustering outcomes.

From Table 7.13, the distances between cluster "Low risk" to "High risk" and "Low risk" classes (expectations) are negligible (nearly 0.00). Contrastingly, the distances between cluster "High risk" to both these classes are quite far (2.01). This shows the reason for the poor clustering "High risk" performance as in their original distributions the patterns in the expected "High risk" class are not close to each others.

Table 7.14 shows the negligible distance (0.01) between correct "High risk" group and incorrect "Low risk" group (in the "High risk" cluster). This means these patterns are very closely distributed in data space. In other words, they have similar pattern forms. The same explanation might be used for the gap between correct "Low risk" group and incorrect "High risk" group (with a distance of 1.01).

Contrastingly, the distances between alternative clustering groups are more than double the two above distances. For example, the distance from the correct "High risk" patterns to the incorrect "High risk" is 3.02. This means their distributions in data space are quite far from each other. Hence, the correct "High risk" pattern forms are different to the incorrect "High risk" ones. However, their patterns have the same outcomes as labelled using the heuristic

rules indicated in Chapter 6. From this result, it is strongly suggested that the natural structure of the data (similar pattern forms) does not support the labelled outcomes from the heuristic clinical models. In other words, the nature of the problem and the difficulty of measuring influential parameters cause the poor performance for classifiers on the data, and in particular for the "High risk" patterns.

Groups	**Distances**
Correct High – Incorrect High (True Positive – False Negative)	3.02
Correct High – Correct Low (True Positive – True Negative)	2.03
Correct High – Incorrect Low (True Positive – False Positive)	0.01
Correct Low – Incorrect High	1.01
Correct Low – Incorrect Low (True Negative – False Positive)	2.02
Incorrect Low – Incorrect High (False Positive – False Negative)	3.01

Table 7.14: The distances between alternative groups in confusion matrix.

To discuss the relationship between all supervised classifiers and clustering classifiers, the model CM3a is chosen as the representative for all the above models. A rule to build the confusion matrix for all supervised classifiers (multilayer perceptron; radial basis function; and support vector machine) is given by:

IF the outcome is “High risk”, and for any supervised classifiers (MLP, RBF, and SVM) it is predicted as “High risk”, it can be seen as true positive “High risk”

Otherwise, it can be seen as false negative “High risk”.

The same is done for the “Low risk” patterns. The results can be seen in Table 7.15.

	High risk	**Low risk**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
High risk	14	125	0.76	0.10	0.89	0.15	0.83
Low risk	78	622					

Table 7.15: The confusion matrix for CM3a model with all supervised classifiers.

A confusion matrix for the model CM3a in clustering classifier (see in Table 7.5 above) can rewritten as:.

Risk	**High risk**	**Low risk**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
High risk	48	91	0.60	0.35	0.65	0.16	0.83
Low risk	248	452					

Table 7.16: The confusion matrix for CM3a model with clustering classifier (KMIX).

The confusion matrix resulted by combination of the supervised and unsupervised (clustering) classifiers can be seen in Table 7.17. The rule to generate the outcomes as follows:

If the outcome is “High risk”, and for any supervised classifiers (MLP, RBF, and SVM) it is predicted as “High risk”, or the clustering resulted as “High risk”, it can be seen as true positive “High risk”

Otherwise, if it is predicted as “High risk” it can be seen as false negative “High risk”.

A similar rule exists for the “Low risk” patterns:

IF the outcome is “Low risk”, and for any supervised classifiers (MLP, RBF, and SVM) it is predicted as “Low risk” or the clustering result is “Low risk”, it can be seen as true negative “Low risk”

Otherwise, if it is predicted as “Low risk” it can be seen as false positive “Low risk”.

	High risk	**Low risk**	**ACC**	**Sen**	**Spec**	**PPV**	**NPV**
High risk	58	81	0.55	0.42	0.57	0.16	0.83
Low risk	300	400					

Table 7.17: The results of a combination between clustering classifier (KMIX) and supervised classifiers.

Although the accuracy rate in Table 7.15 is highest (0.76) but the sensitivity rate and positive predictive value is poorest (0.10 and 0.15). This means all supervised classifiers poorly predicted for “High risk” patients. Table 7.16 shows the improvement of predicted sensitivity rate (0.35) although the accuracy rate is poorer (0.60). The big improvement for predicted “High risk” patient via sensitivity rate can be seen in Table 7.17 (0.42). This result is generated by a combination of the supervised classifiers (multilayer perceptron, radial basis function, and support vector machine) and the clustering classifier (KMIX). As indicated above the natural structure of the data affects the supervised classifiers’ results. Therefore, the combination between supervised and unsupervised might open the new direction for the prediction process. This is left for further research.

7.4. Summary

This chapter shows and discusses the classification results for the combined data from the Hull and the Dundee sites. The discussions about both "clinical - structure" classifiers (clinical and scoring risk models) and “natural – structure” classifiers (clustering models)

are shown in greater detail in section 7.3 above. The "clinical - structure" classifiers produced similar rates of accuracy, specificity, and negative predictive value. However, all produced poor performance for “High risk” predictions (sensitivity and positive predictive value). Furthermore, there are big gaps between the sensitivity and the positive predictive value. This is also true for the use of the KMIX algorithm. Therefore, an investigation of the distances of the risk groups in KMIX confusion matrix was needed. The resultant distances show that the patterns in the same KMIX cluster have similar forms (small distances) whereas the expected classes (clinical heuristic outcome) have quite different pattern forms (high distances). Furthermore, the “natural - structure” classifiers produced high performance in the reuse of neural network techniques to replicate the clustering outcomes. Therefore, the poor performance for “High risk” predictions in "clinical - structure" classifiers can be explained by the nature of the problem and the difficulty of measuring influential parameters.

In all the above classifiers, each attribute is treated equally. The next chapter investigates whether certain attributes might be more important than the others in determining the model outcomes. Furthermore, the stressing of significant attributes during the clustering process makes for an interesting investigation.

Chapter 8

Feature Selection and Mutual Information

8.1. Introduction

Feature selection can help in data mining by reducing the number of irrelevant and redundant features, which often degrade the performance of classification algorithms in both speed and prediction accuracy. Most feature selection methods use evaluation functions and search procedures to achieve their targets. These evaluation functions measure how good a specific subset of input attributes is in discriminating between the outcome classes. Feature selection methods can be seen as belonging to two main groups: filters and wrappers. The former will be used in this thesis as they are faster and simpler than the wrapper techniques (Dash and Liu, 1997). Furthermore, they are useful for clustering, as they present the distance between the attributes to the outcomes. The concept of mutual information is also introduced in this chapter. It measures the dependencies between random variables. Therefore, it is suitable for assessing "information content" of the attribute contributions to outcome classes in the data domain. For example, it was used in O'Connor and Walley (2000) for measuring the quality of a self organizing map. It is applied here with the KMIX algorithm as the attribute weights in the clustering process; in an attempt to improve the quality of the clustering performance.

8.2. Feature Selection

This section introduces the two feature selection methods of filters and wrappers. Filters measure independently the relevance of feature subsets to classifier outcomes whereas wrappers use the classifier's performance as evaluation function. A representative of the filter concept, the Relief algorithm (Kira and Rendell, 1992), is also introduced in this section. This is for a comparison purpose with mutual information in the following case study.

8.2.1. Filter Method

Each feature is evaluated with a measure such as the distance to outcome classes. All features in the data set are then ranked according to these measures. The first m features, from the ranked list, can be chosen by the user. The methods for choosing m features are not described in detail in this thesis. More detail about this method, and the definition of m features, can be seen in Liu and Motoda (1998).

8.2.2. Wrapper Method

The wrapper method is used as an inductive algorithm to estimate the value of a given feature subset (e.g via cross-validation). This means its goal is to return a subset of features that gives the lowest prediction error. However, according to Dash and Liu (1997) the algorithms exhibit a moderate complexity, because the number of executions requires a high computational cost, in particular when used with exhaustive search strategies. Therefore, this method is not used for the thesis data. Further details about the wrapper method can be found in Dash and Liu (1997); Liu and Motoda (1998); and Talavera (2005).

8.2.3. Relief Algorithm

The Relief algorithm (Kira and Rendell, 1992) is a filter method that estimates the usefulness of attributes according their values in distinguishing samples that are near each other. The algorithm searches for two nearest neighbors of each sample in the data domain in the following way: Firstly, it compares one pattern from the "nearest hit" class with another from the "nearest miss" mis-class. It updates the quality estimation according to the "miss" and the "hit" value. This process is repeated until a best ordering for the attribute set is found; or is terminated by some user defined parameter. The pseudo code for the Relief algorithm can be seen in Figure 8.1.

Algorithm Relief

Input: For each training instance a vector of attribute values and the class value

Output: The vector W of estimations of the qualities of attributes

set all weights W[A]:=0.0;

for i:=1 to m **do begin**

randomly select an instance R_i;

Figure 8.1: Pseudo code of original relief algorithm (Kira and Rendell, 1992).

In Figure 8.1, A is the current attribute; $W[A]$ is the weight of the currently considered attribute; R_i is the i^{th} sample; H is the *"hit"*; M is the *"miss"*; $diff()$ is the probability function; and m is number of the neighbors.

More detail on the Relief-family of algorithms can be seen in Kononenko(1994); and Lopez(2002).

8.3. Mutual Information

This section introduces the concepts of mutual information, entropy, and a measure of mutual information based on Bayes' theory.

Mutual Information has been applied to many areas to help reduce the feature set by evaluating the significant attributes in the feature set. For example, Battiti (1994) produced the MIFS algorithm (Mutual Information based Feature Selection) to classify sonar data (Gorman and Sejnowski, 1988), Iris data (Fisher, 1936), and for use in optical character recognition. Recently, mutual information has been used in feature selection algorithms such as in Huang et al (2006), Sanchez et al (2008), Liu et al (2008), and Schaffernicht et al (2009). Huang et al (2006) produced a wrapper method to find a subset of features applied to the benchmark data sets (UCI, Merz & Merphy, 1996). Sanchez et al (2008) used mutual information to discover the most important variables for a fuzzy rule-based system for use with benchmark imprecise data. Liu et al (2008) shown the advantage of using mutual information in a feature selection algorithm compared to other filter methods over thirty-three datasets. Schaffernicht et al (2009) proposed a feature selection algorithm using residual mutual information as selection criterion to evaluate the output and the input features with the UCI data sets (Merz & Merphy, 1996). They concluded that the resulting performance is on par with the other approaches, but it needed fewer adaptation cycles. All of these researches are the motivation for the use of mutual information in the feature selection algorithm with the thesis data.

8.3.1. Notations

Entropy

In information theory, Shannon (1948) defined entropy or information entropy as a measure of the uncertainty associated with a discrete random variable. In other words, entropy is a measure of the average information content of the missing recipients when the system does not know the value. Mathematically, entropy can be written as:

$$H(X) = -c\sum_{i=1}^{n} p_i(x)\log\, p_i(x) \quad (8.1)$$

where $p_i(x)$ is probabilities of occurrence in a set of possible events x (i.e. the transaction in cardiovascular risk prediction), n is number of transactions, and c is a positive constant (usually consider $c=1$).

Joint Entropy

Suppose there are two discrete variables X and Y. $H(X,Y)$ is the joint entropy, given by:

$$H(X,Y) = -c\sum_{i=1}^{n}\ \sum_{j=1}^{m} p_{i,j}(x,y)\log\, p_{i,j}(x,y) \quad (8.2)$$

where $p_{i,j}(x,y)$ is the probability of the joint occurrence of x and y.

Condition Entropy

The conditional entropy of Y is $H_X(Y)$ defined as average of the entropy of Y for each value of x, weighted according to the probability of that particular x.

$$H_X(Y) = -c\sum_{i,j=1}^{n} p_{i,j}(x,y)\log\, p_i(y) \quad (8.3)$$

$$H_Y(X) = -c\sum_{i,j=1}^{n} p_{i,j}(x,y)\log\, p_j(x) \quad (8.4)$$

where $p_{i,j}(x,y)$ is the probability of the joint occurrence of x and y; and $p_i(y)$ and $p_j(x)$ are conditional probabilities of X, and Y.

$$p_j(y) = \frac{p_{i,j}(x,y)}{\sum_j p_{i,j}(x,y)} \quad (8.5)$$

Relative Entropy

The relative entropy is a measure of the statistical distance between two distributions. It is originally introduced by Solomon Kullback and Richard Leibler in 1951. It is known as the Kullback Leibler distance; or Kullback Leibler divergence (Cover and Thomas, 1991).

$$K(p,q) = \sum_{x\in A} p(x)\log(\frac{p(x)}{q(x)}) \quad (8.6)\,,$$

where $p(x)$, and $q(x)$ are the distributions in the data set A.

Quantization

Quantization, also called discretization, is the process of converting continuous variables into discrete variables. The discretized variable has a finite number of values which is considerably smaller than the number of possible values in the original data set. The continuous values are mapped to alternative intervals (bins) in the attribute value range. According to Venables and Ripley (1994); Yang et al (2001); and Tourassi et al (2001), the proper number of the *bin* is given by:

$$bin = \log_2 N + 1 \qquad (8.7),$$

where N is total number of patterns.

8.3.2. Mutual Information

The mutual information between discrete random variables X and Y, $MI(X,Y)$, is a measure of the amount of information in X that can be predicted when Y is known.

For the case where X and Y are discrete random variables, $MI(X,Y)$ can be written as:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) \\ &= \sum_i \sum_j p_{i,j}(x,y) \log[\, p_{i,j}(x,y) \,|\, p_i(x)\, p_j(y)] \end{aligned} \qquad (8.8)$$

where $H(X)$ is the entropy of X, $H(X|Y)$ (or $H_X(Y)$) is the conditional entropy, which represents the uncertainty in X after knowing Y.

The concept of entropy can be extended to continuous random variables (Shanon, 1948; Cover and Thomas, 1991) by:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) \\ &= \int p_{i,j}(x,y) \log[\, p_{i,j}(x,y) \,|\, p_i(x)\, p_j(y)]\, dx\, dy \end{aligned} \qquad (8.9)$$

Based on the Kullback Leibler divergence given in Equation (8.6), the Equation (8.8) and Equation (8.9) can be written as follows:

$$MI(x,y) = K(P(x), P(x).P(y)) = \sum_{x \in X} \sum_{y \in Y} P(x,y).\log \frac{P(x,y)}{P(x).P(y)} \quad (8.10)$$

$$MI(x,y) = K(P(x), P(x).P(y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x,y).\log \frac{P(x,y)}{P(x).P(y)} dxdy \quad (8.11)$$

In this thesis, the continuous values are transformed into discrete ones by using the above quantization method. So from now on, the mutual information formula is written in the discrete case only.

Bayes' Theorem for Mutual Information in Pattern Recognition

Assume that the output classes set $C= \{C_i\}, i=1,..,c$ and the attributes set $X= \{x_j\}, j= 1,..m$ are two sets of random variables in the data domain. According to the Bayes theorem:

$$p(C_i \mid x_j) = \frac{p(x_j \mid C_i)\, p(C_i)}{p(x_j)} \quad (8.12)$$

Therefore,

$$p(C_k, x_{jk}) = p(C_k \mid x_{jk})\, p(x_{jk}) \quad (8.13);$$

where $p(C_k,x_{jk})$ can be seen as the probability of finding attribute x_j in class C_k. in the k^{th} state.

To evaluate the relevance between classes and attributes in the data set, the mutual information is calculated. By applying Equation (8.10) for class C_i $(i=1,..c)$, and attributes x_j $(j=1$ to $m)$, the mutual information is calculated as:

$$MI(C, x_j) = K(P(C), P(C).P(x_j)) = \sum_{i=1}^{c} \sum_{k=1}^{s} p(C_i, x_{jk}) \log \frac{p(C_i, x_{jk})}{p(C_i).p(x_{jk})} \quad (8.14)$$

where $p(C_i,x_{jk})$ is probability of finding attribute x_j in class C_i in the k^{th} state; $p(C_i)$ is prior probability of class C_i; $p(x_{jk})$ is prior probability of finding attribute x_j in the k^{th} state; c is number of classes, s is number of states in the considered attribute, and m is number of attributes.

It clear that the probability of finding attribute x_j in class C_i in k^{th} state, ($p(C_i,x_{jk})$), is the probability of finding number of patterns in class C_i with attribute x_j when considering the k^{th} state.

In short, Equation (8.14) can be rewritten as follows:

$$MI(C, x_j) = \sum_{i=1}^{c} \sum_{k=1}^{s} p_{ijk} \log(\frac{p_{i\,jk}}{q_i.r_{jk}}) \quad (8.15),$$

where $p_{ijk} = \frac{sum_{ijk}}{sum}$, sum_{ijk} is total number of patterns in class C_i with attribute x_j and in the k^{th} state, and sum is total number of patterns;

$q_i = \frac{sum_i}{sum}$, and sum_i is total number of patterns belong to class C_i;

$r_{jk} = \frac{sum_{jk}}{sum}$, and sum_{jk} is total number of patterns in attribute x_j and in the k^{th} state.

Example 8.1

Table 8.1 below shows information about cardiovascular patients. The input attributes are labeled as *30D_St/D* (x_1), *Diabetes* (x_2), *SEX* (x_3), and *Age* (x_4). The output attribute is

labeled as *Risk*. Assume that three output classes are labeled as C_1*="High",* C_2*="Medium",* and C_3 *= "Low"*.

The calculation for the mutual information for the discrete attribute"30D_str/D" is as follows: $p(C_1,x_{11})$ (finding *30D_St/D* = "Y" (x_{11}) in "High" class (C_1)) can be seen as the ratio of the number patients with *Risk* = "*High*", and *30D_St/D* = "*Y*" over the total number of patients in the data set. Hence, $p(C_1,x_{11}) = 2/10$. Similarly, the prior probability of class C_i, $p(C_j)$, is probability of finding the number of patterns in class C_i; and $p(x_j)$ is the probability of finding the number of patterns of attribute x_j in the k^{th} state. Therefore, $p(C_1) = 3/10$, $p(x_{11})=6/10$.

30D_St/D (x_1)	**Diabetes (x_2)**	**SEX (x_3)**	**AGE**	**AGE_BIN (x_4)**	**Risk (C)**
Y	N	M	90	4	Medium
Y	Y	F	45	2	High
Y	Y	M	40	2	High
N	Y	F	74	4	High
Y	N	M	20	1	Medium
Y	N	F	42	2	Medium
N	N	M	25	1	Low
N	N	F	45	2	Low
Y	N	F	86	4	Medium
N	N	M	68	2	Low

Table 8.1: Cardiovascular patient information.

$MI(C,x_1)$ is given by:

$$MI(C, x_1) = \sum_{i=1}^{3} (p(C_i x_{11}) \log \frac{p(C_i, x_{11})}{p(C_i).p(x_{11})} + pC_i x_{12}) \log \frac{p(C_i, x_{12})}{p(C_i).p(x_{12})})$$

Hence,

$$MI(C,x1)= (0.030-0.126) + (0.295+0) + (0+0.396) =0.595$$

Similarly, $MI(C,x_2)$, and $MI(C,x_3)$ are calculated as:

$$MI(C,x2)= (0.521+ 0) + (0 + 0.206) + (0+0.29) = 0.757$$

$$MI(C,x3)= (0.083- 0.058) + (0 + 0) + (-0.1 + 0) =-0.075$$

Therefore, it is clear that x_2 (*Diabetes*) has a stronger relevance than other attributes (0.757), and the least relevant attribute to the output classes is the *SEX* attribute (-0.075). As indicated above, the discretization method is used for attribute "Age" to convert its continuous values into discrete values. The number of *bins* is calculated as Equation (8.7). Hence, we have *4 bins* ($\approx log_2 10$).

$$MI(C, x_4) = \sum_{i=1}^{3} (p(C_i x_{41}) \log \frac{p(C_i, x_{41})}{p(C_i).p(x_{41})} + pC_i x_{42}) \log \frac{p(C_i, x_{42})}{p(C_i).p(x_{42})} + pC_i x_{43}) \log \frac{p(C_i, x_{43})}{p(C_i).p(x_{43})})$$

$$MI(C,x_4)= (0 + 0.083 + 0.015) + (0.032 – 0.1 +0.147) + (0.074 + 0.082 +0) = 0.334.$$

8.4. Case Study V

This section demonstrates the use of mutual information for the thesis data. The results are discussed by comparison to the use of the Relief algorithm. More detail can be seen in section C.6 in Appendix C.

Data

Two models are used in this experiment, CM3aD and CM2. CM3aD contains 16 input attributes and 341 cases whereas model CM2 contains 26 input attributes and 839 cases.

All data is cleaned by supplying missing values, and transformed to appropriate categorical types. The number of "continuous to discrete value" bins for the models CM3aD and CM2 are 9 and 11 respectively.

Method

Both these models are used with the WEKA software package (WEKA, 2005) for the Relief algorithm. Mutual information calculations are applied to both these models using Equation (8.15).

Result

The comparison result for the use of the Relief algorithm and mutual information can be seen in Figures 8.2 and 8.3.

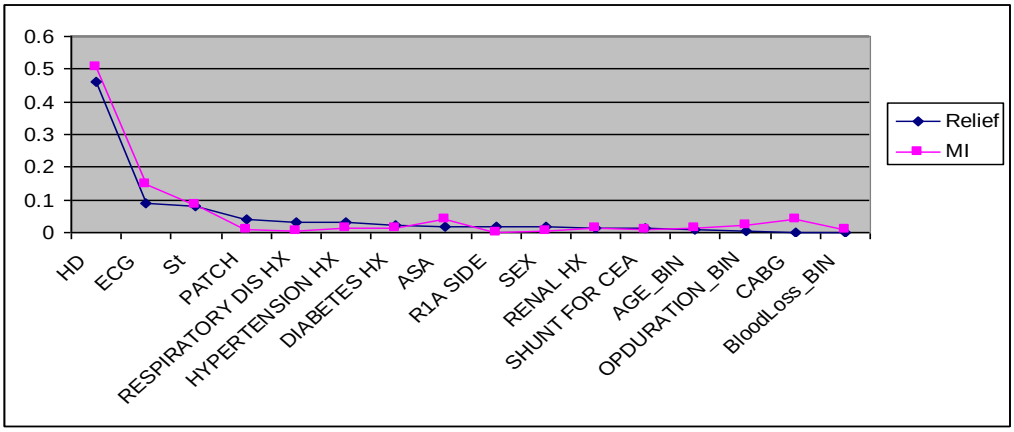

Figure 8.2: A comparison of mutual information and Relief for the CM3aD model.

Discussion

From Figure 8.2, the use of mutual information for ranking attributes produces a slightly higher value compared to the Relief algorithm. The results in Figure 8.3 show that some attributes have high rank according to the Relief whereas their ranks are low corresponding to the mutual information calculations. For example, the Relief algorithm

ranked the attribute “R1 PAT” as the first whereas the first ranked by the mutual information is the “CARDIAC_FAIL” attribute.

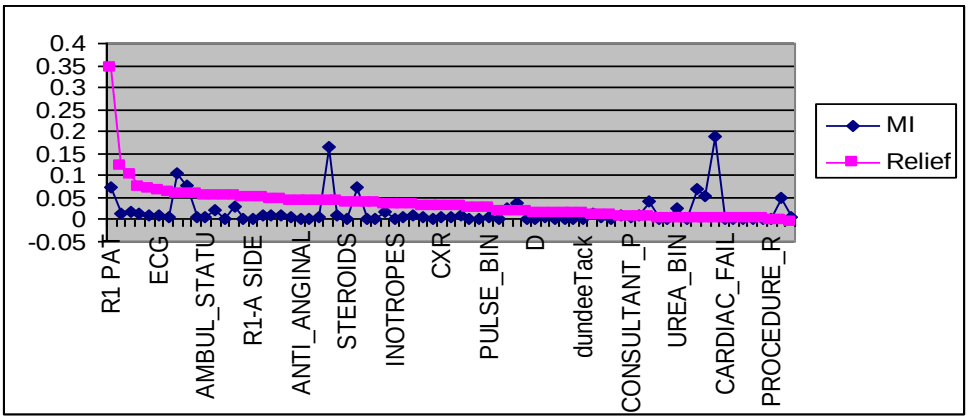

Figure 8.3: A comparison of mutual information and Relief for the CM2 model.

Overall, in Figures 8.2 and 8.3, the measurements, between each attribute in the data set (CM2, CM3aD) for the outcomes, are very similar using either Relief or mutual information algorithms. This means the rank from the use of mutual information calculations is nearly the same as the popular Relief algorithm except The advantage of using mutual information over Relief is that this algorithm can show the weight values from each attribute directly to outcome classes whereas Relief weight values are based on the distinguishing samples that are near each other in the same class. Hence, mutual information seems to be simpler to use than the Relief algorithm. Moreover, mutual information algorithm represents an interesting combination between pattern recognition concepts (a pattern is represented in the attribute dimensional space), Bayes' theory, and mutual information.

8.5. Mutual Information and Clustering

This section demonstrates the use of mutual information in the KMIX clustering algorithm. The hope is that the KMIX results can be improved by using the attribute weights (mutual information values) inside the clustering process.

8.5.1. The Weighted KMIX Algorithm (WKMIX)

The idea behind the Weighted KMIX Algorithm (WKMIX) is derived from the contributions of Huang (1997) where the weights are applied to the categorical attributes. According to Huang (1997), the choice of the weight depends on how many numeric attributes are allocated in the data domain. The weight is normally chosen as the overall average standard deviation of numeric attributes. Therefore, the weights do not clearly reflect the relationship between the data attributes and the clusters. Moreover, there always exists an influence of data attributes to the outcome risks for patients in medical domains. Therefore, the data attributes will have an influence on the clusters in the clustering process. The use of mutual information enhances the alternative significant levels of the data attributes contributing to the outcomes as shown in the previous section. It is suggested that the combination of the KMIX algorithm and the weights derived from the mutual information might improve the clustering process. This is like supervised clustering, where the attributes' contributions to the outcomes are considered during the clustering process. This algorithm can be seen as the first instance of the idea being used in medical risk prediction.

The algorithm steps are the same as the KMIX algorithm. However, the distance between each pattern to the centre vector is calculated as follows:

$$d(X_i,Q_j) = W_{iN}d^N(X_i,Q_j) + W_{iC}d^C(X_{i,}Q_j); j=1,2,..k; \qquad (8.16)$$

where W_{iN}, W_{iC} are the mutual information values of individual numerical, or categorical attributes.

In other words, Equation (8.16) can be rewritten as:

$$d(X_i,Q_j) = MI_{iN}d^N(X_i,Q_j) + MI_{iC}d^C(X_{i,}Q_j); j=1,2,..k; \qquad (8.17)$$

The detail steps of WKMIX algorithm are rewritten as:

Step 1: Initial K clusters according to K partitions of data set.

Step 2: Update K centre vectors in the new data set (for the first time the centre vectors are calculated)

$Q_j = (q^N_{j1}, q^N_{j2}, \ldots, q^N_{jp,} q^C_{jp+1,} \ldots, q^C_{jm}),\ j = 1, 2, \ldots, k$

where $\{q^N_{ji}\}_{i=1,2..p} = \{mean^N_{ji}\}$ (mean of i^{th} attribute in cluster j);

and $\{q^C_{ji}\}_{i=p+1,..m} = \{mode^C_{ji}\}$ (max freq value in attribute i^{th} in cluster j).

Step 3: Update clusters as the following tasks:

Calculate the distance between X_i in i^{th} cluster to K centre vectors:

$d(X_i,Q_j) = MI_{iN}d^N(X_i,Q_j) + MI_{iC}d^C(X_{i,}Q_j); j=1,2,..k;$

Allocate X_i into the nearest cluster such that $d(X_i,Q_j)$ is minimum.

Repeat for the whole data set, and save them to the new data set with K new centre vectors.

Step 4: Repeat step 2 and 3 until no more change in the distance between X_i and new K centre vectors.

The computational cost of this algorithm is O(TmnK), where T is the number of iterations of the reallocation process; m is number of data attributes; n is the number of objects; and K the number of clusters. Like the K-means algorithm, this algorithm produces potentially problematic locally optimal solutions. To deal with this, techniques such as genetic algorithm can be applied to produce globally optimal solutions. Optimization via genetic algorithms is not discussed further in this thesis.

The WKMIX algorithm is KMIX with a small change in the 3rd step in the algorithm, where mutual information is used to weight the data attributes. Hence, WKMIX might be appropriate for medical data domains, in particular the thesis data domain, where the mutual information between the attributes and the prediction risks reflects important dependencies. The next section will show the demonstration on this algorithm in a case study with the thesis data domain.

8.5.2. Case Study VI

This section demonstrates the use of the WKMIX algorithm for the thesis data. The results are compared to the KMIX algorithm results.

Data

The experimental data is model CM3aD with 16 input attributes and 341 cases. The preparation of the data such as cleaning, transformation and so on is the same as in Case Study V above (in section 8.4).

Method

The mutual information values are used as the attributed weights for the WKMIX algorithm. The clustering results are compared to the KMIX results (taken from Table 6.16 in Chapter 6). Alternative clustering models are then built for the use of neural network techniques and the J48 decision tree algorithm. The classification results will be discussed using the standard measures.

Result and Discussion

The results can be seen in Table 8.2. Note that the outcome clusters of "C2H" and "C1L" are regarded as "High risk" and "Low risk" respectively.

From Table 8.2, the WKMIX sensitivity is a little poorer than the KMIX sensitivity (0.53 and 0.68) whereas the prior positive predictive value is nearly the same (0.22 and 0.20). However, the WKMIX achieves a considerable improvement in accuracy compared to the KMIX algorithm (0.61 and 0.48 respectively). Therefore, the WKMIX achieves higher performance compared to the KMIX in overall. Importantly, the WKMIX gap between sensitivity and positive predictive value is a little smaller compared to the KMIX (about 0.30 and 0.48 respectively).

Algorithms	Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV
WKMIX	High risk	30	27	0.61	0.53	0.63	0.22	0.87
	Low risk	105	179					
KMIX	High risk	39	18	0.48	0.68	0.44	0.20	0.88
	Low risk	158	126					

Table 8.2: The results of alternative weights for CM3a model.

Like the framework of the Case Study IV for discovering the data structure, the clustering model CM3aDC is generated with the outcomes derived from the WKMIX results. This new model is then used with the same neural network techniques as in the Case Study IV.

For example, the multilayer perceptron with a topology of 16-0-1 (16 input nodes; 0 hidden node; and 1 output - 2 class nodes) is used; its learning rate is 0.3; 100 training cycles; and ten-fold cross-validation (see detail in Table 8.3). The use of the J48 decision tree technique is also applied to this new model.

Classifiers	Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3aDC-MLP (16-0-1; 0.3;100 epochs)	C2H	135	0	1	1	1	1	1	0
	C1L	0	206						
CM3aDC-RBF (c=1)	C2H	132	3	0.97	0.97	0.98	0.99	0.96	0.02
	C1L	6	200						
CM3aDC-SVM (poly; p=1)	C2H	135	0	1	1	1	1	1	0
	C1L	0	206						
CM3aDC-J48 (binary tree)	C2H	135	0	1	1	1	1	1	0
	C1L	0	206						

Table 8.3: The results of CM3aDC used alternative techniques.

Table 8.3 shows that all classifiers achieve nearly ideal results over all standard rates. For example, at the confusion matrix of the poorest classifier CM3aDC-RBF, there are only 3 mis-classified patterns ("C2H") in the "C1L" class whereas there are 6 mis-classified "C1L" in the "C2H" class. Importantly, there is no distance between the sensitivity versus positive predictive value, and the specificity versus negative predictive value except the very negligible gap (0.02) of the classifier CM3aDC-RBF. Again, these results demonstrate that pattern recognition techniques (neural networks and decision tree of J48) can replicate the clustering results derived from the WKMIX outcomes. Therefore, as discussed in Chapter 7, the nature of the problem and the difficulty of measuring influential parameters again influence the clustering WKMIX results in Table 8.2.

8.6. Discussion

Figure 8.2 and Figure 8.3 show that the mutual information performance is very similar to the popular Relief algorithm. However, it is simpler to use than the Relief algorithm.

Adding attribute weights to the clustering algorithm does not change the computational cost of the algorithm compared to the KMIX. The WKMIX results (Table 8.2) clearly show an improvement in accuracy compared to KMIX. Furthermore, the high performances (ideal rates) in Table 8.3 show the supervised techniques can perfectly replicate the outcomes of the WKMIX clustering models. This means the data is formed into well defined clusters by the WKMIX algorithm. This suggests that by adding weights in the clustering process, the quality of the algorithm is improved.

8.7. Summary

This chapter presented the concept of feature selection. A simple (filter) feature selection method is applied to the thesis data to measure the relevant level of attribute set to the outcomes. The improved WKMIX algorithm provides an investigation of using attribute weights (mutual information calculations) in the clustering process. The Case Study VI results proved this by showing the increasing performance compared to the KMIX.

Chapter 9

Conclusions and Further Research

9.1. Introduction

This chapter presents the conclusions for each research question initially presented in Chapter 1. The main contributions of the thesis are shown for the theory and practice in the use of pattern recognition and data mining techniques for generating risk prediction models. The limitations of the research are discussed in order to propose directions for further works.

9.2. Concluding Remarks

Chapter 1 introduced the reasons for the use of pattern recognition and data mining techniques in the cardiovascular domain. From this, a list of questions was generated (Section 1.3 in Chapter 1) to outline specific goals for this research. This section discusses the research findings in the light of these questions.

9.2.1. How Able Are The Existing Systems In Dealing With Risk Prediction For Patients?

It is not to deny that the POSSUM and PPOSSUM systems have a dramatic impact on risk prediction in the cardiovascular domain for the morbidity and mortality of patients. From the reduction to 12 physiological factors and 6 operative and postoperative factors (Copeland et al, 1991), for each patient, the mortality and morbidity rates are easily

calculated using Equations (2.1), (2.2), and (2.3) in Chapter 2. However, these systems do not consider all possible attributes. The POSSUM equations concentrate on the attributes defined by Copeland et al (1991) to calculate the physiological scores and the operative severity scores whereas other attributes might contribute to the prediction of cardiovascular risk for patients. For example, according to Kuhan et al (2001), the list of attributes in Table 6.1 in Chapter 6 is seen as significant in risk prediction for cardiovascular patients. Some of these are not used in POSSUM calculations, but they are correlated to the heuristic outcomes with the thesis models via the mutual information evaluations. For example, "Heart disease" attribute is significant (Kuhan et al, 2001), and it is not being used in POSSUM calculations. This attribute is very much correlated to the models CM3aD outcomes (its value (HD) is 0.51- see in Figure 8.2 in Chapter 8) according to mutual information evaluation. Hence, this attribute can be seen as a significant indicator in the prediction of patient risk. This is to be expected as the POSSUM is a generic medical risk prediction system. Moreover, in POSSUM and PPOSSUM and other logistic regression systems, the transformed categorical risks, from the numerical risk threshold, might lead to an ambiguous interpretation for patient outcome. For example, the analysis of an example from the use of the INDANA system (Pocock et al, 2001) can be seen as the representative of logistic regression system. At this point, although the POSSUM and PPOSSUM risk assessment systems present some advantages in the risk prediction for patients, they are not totally satisfactory for use according to the needs of the clinicians in this data domain.

9.2.2. Are Linear Models Adequate For Use With The Data Domain?

Linear models are usually a first choice for classification problems. However, they are not adequate for use with the cardiovascular data, because of the limitations indicated in section 3.4 in Chapter 3. For example, linear models provide poor prediction for data outliers, and struggle to deal with noisy data whereas these data characteristics are typical in the thesis data domain (see in section 5.3.1 in Chapter 5). The use of linear models might lead to poor classification performance as they just use linear boundaries to separate the patterns in the data space. However, the thesis data is difficult to separate clearly in to alternative classes as their patterns are non-linear distributions in data space. For example, Figure 6.5 in the Case Study III in Chapter 6 showed that the patient risks ("High risk", "Medium risk", and "Low risk") are very hard to separate by linear boundaries in the map. Therefore, nonlinear models might be the better choice for use with the cardiovascular data used in the thesis.

9.2.3. What Are The Different Ways To Classify The Data?

Four pattern recognition and data mining techniques, template matching; statistical classification; syntactic or structural matching; and neural network, can be used separately or in combination to solve classification problems. For example, according to Tsai and Fu (1980), the neural network approach sometimes might be seen as an implementation derived from statistical pattern recognition and syntactic pattern recognition approaches.

In this thesis, two main approaches to pattern recognition and data mining, supervised (multilayer perceptrons; radial basis functions; support vector machines; and J48 decision tree) and unsupervised (self organizing maps; and the KMIX/WKMIX clustering algorithm) methods are used.

Of the supervised methods, radial basis function classifiers offered the poorest results. For example, the radial basis function classifiers predicted all "High risk" patterns to belong to the "Low risk" class for all clinical risk models (CM1-RBF, CM2-RBF, CM3a-RBF, CM3b-RBF, CM4a-RBF, and CM4b-RBF) in Chapter 7 (Sen=0 in Table 7.9). The reason for this may arise from the disadvantages indicated in section 4.2.2 in Chapter 4. The decision tree technique (using the representative J48 classifier) performs as well as the neural network classifiers through the Case Study VI in Chapter 8 (see Table 8.3). However, its use is limited, because of the difficulties in finding clinically meaningful structural rules.

Although the self organizing maps offer the benefit of a visual presentation for the data, they are not suitable for use in this thesis as they require numerical data maps for input. This is not convenient for the thesis data domain where a mixture of data types is given.

KMIX and the improved WKMIX showed advantages in dealing with both numerical and categorical data via the use of Euclidean and Hamming distances. By using this clustering approach, and the resulting confirmatory classification results, the nature of the problem and the difficulty of measuring influential parameters are suspected for the earlier poor supervised classification results. For example, KMIX accuracy rates were at 0.60 and 0.45, and there is a big distances between the sensitivity and the positive predictive value (see in Tables 7.5, 7.6 in Chapter 7), for the models CM3a and CM3b. The high performance in the use of neural network techniques with clustering outcome models highlighted the nature of the problem and the difficulty of measuring influential parameters (see results in Table 7.7, 7.8). The improvement in the accuracy rate of WKMIX compared to the KMIX (see Table 8.2 in Chapter 8) might open a new direction

in the use of attribute weights in the clustering process. Generally, there is an influence of the attributes to the patient's outcomes in medical domain. For example, the *INDANA trial (Pocock et al, 2001)* indicated that smoking attribute is one of significant attributes in the predicting the patient risk (see section 2.2 in Chapter 2). Another example is the list of attributes in Table 6.2 in Chapter 6 (Kuhan et al (2001), where they are estimated as the significant attributes to predict the cardiovascular risks. Therefore, attribute weights might help the clustering process place data in the right classes.

9.2.4. Which Method Of Clustering Data Is Appropriate For This Medical Domain?

The thesis focuses on the use of the partitional clustering method, because of the advantages highlighted in the literature review in Chapter 3. The K-means algorithm (Forgey, 1965; Jancey, 1966; MacQueen, 1967; Hartigan, 1975; Hartigan and Wong, 1979) is the most popular clustering tool used in scientific and industrial applications (Berkhin, 2002). An example of K-means, KMIX, is used with the thesis data. The results (see in Table 4.5 in Chapter 4) show an improvement in the ability to deal with the mixture of numerical, categorical, and Boolean attributes in the data set, by using both Euclidean and Hamming distances to the appropriate clustering centre. The centre vector types and its measures can be listed according to attribute types as follows:

- Numerical (continuous) attributes: The centre is the average value of all attributed values. The measurement method used here is Euclidean distance.
- Categorical (discrete) attributes: The centre is the mode (maximum frequency value) in the attribute. The measurement method is Hamming distance (see in Equation (4.17) in Chapter 4).

- Boolean attributes: They can be viewed as either numerical or categorical attributes. In the thesis, they are, in most cases, treated as categorical attributes.

The KMIX algorithm is compared to the publicised and K-means results through the use of standard data sets from the UCI repository (Merz & Merphy, 1996). The better results for KMIX (see in Table 4.5 in Chapter 4) lead to confidence in its use for the thesis data.

9.2.5. Can The Attribute Set Be Decreased By Defining The Significant Attributes For Data Domain?

The use of the filtering and ranking methods based on mutual information can reduce the number of attributes for the data domain. The mutual information between the attributes and output classes can be calculated as Equation (8.15) in Chapter 8. This is based on Bayes' theorem (Bayes, 1763); the entropy (Shannon, 1948); and Kullback Leibler divergence (Cover and Thomas, 1991); given by:

$$MI(C, x_j) = \sum_{i=1}^{c} \sum_{k=1}^{s} p_{ijk} \log\left(\frac{p_{i\,jk}}{q_i . r_{jk}}\right),$$

where $p_{ijk} = \frac{sum_{ijk}}{sum}$ with sum_{ijk} is number of patterns in class $C_{i,}$ with attribute jth and in the k^{th} state, $q_i = \frac{sum_i}{sum}$, where sum_i is number of patterns belong to class $C_{i,}$ and $r_{jk} = \frac{sum_{jk}}{sum}$, where sum_{jk} is number of patterns in attribute jth and in k^{th} state.

Case Study V in Chapter 8 showed the same performance of mutual information calculations compared to the popular Relief algorithm (see comparison results in Figures 8.2 and 8.3). These slight improvements, plus its simpler calculation, lead to mutual information being chosen for use with the thesis data.

9.3. Contributions

From the academic point of view, the main contributions in this thesis can be identified as follows:

- The implementation of modeling the cardiovascular data based on clinical knowledge.
- Investigating and implementing the use of pattern recognition and data mining techniques instead of the use of other methods such as POSSUM and PPOSSUM.
- The investigation and verification of a data mining methodology for evaluating individual risk prediction in alternative risk prediction models by using alternative pattern recognition and data mining techniques.
- The improvement of K-means algorithm, as KMIX, to use alternative attribute types in the data domain.
- The definition of a calculation based on mutual information and Bayes' theorem, and its use as attribute weights in the WKMIX algorithm.

Data from both clinical sites (Hull and Dundee) is viewed using 6 clinical models based on clinical expert advice. Three other scoring risk models were built based on the Hull site data. The clinical model outcomes for individual patients are labeled via heuristic formulas (see section 6.3 in Chapter 6) whereas the scoring risk outcomes are based on

the model threshold values (see Table 6.6 in Chapter 6). Model CM2 was derived from the model CM1 with different outcome set (based on the "PATIENT STATUS" and "30D stroke/death" attributes). The CM3a and CM4a outcomes are derived from CM2 outcomes with smaller input sets. The other models, CM3b and CM4b, differ to CM3a and CM4a in an expansion of the scale for the outcomes. These alternative outcomes were used here with the hope to determine more detailed risk predictions for individual patients. However, all the above models seem to be fail, as indicated by the poor performances in the supervised classifiers and the high gaps of the sensitivity rates versus positive predictive values (for "High risk" predictions) in all thesis experiments. The suggested reason for these failures is the nature of the problem and the difficulty of measuring influential parameters for the models.

POSSUM and PPOSSUM can predict individual risk for patients via the mortality, morbidity, and death rate scores. The numeric output moving from 0 (0%) to 1 (100%) supported the patient risks being located in alternative risk bands. Categorical risks, such as "High risk", "Low risk" and so on, based on the POSSUM and PPOSSUM bands, are familiar and easily realized for individual patterns. This might help the clinicians be aware, and make more exact calls about the risk predictions for individual patients. Moreover, the categorical risks enable the use of standard measurement evaluations in order to analyse results from the use of alternative classifiers with these outcomes. However, the individual categorical risks according to this are ambiguous in their interpretations and dependent on the threshold between the numerical bands. For example, the first band, from 0% (0) to 10% (0.1) of the POSSUM and PPOSSUM results, can be seen as an implicit "Very Low risk" group (see in Tables 6.9, 6.10 in

Chapter 6). Therefore, a patient with the results of 10.5% might be belonged to this band or the next band depending on the user.

The primary research and the literature review in Chapter 3 showed clearly that the use of pattern recognition and data mining in medical areas is of wide interest. However, there is a lack in the methodology for generating and deploying the models for the cardiovascular data domain. The popular methodologies of CRISP_DM (Shearer, 2000), and SEMMA (SAS, 2008) are not suitable to use in this thesis, because they are too big and too complicated. The thesis methodology proposed here (see in Chapter 5) is based on Davis (2007). This framework satisfies data mining criteria such as the right technique choices (supervised and unsupervised pattern recognition methods), and the use of correct measurement methods (mean square error, confusion matrix and so on). It also provides a systematic approach in transforming raw data to a state where it can be reliably used with the chosen classifier techniques.

The supervised approach makes use of multilayer perceptron; radial basis function; and support vector machine classifiers whereas the unsupervised method uses self organizing maps and K-means clustering algorithms. The KMIX algorithm is shown to be an improvement over K-means, and can be applied to the alternative attribute types (categorical, Boolean and numerical) in the thesis data.

The mutual information calculations are based on the probability of the number of patterns falling to alternative classes in alternative output states. The significance of input attributes to the model outcomes are based on the mutual information values. The WKMIX algorithm was proposed based on the KMIX algorithm plus the use of mutual information values (as the attribute's weights) (see Chapter 8). However, as the nature of

the problem and the difficulty of measuring influential parameters are suspected for the poor results in the thesis experiments, this approach may need to be investigated further in future work.

The above issues apply to the cardiovascular data derived from both the Hull and Dundee clinical sites. The nature of the problem and the difficulty of measuring influential parameters caused poor performance for the classifiers though all thesis experiments. According to Haykin (1999), the performances of neural network classifiers are influenced strongly by the quality of the training data. Towards the end of the thesis research, an interesting comment was made by one of the cardiovascular clinicians. The thesis data from the Hull and Dundee sites is in fact not the raw, original data. It is derived and interpreted data, with many attributes representing clinical interpretations of numeric measurements and categorical values. Therefore, these interpretations of the original medical data (held on paper patient records), the nature of the problem and the difficulty of measuring influential parameters may be the major reasons for the poor classification results across many of the experiments presented in this thesis. The original uninterpreted data is currently not available, so further investigation of this was not possible. It is, however, a message to those applying data mining to clinical domains: make sure the original uninterpreted data is available!

9.4. Summary and Further Research

The thesis presented an investigation in the use of pattern recognition and data mining techniques to produce and verify risk prediction models in medicine, and in particular in the cardiovascular domain. Firstly, the existing risk assessment systems, and in particular POSSUM and PPOSSUM, were introduced and discussed to show the need for using

pattern recognition and data mining techniques (and in particular, neural networks) for the thesis data. A data mining methodology was proposed as the framework for use with the thesis data. Alternative supervised (neural networks) and unsupervised techniques (self organizing map and KMIX clustering algorithm) were used with the thesis data. The standard measurements such as mean square error, confusion matrix, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value are used to analyse the performance of the classification process.

Overall, the radial basis function classifiers produced worst results compared to the multilayer perceptron and support vector machine classifiers. This is not surprising, given the disadvantages shown in the literature review in Chapter 3 (section 3.6.1), and in the case studies in Chapter 6.

In Chapter 7, all neural network classifiers produced very poor sensitivity and positive predictive values (about less than 0.24 and 0.30 - see Table 7.9 in Chapter 7) over all experiments with the thesis data whereas the specificity rates and negative predictive values are high (over 0.82 and 0.83). However, in the context of the thesis, the specificity and negative predictive values provide poor information content. This is because these rates reflect the classification results for the "*Low risk*" class, which was the minor predictive outcome considered in this thesis.

The scoring risk classifiers (Mortality, Morbidity, and Death rate) produced better results in “High risk” patterns compared to the neural network classifiers, with higher rates for sensitivity and positive predictive values.

However, all classifiers (neural network and scoring risk models) show big gaps in the sensitivity versus the positive predictive value. Furthermore, the specificity rates and

negative predictive values are much higher than the sensitivity rates and positive predictive values. This suggests many distributions of "Low risk" patterns in data space; or the poor quality of the classification techniques. The comparison between neural network classification and random classification results (see Table 7.10 in Chapter 7) suggests that the former is likely.

Unsupervised techniques, such as self organizing maps and K-means algorithm (KMIX), were used to find out what pattern structure existed in the data set. Although self organizing maps can visualize data, they could not define clearly the data structure, because not all patterns can be seen in the clustering map. The KMIX algorithm was used with the thesis data due to its ability in dealing with many data types of the cardiovascular data. However, the gaps between the sensitivity and positive predictive value for KMIX are at about 0.19 and 0.71 respectively (Tables 7.5 and 7.6 in Chapter 7). The gaps between these results means the clustering outcomes are very different to the predicted expectations derived from heuristic rules. By using Euclidean and Hamming distances in risk prediction groups (see Tables 7.13, and 7.14 in Chapter 7), the results suggested that the data did not support the outcome labeling (for the clinical risk models). Therefore, it is suggested that the nature of the problem and the difficulty of measuring influential parameters are the main reasons for the poor performance for all classifiers.

Although the thesis classifiers produced poor results, the thesis methodology can be seen as a viable strategy for the practical implementation for a complete decision support tool in the cardiovascular data domain.

The results, techniques and methods developed in this research could not be used in clinical diagnosis, without trials that have ethical clearance. However, the thesis has

demonstrated that the theory, from the data mining and pattern recognition research area, might be applied in to this medical domain with good results, given reliable data. There needs to be further collaborative research, between clinicians and computer scientists, to verify the interpretation of cardiovascular data, and to build a more complete decision support tool in the cardiovascular area.

Furthermore, the strategy of combining clinical knowledge and the data mining techniques used in the thesis framework via fuzzy logic theory might produce a realistic tool for the risk prediction process (Warren et al, 2000). The combinations of neural networks and fuzzy systems are popular (Nauck et al, 1993; Nauck, 1994; Kosko and Burgess, 1998; Gegov et al, 2007). According to Wang and Mendel (1992), fuzzy systems are capable of approximating any nonlinear function over a compact set to arbitrary accuracy. Hence, the design methods for fuzzy systems are developed which determine fuzzy systems based on desired input-output pairs and fuzzy IF- THEN rules from human experts. Wang and Mendel (1992) proved that the performance of these fuzzy identifiers is much better than the neural network identifiers. Also according to Kochurani et al (2007), the utility of fuzzy systems lies in their ability for modeling uncertain or ambiguous, multi-parameter data often encountered in complex situations like medical diagnosis. Kochurani et al (2007) proposes a new model, a combination between the rule structure obtained from decision tree and TSK fuzzy model (Takagi and Sugeno, 1985; Sugeno and Kang, 1988), to predicting the risk for medical decision making situations.

For the problem indicated in the thesis, first of all, the heuristic labeling for the outcome need to be further investigated. For example, the expected outcomes produced from the

clustering WKMIX might be the reliable risks for patients. Based on the given ideas from the experts about the significant attribute set, and the status of patients in the given data domain, an estimation of the attributes weights via mutual information evaluations for the data attributes is obtained. The clustering WKMIX used with the attributes weights might produce a reliable outcome set for the data domain. However, further collaborations between clinicians and computer scientists are needed to verify this.

As indicated in this thesis the use of multilayer perceptron classifiers produced the better results compared to other neural network classifiers. Moreover, the combination between supervised and unsupervised classifiers might ensure the improvement of correctly prediction for patients in particular "High risk" ones (see Table 7.17 in Chapter 7). Therefore, by using neuro-fuzzy classification techniques (Wang and Mendel, 1992; Nauck and Kruse, 1995) building on the combination of supervised classifiers such as multilayer perceptron and the clustering classifier such as WKMIX (see in Chapter 8), where attribute weights are used, the resultant classifiers might enhance the thesis results. All of these suggestions are left for the further research.

Bibliography

Aikins, J.S., Kunz, J.C., Shortliffe, E.H., Fallat, R.J. (1983). PUFF: An Expert System for Interpretation of Pulmonary Function data. Computer Biomedicine Research. Vol.16 (3). Pp:199-208.

Alhoniemi, E., Himberg, J., Parhankangas, J., Vesanto, J. (2005). SOM tool box. Available online at website: http://www.cis.hut.fi/projects/somtoolbox/.

Altman, D.G., Bland JM (1994a). Diagnostic Tests 1: Predictive Values. British Medical Journal. Vol. 308.

Altman, D.G., Bland JM (1994b). Diagnostic Tests 2: Predictive Values. British Medical Journal. Vol. 309.

Anderson, J. A. (1982). Logistic Discrimination. In Handbook of Statistics, Classification, Pattern Recognition and Reduction of Dimensionality (Eds P.R. Krishnaiah, and L.N. Kanal). Vol .2. Pp. 169–91. North-Holland, Amsterdam.

Anzai, Y. (1992). Pattern Recognition and Machine Learning. Academic Press Inc. London.

Assmann, G., Cullen, P., and Schulte, H. (2002). Simple Scoring Scheme for Calculating The Risk of Acute Coronary Events Based on The 10-year follow-up of Prospective Cardiovascular Münster (PROCAM) Study. Circulation. Vol. 105. (3). Pp: 310-315.

Balakin, K.V, Ivanekove, Y.A., Savchuk, N.P., Ekins, S. (2005). Comprehensive Computational Assessment of ADME Properties Using Mapping Techniques. Current Drug Discovery Technology. Vol 2(2). Pp: 99-113.

Barach, P. and Small, P.D. (2000). Reporting and Preventing Medical Mishaps: Lessons from Non-medical Near-miss Reporting Systems. British Medical Journal. Vol. 320. Pp. 759-763.

Battiti, R (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transaction on Neural network. Vol.5, No. 4.

Baxt, W. G. (1991). Use of an Artificial Neural Network for The Diagnosis of Myocardial Infarction. Annual Internal Medicine, Vol. 115. Pp: 843-848.

Baxt, W. G. (1992). Use of an Artificial Neural Network for Data Analysis in Clinical Decision Making: The Diagnosis of Acute Coronary Occlusion. Neural Computing, Vol. 2.

Baxt, W.G., Skora, J. (1996). Prospective Validation of Artificial Neural Network Trained to Identify Acute Myocardial Infarction. Lancet. Vol. 347. Pp:12-5.

Bayes, T. (1763), "An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.", Philosophical Transactions, Giving Some Account of the Present Undertakings, Studies and Labours of the Ingenious in Many Considerable Parts of the World. Pp: 370–418

Bennet, M. (2006). A Risk Score for Cardiovascular Disease. Available online at website: http://www.riskscore.org.uk/.

Berkhin, P. (2002). Summary of Clustering Data Mining Techniques, Springer Berlin Heidelberg. Pp: 25-71.

Bigi, R., Gregori, D., Cortigiani, L., Desideri, A., Chiarotto, F.A., Toffolo, G.M. (2005). Artificial Neural Networks and Robust Bayesian Classifiers for Risk Stratification Following Uncomplicated Myocardial Infarction. International Journal of Cardiology. Vol.101.(3). Pp: 481-487.

Bishop, C. (1995). Neural Network for Pattern Recognition. Oxford University Press.

Boser, B., Guyon, I., and Vapnik, V.N. (1992). S Training Algorithm for Optimal Margin Classifiers. Fifth Annual Workshop on Computational Learning Theory. Pp: 144-152.

Bramer, M. (2007). Principle of Data Mining. Springer Press.

Breiman, L., Freidman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

Cacciafesta, M., Campana, F., Piccirillo, G., Cicconetti, P., Trani, I., Leonetti-Luparini, R., Marigliano, V., Verico, P. (2001). Neural Network Analysis in Predicting 2-year Survival in Elderly People: A New Statistical– Mathematical Approach. Gerontology Geriatrics.Vol. 32. Pp: 35–44.

Cacciafesta, M., Campana, F., Trani, I., Annichiarico, R., Leonetti-Luparini, R., Gianni, W., Cicconetti, P., Piccirillo, G. (2000). A Neural Network Study of The

Correlations between Metabolic-Cardiovascular Diseases and Disability in Elderly People. Gerontology. Geriatrics. Vol. 31.Pp. 257–266.

Chan, I., Wells, W., Mulkern, R.V, Haker, S., Zhang. J., Zou, K.H., Maier, S.E., (2003). Tempany CM. Detection of Prostate Cancer by Integration of Line-Scan Diffusion, T2-mapping and T2-weighted Magnetic Resonance Imaging; a Multichannel Statistical Classifier. Medical Physics.Vol. 30 (9). Pp:2390-8.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). The CRSIP – DM Process Model. Available online at website: http://www.industrieldatamining.dk/images/crisp-dm.pdf.

Cheung, Y-M. (2003). "*k**-means: A New Generalized *k*-means Clustering Algorithm," Pattern Recognition Letters," Vol. 24. Pp: 2883-2893.

Coiera, E. (2003). Guide to Heath Informatics (2nd). Chapter 25. Available online at website: http://www.coiera.com/.

Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., Jaulent, M. C. (2000). Models to Predict Cardiovascular Risk: Comparison of CART, Multilayer perceptron and Logistic Regression. Proceeding of AMIA Symposium. Pp:156-60.

Copeland G P, Jones D, Walters M. (1991). POSSUM: A Scoring System for Surgical Audit. British Journal Surgery. Vol. 78. Pp: 355-360.

Copeland G P. (2002). The POSSUM System of Surgical Audit. Archives of Surgery;
Vol.13. Pp: 15-19.

Cortes, C., and Vapnik, V.N. (1995). Support Vector Networks. Machine Leanring, Vol.20. Pp: 273-297.

COSEHC (2003). Consortium for Southeastern Hypertension Control. Available online at website: http://www.cosehc.org/about.asp.

Courant, R., and Hilbert, D. (1970). Methods of Mathematical Physics. Vols. I and II**,** Wiley Interscience, New York.

Cover, T. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities and Applications in Pattern Recognition. IEEE Transaction Electronic. Computing. Vol.14. Pp. 326–334.

Cover, T., & Thomas, J. (1991). The Elements of Information Theory. NewYork: Plenum Press.

Cox, D.R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34. (2). Pp: 187-220.

Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines. Cambridge Press.

Cross, S.S., Harrison, R.F., Kennedy, R.L. (1995). Introduction to Neural Networks. Lancet 346. Pp. 1075–1079.

Dash, M., and Liu, H. (1997). Feature Selection for Classification. Intelligent Data Analysis. Vol. 1. No.3.

Davies, D. L. & Bouldin, D. W. (1979) A Cluster Separation Measure. IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. PAMI-1 (2). Pp. 224-227.

Davis, D. N. (2007). Data Mining and Decision Systems. Lecture notes. Available online at website: http://intra.net.dcs.hull.ac.uk.

Dunham, M. H. (2002). Data Mining Introductory and Advance Topics. Prentice Hall Pearson Education.

Ellenius, J., Groth, T., Lindah, B., and Wallentin, L. (1997). Early Assessment of Patients with Suspected Acute Myocardial Infarction by Biochemical Monitoring and Neural Network Analysis Clinical Chemistry. Vol 43:10. Pp.1919–1925.

Everitt, B. S. (1994). Cluster Analysis. (3rd Edition), John Wiley & Son, New York.

Fausett, L. (1994). Fundamentals of Neuron Networks: Architectures. Algorithms and Applications, Prentice-Hall, Englewood Cliffs.

Fielding, A.H. (2007). Cluster and Classification Techniques for the Biosciences. Cambridge University Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics. Vol. 7. No. 2. Pp: 179-188.

Forgey, E.W. (1965). Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications. Biometric Society Meetings. Riverside, California. (Abstract in Biometrics, Vol.21(3)).

Framingham Heart Study. (1948). Framingham Heart Study. Available online at website: http://www.framinghamheartstudy.org/.

Fu, K.S. (1982). Syntactic Pattern Recognition and Application. Prentice- Hall, Englewood Cliffs.

Gegov, A., Lotfi, A., Garibaldi, J., Angelov, P., Kaymak, U. (2007). Fuzzy Rule Base Networks - Overview and Perspectives. IEEE International Conference on Fuzzy Systems. Available at: http://www.fuzzieee2007.org/panels.

Gonzalez , R. C. and Thomason, M. G. (1978). Syntactic Pattern Recognition: An Introduction, Addison-Wesley Publishing Company, Massachusetts.

Gorman, R. P. and Sejnowski, T. J.. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks. Vol. 1. Pp: 75-89.

Gower, J. C. (1985). Measure of Similarity, Dissimilarity and Distance. In Encyclopedia of Statistical Sciences, Vol. 5. John Wiley & Son, New York.

Graupe, D. (2007). The Principle of Artificial Neural Networks. Second Edition. World Scientific Publishing Company.

Gueli, N., Piccirillo, G., Troisi, G., Cicconetti, P., Meloni, F., Ettorre, M., Verico, P., D'Arcangelo, E., Cacciafesta, E. (2005). The Influence of Lifestyle on Cardiovascular Risk Factors Analysis Using a Neural Network. Gerontology and Geriatrics. Vol. 40. Pp: 157–172.

Gupta, A., Wroe, C., Mi, H., Asher, J., Gok, M.A., Shenton, B.K., Ward, M., and Talbot, D. (2005). Cardiovascular Risk Assessment Scoring System for the Determination of Cardiovascular Mortality in Renal Transplant Patients. Elsevier Inc. Transplantation Proceedings, 37. Pp: 3290-3291.

Han, J. and Kamber, M., (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques. (2nd Edition). Morgan Kaufmann Publishers.

Hand, D.J., Mannila, H., Smyth, P. (2001). Principles of Data Mining. (Eds.), MIT Press, Cambridge.

Harrison R. F., Marshall, S. J., and Kennedy, R. L. (1991). The Early Diagnosis of Heart Attacks: A Neuro-computational Approach. Proceeding of the international Joint Confference on Neural Networks, Seatle. Vol 1. Pp: 1-5.

Hartigan, J.A. (1975). Clustering Algorithms, Wiley & Sons, New York.

Hartigan, J.A., and Wong, M. (1979). Algorithm AS136: A *k*-means Clustering Algorithm, Applied Statistics, Vol. 28. Pp: 100-108.

Hawkins, R.G., Houston, M.C., Ferrario, C.M., Moore, M.A., and Bestermann, W.H. (2005). A Second Generation Approach to Cardiovascular Risk Assessment: The COSEHC Cardiovascular Risk Assessment Tool. Poster in American Journal Hypertension. Vol. 18. No. 5, PART 2. Pp: 144A.

Haykin, S. (1994). Neural networks: A Comprehensive Foundation, Macmillan College Publishing Company, Inc.

Haykin, S. (1999). Neural networks: A Comprehensive Foundation, Macmillan College Publishing Company, Inc.

Hebb, D.O. (1949). The Organization of Behaviour. New York: Wiley.

Heden B., Ohlsson M., Rittner R. Pahlm O., Haisty W. K., Peterson C., and Edenbranst L. (1996). Agreement Between Artificial Neural Networks and Human Expert for the Electrocardiographic Diagnosis of Healed Myocardial Infarction. Journal of the American College of Cardiology. Vol. 28. Pp: 1012-1016.

Huang Z. (1997) A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.

Huang Z. (1998): Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining Knowledge Discovery. Vol. 2. (2). Pp: 283–304.

Huang, J., Cai, Y., Xu, X. (2006). A Wrapper for Feature Selection Based on Mutual Information. ICPR 2006 International Conference on Pattern Recognition. HongKong.

Jain A.K., Duin R.P.W. & Mao J. (2000) Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 22, No. 1. Available online at website: http://www.mts.jhu.edu/~priebe/COURSES/FALL2003/550.730/jdm00.pdf.

Jain, A. K. & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall Advanced Reference Series. Prentice-Hall, Inc., Upper Saddle River, NJ

Jain, A. K. (1999). Data Clustering: A Review. Association Computing Machinery (ACM) Computing Surveys. Vol. 31. (3).

Jain, A.K., Duin, R. P.W., and Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. 22. (1).

Jancey, R.C. (1966). Multidimensional Group Analysis. Australian Journal. Botany, Vol. 14. Pp: 127-130.

Jeff , S. (1987). Concept Acquisition Through Representational Adjustment. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA.

Jensen, S. (2001). Mining Medical Data for Predictive and Sequential patterns. Available online at web site: http://lisp.vse.cz/challenge/pkdd2001/jensen.pdf.

Jorgensen, J.S., Pedersen, S.M., and Pedersen J.B. (1996). Use of Neural Networks to Diagnose Acute Myocardial Infarction. I. Methodology. Clinical Chemistry. Vol 42. Pp: 604-12.

Kamruzzaman, J., and Begg, R. R. (2006). Support Vector Machines and Other Pattern Recognition Approaches to the Diagnosis of Cerebral Palsy Gait. Biomedical Engineering. Vol. 53. Pp: 2479-2490.

Kang, S., Seon, SK., Yang, YJ., Lee, A., Bae, JM. (2006). Estimation of a Nationwide Statistics of Hernia Operation Applying Data Mining Technique to the National Health Insurance Database. Journal Prevention Public Health. Vol. 39. (5). Pp: 433-7.

Kaski, S. (1997). Example of Application of the Self Organzing Maps. World Poverty Map. Available online at website: *http://www.cis.hut.fi/research/som-research/worldmap.html*

Kaufman, L. & Rousseeuw, P.J. (1990). Finding Groups in Data-An Introduction to Cluster Analysis. Wiley and Son Press.

KDNuggets,(2007). Available online at website: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

Kira, K., and Rendell, L.A. (1992). The Feature Selection Problem: Traditional ,ethods and a New Algorithm. Proceedings of the 9th National Conference on Artificial Intelligence. Pp: 129-134.

Kiviluoto, K. (1996). Topology Preservation in Self-Organizing Maps. In Proceeding of International Conference on Neural Networks (ICNN'96). Vol.1. Pp: 294 299. New York.

Knuiman, M.W., Vu, H.T., and Bartholomew, H. C. (1998). Multivariate Risk Estimation for Coronary Heart Disease: the Busselton health study. Australian and New Zealand Journal of Public Health. Vol. 22. Pp: 747-753.

Kochurani, O.G., Aji, S., Kaimal, M.R. (2007). A Neuro Fuzzy Decision Tree Model for Predicting the Risk in Coronary Artery Disease. Intelligent Control, ISIC 2007. IEEE 22nd International Symposium. Pp:166 - 171.

Kohonen, T. (1981). Self-organized Formation of Generalized Topological Maps of Observations in a Physical System. Report TKK-F-A450, Helsinki University of Technology, Espoo, Finland.

Kohonen, T. (1990a). The Self- Organizing Map. Proceeding of the IEEE. Vol. 78. Pp: 1464-1480.

Kohonen, T. (1990b). Pattern Recognition by the Self-Organizing Map. In Proceeding of the third Italian Workshop on Parallel Architectures and Neural Networks. Pp:13–18.

Kohonen, T. (1995). Self-Organizing Maps. Springer, Berlin, Heidelberg.

Kohonen, T. (1996). Advances in the Development and Application of Self-organizing Maps. Proceeding of the 5th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'96). Pp: 3–12.

Kohonen, T.J. (2001). Self-Organizing Maps. Springer Series in Information Sciences, 3rd ed., Springer-Verlag, Berlin Heidelberg.

Kononenko, I., and Kukar, M. (2007). Machine Learning and Data Mining. Horwood Publishing Ltd.

Kosko, B., Burgess, J.C. (1998). Neural Networks and Fuzzy Systems. Journal of Acoustical Society of Americal. Volume 103, Issue 6. Pp: 3131-3131.

Kuhan G, Gardiner ED, Abidia AF, Chetter IC, Renwick PM, Johnson BF, Wilkinson AR, McCollum PT. (2001). Risk Modelling Study for Carotid Endarterectomy. British Journal Surgery. Vol. 88. (12). Pp: 1590-4.

Kuhan, G., Davis, DN., Chetter, IC, McCollum, CN., McCollum, PT. (2003). The Use of Artificial Neural Networks for Risk Prediction Following Carotid Endarterectomy. Internal Report.

Langdell, S.J., and Mason, J.C. (1998). Classifying Spinal Measurements using a Radial Basic Function Network. Artificial Neural Network in Biomedicine (2000). Springer-Velag London. Chapter 7. Pp: 93-104.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.(1989). Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation. Vol. 1. (4). Pp:541-551.

LeCun, Y., Jackel L. D., Bottou L., Brunot A., Cortes C., Denker J. S., Drucker H., Guyon I., Muller U. A., Sackinger E., Simard P. and Vapnik V., (1995). Comparison of Learning Algorithms for Handwritten Digit Recognition, International Conference on Artificial Neural Networks, Fogelman, F. and Gallinari, P. (Ed.), Paris. Pp: 53-60.

Lewis, R. J., (2000). An Introduction to Classification and Regression Tree (CART) Analysis. Available online at: www.saem.org/download/lewis1.pdf

Lin, T.H., Li, H.T., and Tsai, K.C. (2004). Implementing the Fisher's Discriminant Ratio in a *k*-means Clustering Algorithm for Feature Selection and Dataset Trimming. Journal of Chemical Information and Computer Sciences. Vol. 44(1). Pp: 76-87.

Lingras, P. and Y.Y. Yao (2002). "Time Complexity of Rough Clustering: GAs versus *k*-means," Lecture Notes in Artificial Intelligence 2475. Pp: 263-270.

Lisboa P.J.G (2002). A Review of Evidence of Health Benefit from Artificial Neural Networks in Medical Intervention. Elsevier: Neural Network. Vol. 15. Pp: 11-39.

Lisboa, P.J.G. (2004). Neural Networks in Medical Journals: Current Trends and Implications for BioPattern. Open Clinal Documents. Available online at: http://www.openclinical.org/docs/ext/ann/lisboa2004.pdf.

Liu, H., and Motoda, H. (1998). Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer Academic Publishers.

Liu, H., Liu, L., and Zhang, H. (2008). Feature Selection Using Mutual Information: An Experimental Study. Lecture Notes in Computer Science.

Lotte, F., Congedo, M., L´ecuyer, A., Lamarche, F., and Arnaldi, B. (2007). A Review of Classification Algorithm for EEG-based brain–computer interfaces. Topical Review. Journal of Neural Engineering. Vol. 4. Pp: R1-R13.

Lowe, H. J. (2003). Introduction to Clinical Informatics: History, Agenda, and Future directions. Available online at website: http://clinicalinformatics.stanford.edu/

Luan, F., Ma, W., Zhang, H., Zhang, X., Liu, M., Hu, Z., and Fan, B. (2005). Prediction of pKa for Neutral and Basic Drugs Based on Radial Basis Function Neural Networks and the Heuristic Method. Pharmaceutical Research, Springer Netherlands, Vol. 22.

MacQueen, J.B. (1967). Some Methods of Classification and Analysis of Multivariate Observations, Proceeding Symposium. Mathematic. Statistic. and Probability, 5th Berkeley, Vol.1. Pp: 281-297.

Maglaveras, N. Stamkopoulos, T. Diamantaras, K. Pappas, C. and Strintzis, M. (1998). ECG Pattern Recognition and Classification Using Nonlinear Transformations and Neural Networks: A Review, International Journal of Medical Informatics 52. Pp. 191–208.

Manning, C.D., Raghvan, P., and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. Available at website: http://informationretrieval.org/

Marckmann, G. (2001). Recommendations for the Ethical Development and Use of Medical Decision Support Systems. Medscape General Medicine. Available at website: http://www.medscape.com/viewarticle/408143.

Maschewsky-Schneider U, Greiser B. (1989). Primary Prevention of Coronary Heart Disease versus Health Promotion-a Contradiction? Annual Medicine. Vol.21. Pp:215-8.

Mathworks. (1994). Available online at web site: http://www.mathworks.com/products/matlab/.

Matsopoulos, G.K.., Asvestas, P.A., Mouravliansky, N.A., d Delibasis, K.K. (2004). Multimodal Registration of Retinal Images using Self organizing Maps. IEEE Trans Med Imaging, Vol. 23. (12). Pp: 1557-1563.

Matsui, Y., Egawa, S., Tsukayama, C., Terai, A., Kuwao, S., Baba, S., Arai, Y. (2003). Artificial Neural Network Analysis for Predicting Pathological Stage of Clinically Localized Prostate Cancer in the Japanese Population. Japanese Journal Clinical Oncology. Vol. 32. (12). Pp. 530-5.

McCulloch, W. S. and Pitts, W. H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, Vo.5. Pp:115-133.

Melle, W.V. (1979). A Domain-Independent Production-Rule System for Consultation Programs. Heuristic Programming Project, Department of Computer Science, Stanford University.

Mercer, J. (1909). Functions of Positive and Negative Type and Their Connection with the Theory of Integral Equations. Philosophical Transactions of the Royal Society. London, A 209 .Pp: 415–446.

Merz, C. J. & Merphy, P. (1996); Richard, (1990). UCI Repository of Machine Learning Database. Available online at website: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Mia, K. M., Joseph, Y.L., Georgia, D.T., Carey, E.F. Jr. (2003). Self-Organizing Map for Cluster Analysis of a Breast Cancer Database. Artificial Intelligence in Medicine. Vol. 27. Pp: 113–127.

Michalski, R.S. and Chilausky, R.L. (1980). Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soy-bean Disease Diagnosis. International Journal of Policy Analysis and Information Systems. Vol.4(2). Pp. 125-161. Soybean Databases.

Midwinter, M.J., Tytherleigh, M., Asley, S. (1999). Estimation of Mortality and Morbidity Risk in Vascular Surgery using POSUM and the Portsmouth Prediction Equation. British Journal Surgery, Vol. 86. Pp. 471-474.

Miller, R.A., Pople, H.E. Jr. and Myers, J.D. (1982). INTERNIST-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. New England Journal Medicine. Vol. 307. Pp: 468-676.

Minsky, M.L., Papert, S.A. (1969). Perceptrons, MIT Press, Cambridge.

Mirkin, B. G. (2005). Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall/CRC Press.

Mitra, S., Pal, S., K., & Mitra, P. (2002). Data Mining in Soft Computing Frame work: A Survey. IEEE transaction on Neural Networks, Vol. 13. (1).

Moody, J., and Darken, C. (1989). Fast Learning in Networks of Locally-tuned Processing Units. Neural Computation, Vol.1.Pp: 281-294.

Nauck, D. (1994). Fuzzy neuro systems: An overview. Fuzzy Systems in Computer Science, Artificial Intelligence. Pp: 91-107.

Nauck, D., and Kruse, R. (1995). NEFCLASS A Neuro-Fuzzy Approach For The Classification Of Data. Association for Computing Machinery Symposium on Applied Computing. Pp 26-28.

Nauck, D., Klawonn, F., and Kruse, R. (1993). Combining neural networks and fuzzy controller. In Fuzzy Logic in Artificial Intelligence (FLAI93). Pp: 35-46. Springer-Verlag.

New York Wiley, 1973.

NIST/SEMATECH. (2006). Engineering Statistics Handbook. NIST/SEMATECH e-Handbook of Statistical Methods, Available at website: http://www.itl.nist.gov/div898/handbook/.

Nurettin, A. (2006). A Support Vector Machine Classifier Algorithm Based on a Perturbation Method and its Application to ECG Beat Recognition Systems. Expert Systems with Applications. Vol.31. (1). Pp. 150-158.

O'Connor M.A. and Walley W.J. (2000) An Information Theoretic Self-Organising Map with Disaggregation of Output Classes. Second International Conference on Environmental Information systems, Stafford, UK.

Ohn, M. S., Van-Nam, H., and Yoshiteru, N. (2004). An Alternative Extension of the K-means Algorithm for Clustering Categorical Data. International Journal Mathematic Computer Science, Vol. 14. (2). Pp: 241-247.

Papik, K., Molnar, B., Schaefer, R., Dombovari, Z., Tulassay, Z., Feher, J. (1998). Application of Neural Networks in Medicine - a Review. Medicine Science Monitoring. Vol. 4 (3). Pp. 538-546.

Parisi, G. (1986). Asymmetric Neural Networks and the Process of Learning. J. Phis. A: Mathematic Generations. Vol. 19. Pp.675–680.

Pedersen, S.M. Jorgen, S.J. Gensen, R. and Pedersen J.B. (1996). Use of Neural Networks to Diagnose Acute Myocardial Infarction. II. A Clinical Application. Clinical Chemistry Vol. 42. Pp: 613-617.

Perlovsky, L.I. (1998). Conundrum of Combinational Complexity. IEEE Transaction on Pattern Analysis and machine Intelligence. Vol. 20. No. 6. Pp: 666-670.

Plaff, M., Weller, K., Woetzel, D., Guthke, R., Schroeder, K., Stein, G., Pohlmeier, R., Vienken, J. (2004). Prediction of Cardiovascular Risk in Hemodialysis Patients by Data Mining. Methods of Information in Medicine. Vol.43. Pp: 106-113.

Pocock, S. J., McCormark, V., Gueyffier, F., Bouttie, F., Fagard, R.H., Boissel, J. P. (2001). A Score for Predicting Risk of Death from Cardiovascular Disease in Adults with Raised Blood Pressure, Based on Individual Patient Data from Randomised Controlled Trials. British Medical Journal. Vol. 323. Pp: 75-81.

Prytherch, D.R., Whiteley, M. S., Higgins, B., Weaver, P.C., Prout, W. G., Powell, S. J. (1998). POSSUM and Portsmouth POSSUM for Predicting Mortality. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. British Journal Surgery. Vol 85. Pp: 1217-20.

Pyle, D. (1999). Data Preparation for Data Mining. Morgan Kaufmann Publishers.

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning. Vol.1. Pp: 81-106.

Quinlan, J. R. (1993). C4.5: Programs For Machine Learning. Morgan Kaufmann, Los Altos.

Rabiner, L., and Juang, B. H. (1993). Fundamental of Speech Recognition. Prentice-Hall. Englewood Cliffs.

Reason, J. (2000). Human error: Models and Management. Bristish Medical Journal. Vol. 320. Pp. 768-770.

Richard, (1990). In Merz & Merphy (1996). UCI Repository of Machine Learning Database. Available online at website: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Ripley, B., D. (1996), Pattern Recognition and Neural Network. Cambridge University Press.

Rosenblatt, F. (1958). The perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review. Vol. 65. Pp: 368-408.

Rosenblatt, F. (1962). Principles of Neurodynamics, Spartan Books, Washington DC.

Sanchez, L., Suarez, M. R., Villar, J.R., and Couso, I. (2008). Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data. International Journal of Approximate Reasoning.

SAS, (2008). Available online at web site: http://www.sas.com/technologies/analytics/datamining/miner/semma.html.

Schaffernicht, E., Moller, C., Debes, K., and Gross, H-M. (2009). Forward Feature Selection Using Residual Mutual Information. ESANN 2009. Bruges, Belgium.

Schalkoff, R..J. (1992). Pattern Recognition Statistical, Structural and Neural Approaches, Clemson University, John Wiley & Sons, Inc.

Serretti, A., & Smeraldi, E. (2004). Neural Network Analysis in Pharmacogenetics of Mood Disorders. Biological Medical Central, Medical Genetics. Vol. 5 (27).

Shannon, C. E. (1948). A Mathematical Theory of Communication, Bell System Technical Journal. Vol. 27. Pp. 379-423.

Shearer, C. (2000). The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data warehousing. Vol.5. Pp:13-22.

Shehroz S. K. , and Shri K. (2007). Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation. 20th International Joint Conference on Aftificial Intelligence (IJCAI-07), India.

Shortliffe E. H. (1976). Computer Base Medical Consultations: MYCIN. New York. Elsevier.

Shortliffe, E. H., Perrault, L., Wiederhold, G., and Fagan, L. M. (1990). Medical Informatics. Computer Applications in Health Care. Addison-Wesley.

Simelius, K., Stenroos, M., Reinhardt, L., Nenonen, J., Tierala, I., Makijarvi, M., Toivonen, L., Katila, T. (2003). Spatiotemporal Characterization of Paced Cardiac Activation with Body Surface Potential Mapping and Self-Organizing Maps. Physiological Measurement. Vol 24(3). Pp: 805-16.

Simula, O., Alholinemi, E., . Hollmen, J., Vesanto, J. (1999). Analysis of Complex Systems using the Self-Organizing Map. In Progress in Connection-Based Information Systems. Proceedings of the International Conference on Neural Information Processing and Intellligent Information Systems, Vol. 2. Pp: 1313-1317. Springer, Singapore.

Smulders, Y.M., Spijkerman, A.M., Kostense, P.J., Bouter, L.M., Stehouwer, C.D. (2004). Old and New Scoring Systems for Assessing Cardiovascular Risks: Problems with the Validity, the Precision and the Homogenicity of the Risk Categories. Public Medicine.Vol.148. (50). Pp: 2480-4.

SOM toolbox (2000-2005): Alhoniemi, E., Himberg, J., Parhankangas, J., Vesanto, J. Available online at website: http://www.cis.hut.fi/projects/somtoolbox/.

SOM toolbox (Alhoniemi, A., Himberg, J., Parhankangas, J., and Vesanto, J.). 2000-2005. Laboratory of Information and Computer Science in the Helsinki University of Technology. Available online at website: http://www.cis.hut.fi/projects/somtoolbox/about.shtml

Sondak, N.E., and Sondak, V. K. (1989). Neural Networks and Artificial Intelligence. Pp: 241-245.

Sugeno, M., Kang, G. T. (1988). Structure identification of fuzzy model. Fuzzy Set. Syst., 28. Pp: 15–33.

Syed, Z., Guttag, J., Stultz, C. (2007). Clustering and Symbolic Analysis of Cardiovascular Signals: Discovery and Visualization of Medically Relevant Patterns in Long-Term Data Using Limited Prior Knowledge. EURASIP Journal on Advances in Signal Processing. Vol. 2007.

Takagi, T., Sugeno, M. (1985). Fuzzy identification of systems and its application to modeling and control, IEEE T. Syst. Man Cyb. SMC-15(1). Pp: 116–132.

Talavera, L. (2005). An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering. In Sixth International Symposium on Intelligent Data Analysis, IDA05. Pp. 440-451. Madrid, Spain.

Tan, P. N., Steinbach,M., Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley Companion Book Site.

Tom, F. (2006) An introduction to ROC analysis. Pattern Recognition Letters 27. Pp: 861–874.

Tou, J.T., and Gonzalez, R.C. (1974), Pattern R-ecognition Principle, Addison-Wesley Publishing Company.

Tourassi, G.D., Frederick, E.D., Markey, M.K., and Floyd, C.E. (2001). Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis. Medical Physics. Vol. 28. (12). Pp: 2394- 2402.

Tsai W.H., and Fu K.S. (1980). A Tool for Combining Syntactic and Statistical Approaches to Pattern Recognition, IEEE Transactions on system, Man and Cybernetics, Vol. SMC-10. (12). Pp: 873-885.

Tu, J.V. (1996). Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. Journal of Clinical Epidemiology. Vol.49. (11). Pp: 1225-31.

Twardy, C.R., Nicholson, A.E., Korb, K.B., and McNeil, J. (2005). Data Mining Cardiovascular Bayesian Networks. Available online at website: http://citeseer.ist.psu.edu/cache/papers/cs2/777/http:zSzzSzwww.csse.monash.edu.auzSzpublicationszSz2004zSztr-2004-165-full.pdf/twardy04data.pdf.

Ultsch, A. (1993). Self Organized Feature Maps for Monitoring and Knowledge Acquisition of a Chemical Process. Proceedingof ICANN'93, International Conference on Artificial Neural Networks. Pp: 864-867. London, UK. Springer.

Ultsch, A., and Siemon, H.P. (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. International Proceeding INNC'90. Pp: 305-308. Dordrecht, Netherlands, Kluwer.

Vapnik V., (1995). The Support Vector Method of Function Estimation, in Nonlinear Modeling: Advanced Black-Box Techniques. Pp. 55-86, Kluwer Academic Publishers.

Vapnik V.N. (1998). Statistical Learning Theory. New York: John Wiley & Sons, Inc.

Vapnik, V. N. (1999). An Overview of Statistical Learning Theory. IEEE Transaction on Neural Networks. Vol.10. Pp: 988–99.

Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. NewYork : Springer Velag.

Venables, W.N., Ripley, B.D. (1994). Modern Applied Statistics with S-Plus. Springer, New York.

Vesanto J., Alholiemi E., (2000). Clustering of the Self-Organizing Map, IEEE transactions on Neural Network. Vol 11 (3). Pp: 586-600.

Wang, K. Wang, L. Wang, D. and Xu, L. (2004). SVM Classification for Discriminating Cardiovascular Disease Patients from Non-cardiovascular Disease Controls Using Pulse Waveform Variability Analysis. Lecture Note in Computer Science, Springer Berlin.

Wang. L.-X. , Mendel, J.M. (1992). Back-propagation fuzzy system as nonlinear dynamic system identifiers**.** Fuzzy Systems. IEEE International Conference.

Ward, M. (2005). In Gupta et al (2005). Cardiovascular Risk Assessment Scoring System for the determination of Cardiovascular Mortality in Renal Transplant Patients. Elsevier Inc. Transplantation Proceedings, Vol. 37. Pp: 3290-3291.

Warren, J. Beliakov, G. Van der Zwaag, B. (2000). Fuzzy logic in clinical practice decision support systems. System Sciences, 2000. IEEE.

Watanabe S. (1985). Pattern Recognition: Human and Mechanical, Wiley, New York.

Weingart, S.N., Wilson, R. McL., Gibberd, R.W. and Harrison, B. (2000). Epidemiology of medical error. Bristish Medical Journal, Vol. 320. Pp: 747-777.

WEKA (2005). Weka Machine Learning Project in The University of Waikato. Available at: http://www.cs.waikato.ac.nz/~ml/weka/index.html.

Whitley, M.S., Pytherch, D.R., Higgins, B., Weaver, P.C., Prout, W.G. (1996). An Evaluation of the POSSUM Surgical Scoring System. British Journal Surgery.Vol. 83. (6). Pp: 812-815.

Widrow, B. and Hoff, M.E. (1960). Adaptive Switching Circuits. Neural Computing Research. Convention Record. Pp: 96-104.

Wijesinghe, L.D., Mahmood, T., Scott, D.J.A., Berridge, D.C., Kent, P.J., Kester, R.C. (1998). Comparison of POSSUM and the Portsmouth Predictor Equation for Predicting Death Following Vascular Surgery. Britis Journal Surgery. Vol.85. Pp: 209-212.

Wilson, R.M., Runciman, W.B., Gibberd, Harrison, B.T., Newby, L. and Hamilton, J.D. (1995). The Quality in Australian Healthcare Study. Medical Journal Australia, Vol. 163. Pp: 458-471.

Yang, H.H, Van Vuuren,S., Sharma, S., Hermansky, H. (2001). Relevance of Time Frequency Features for Phonetic and Speaker-Channel Classification. Speech Commun. 31. Pp. 35–50.

Ye, N. (2003). The Hand Book of Data mining. Lawrence Erlbaum Associates Publishers.

Yii, M.K., and Ng. K.J. (2002). Risk-Adjusted Surgical Audit with The POSSUM Scoring System in a Developing Country. British Journal of Surgery. Vol. 89. (1). Pp: 110-113.

Zengyou, H., Xiaofei, X., Shengchun, D. (2005). TCSOM: Clustering Transactions Using Self-Organizing Map. Neural Processing Letters. Vol. 22. Pp: 249–262. Springer 2005.

Appendix A. Data structure

A.1. Hull site

The following table summarises the data from the Hull clinical site. As can be seen there are 30 numeric, 3 discrete numeric, 23 categorical, 38 Boolean, and 4 date/time typed attributes.

Attribute name	Attribute types	Attribute name	Attribute types
UNIT_NO	Categorical	JVP	Boolean
THEATRE_SESSION_DATE	Date/Time	LEG_OEDEMA	Boolean
CONS	Discrete	PULM_OEDEMA	Boolean
DATE_OF_DEATH	Date/Time	CARDIAC_FAIL	Boolean
Combined	Categorical	HAEMOGLOBIN	Numeric
30D MR	Boolean	WCC	Numeric
30D Ipsi CVA	Boolean	PLATELETS	Numeric
CAUSE_OF_DEATH	Categorical	UREA	Numeric
PhysiolScore	Numeric	CREATININE	Numeric
OpSevScore	Numeric	SODIUM	Numeric
P-POSS(2)	Numeric	POTASSIUM	Numeric
P-POSS(1)	Numeric	GLUCOSE	Numeric
POSS	Numeric	INR	Numeric
D	Boolean	PAO2	Numeric
HD	Boolean	ECG	Categorical
St	Boolean	CXR	Categorical
CODE	Categorical	PULM_CXR	Categorical
CAROTID_DISEASE	Categorical	URGENCY	Categorical
ARRHYTHMIA	Boolean	DURATION	Numeric
ANGINA	Boolean	CONSULTANT_PRESNT	Boolean
MYOCARDIAL_INFARCT	Categorical	ASA_GRADE	Discrete
CCF	Boolean	ANAESTHETIC_TYPE	Boolean
DIABETES	Categorical	CRYSTALOID_VOL	Numeric
SEX	Boolean	COLLOIDS	Numeric
PATIENT_STATUS	Boolean	TRANSFUSION	Numeric
INDICATION	Categorical	OTHER_BLOOD	Numeric
PVD	Categorical	BLOOD_LOSS	Numeric
DATE_HISTORY	Date/Time	LOWEST_BP	Numeric
AGE	Numeric	MIN_TEMP	Numeric

HYPERTENSION	Boolean	INOTROPES	Boolean
RENAL_FAILURE	Boolean	PRIMARY_OP	Boolean
HYPERCHOLESTEROLAEMIA	Boolean	OPERATION_DESC	Categorical
ALLERGIES	Boolean	NO_PROCS	Discrete
SMOKING	Categorical	OP_SEVERITY	Categorical
PACK_YEARS	Numeric	PERI_SOILING	Boolean
RESPIRATORY	Boolean	MALIGNANCY	Boolean
AMBUL_STATUS	Categorical	LETTER_TEXT	Categorical
CABG_PLASTY	Boolean	PROCEDURE_RANK	Numeric
THROMBO_EMBOLISM	Boolean	SHUNT	Boolean
EJECT	Numeric	PATCH	Categorical
DIURETICS	Boolean	COMP_GROUP	Categorical
WARFARIN	Boolean	COMPLICATION	Categorical
DIGOXIN	Boolean	SEVERITY	Categorical
ANTIHYPERTENSIVES	Boolean	COMPLICATION_DATE	Date/Time
STEROIDS	Boolean	RESP_SYSTEM	Categorical
ANTI_ANGINAL	Boolean	GCS	Numeric
STATINS	Boolean	BUILD	Boolean
ASPIRIN	Boolean	BP	Numeric
ORTHOPNOEA	Boolean	PULSE	Numeric

A.2. Dundee site

The following table summarises the data from the Dundee clinical site. As can be seen there are 6 numeric, 1 discrete numeric, 19 categorical, 10 Boolean, and 6 date/time typed attributes.

Attribute Name	Attribute type	Attribute Name	Attribute type
ID#	Categorical	HYPERTENSION HX	Boolean
ADMISSION.DATE	Date/Time	RENAL HX	Boolean
Discharge date	Date/Time	SMOKING HX	Categorical
PROCEDURE	Categorical	PACK YRS	Numeric
DATE	Date/Time	RESPIRATORY DIS HX	Categorical
OP DURATION	Numeric	DIABETES HX	Categorical
Surgeon.name.1	Categorical	ARRHYTHMIA	Categorical
surgeon.name.2	Categorical	ANGINA	Boolean
ASA	Discrete	MYOCARDIAL INFARCT	Categorical
EBL	Numeric	CCF	Boolean
SHUNT FOR CEA	Boolean	CABG	Boolean
PATCH	Categorical	Carotid status	Categorical
R1-A SIDE	Boolean	ECG	Categorical
R1 GRAFT	Categorical	Disposal	Categorical
R1 PAT	Categorical	LAST FOLLOW-UP DATE	Date/Time
R1 LOO	Date/Time	DATE OF DEATH	Date/Time
R1 DURATION PATENT	Numeric	CAUSE OF DEATH	Categorical
Aspirin	Boolean	G/S COMPL1	Categorical
Warfarin	Boolean	I/P OP GEN COMPL	Categorical
CROSSCLAMP TIME CEA	Numeric	DATE GENCOMPL 1	Date/Time
Tack	Boolean	Complication	Categorical
AGE	Numeric		

Appendix B.

Experimental Models Overview

This appendix shows all models used in all the case studies plus Chapter 7 experiments. They can be seen in three categories as clinical risk prediction models, scoring models, and clustering models. Note that the classifier techniques are shown in the tables as multilayer perceptron (MLP); radial basis function (RBF); support vector machine (SVM); J48; self organizing map (SOM); KMIX; and WKMIX clustering algorithm.

B.1. Clinical Risk Prediction Models

The Table B1 shows the summary of all clinical models. These models outcomes are based on heuristic rules such as inferring from the "PATIENT STATUS", or "30D stroke/death" attributes.

Models	Input numbers	Pattern numbers	Risk Prediction	Using Classifiers
CM1	26	839	High risk; Low risk.	MLP; RBF;SVM
CM2	26	839	High risk; Low risk.	MLP; RBF;SVM
CM3a	16	839	High risk; Low risk.	MLP; RBF;SVM
CM3b	16	839	Very High risk; High risk; Medium risk; Low risk.	MLP; RBF;SVM
CM4a	14	839	High risk; Low risk.	MLP; RBF;SVM
CM4b	14	839	Very High risk; High risk; Medium risk; Low risk.	MLP; RBF;SVM
Hull_POSS	22	497	High risk; Low risk	MLP; RBF; SVM
CM3aD	16	341	High risk; Low risk	KMIX; WKMIX
CM3bD	16	341	High risk; Medium risk; Low risk.	SOM; KMIX

Table B1: Clinical models summary

B.2. Scoring Risk Models

Table B2 shows the summary of applying scoring risk models to the thesis data. The outcomes are based on heuristic rules derived from POSSUM and PPOSSUM threshold scores.

Models	Input numbers	Pattern numbers	Risk Prediction	Using Classifiers
Mortality	18	499	High risk; Low risk	POSSUM
Mortbidity	18	499	High risk; Low risk	POSSUM
Death rate	18	499	High risk; Low risk	PPOSSUM

Table B2: Scoring risk models summary.

B.3. Clustering Models

Table B3 shows the summary of clustering models used with the thesis data. The outcomes are derived from KMIX and WKMIX clustering algorithms.

Models	Input numbers	Pattern numbers	Risk Prediction	Using Classifiers
CM3aDC	16	341	C2H; C1L	MLP; SVM
CM3bDC	16	341	C3H; C2M; C1L	MLP; SVM
CM3aC	16	839	C2H; C1L	MLP; RBF; SVM
CM3bC	16	839	C4VH; C3H; C2M; C1L	MLP; RBF; SVM
CM3aDC	16	341	WKMIX Outcomes (C2H; C1L)	MLP; RBF; SVM; J48

Table B3: Clustering models summary.

Appendix C. Experimental Data Explanations

This appendix describes the detailed steps of data preparation and experimental explanations for all thesis case studies and the Chapter 8 experiments. These steps are fulfilled following the thesis methodology in Chapter 4 as in the systematic representations.

C.1. Case study I

- **Step 1 (Selection):** Data is taken from Hull site including 4 attributes and 498 cases (see statistical analysis in Table C1). Note that 2 over 4 attributes are the physiological score and operative severe score calculated by POSSUM and PPOSSUM.

		PS	**OS**
N	*Valid*	498	498
	Missing	0	0
Mean		20.36	14.30
Std. Error of Mean		.247	.064
Std. Deviation		5.507	1.421
Minimum		12	13
Maximum		41	23

Table C1: Statistical analyses of PS and OS score in the Hull site.

- **Step 2 (Clean/Transform/Filter)**: All data is cleaned. Therefore, this step is ignored.

- **Step 3 (POSSUM and PPOSSUM Techniques)**: Data is used with the POSSUM and PPOSSUM formulas (Equations 2.1, 2.2, and 2.3 in Chapter 2) to produce the mortality, morbidity, and death rate scores for individual patients. The results are then divided to different groups in the range from 0%-100%. For each group a predicted mean is calculated in order to calculate the number of predicted "*mortality*" or "*death rate"* (see in Tables C2, C3 below).

Range of predicted rate	Mean predicted risk of Mortality (%)	No of operations	Predicted deaths	Reported deaths	The ratio
0-10%	6.75%	274	19	29	1.53
10-20%	14.85%	148	22	28	1.27
20-30%	24.97%	44	11	8	0.73
30-40%	34.90%	11	4	3	0.75
40-50%	43.10%	16	7	7	1.00
>50%	60.85%	5	3	3	1.00
0-100%	13%	498	65	78	1.20

Table C2: Comparison of observed and predicted death from POSSUM logistic equations.

Range of predicted rate	Mean predicted risk of Mortality (%)	No of operations	Predicted deaths	Reported deaths	The ratio
0-10%	3.00%	438	13	60	4.62
10-20%	13.48%	39	5	9	1.80
20-30%	23.25%	13	3	3	1.00
30-40%	32.27%	5	2	4	2.00
40-50%	44.86%	3	1	2	2.00
>50%	58.37%	1	1	-	0.00
0-100%	5%	498	25	78	3.12

Table C3: Comparison of observed and predicted death from PPOSSUM logistic equations.

- **Step 4 (Comparison/ Evaluation):** The comparisons are fulfilled based on the ratios between the predicted and actual rates. For example, the band group of 20%-30% in Table C2 shows that, the predicted mortality is calculated based on the mean (24.97%) and the number of operation cases (44). Therefore, the ratio between the reported mortality (8) and the predicted one (11) is 0.73.

C.2. Case study II

Clinical Model CM3aD

- **Step 1 (Selection):** The data is taken from the Dundee site with a selection of 18 attributes (16 input attributes, and 2 attributes are for the outcome calculations) and 341 patients.

- **Step 2 (Clean/Transform/Filter)**: The method used here is followed the methods indicated in "Data Preparation Strategy" section in Chapter 5. The detail as follows:

- **Cleaning task:** The summary for this task can be seen in Table C4 below. For example, the number of missing values for attribute namely *"OP DURATION"* is 72. The values are in the range of [0.7, 3]. Therefore, missing values will be filled by the mean of non-missing values (1.50). The number of missing values is 10 in the attribute *"PATCH",* and the most frequency value of *"PTFE"* is 170. Therefore, attribute missing values will be filled as "*PTFE"*.

Attribute Name	Number of Missing	Attribute values	Mean/Max freq values
OP DURATION	72	[0.7,3]	1.50
ASA	37	[1,4]	2.45
EBL	243	[0,2000]	214.18
SHUNT FOR CEA	5	Yes/No	213 (No)
PATCH	10	None/Other/PTFE/Vein	170 (PTFE)
R1-A SIDE	0	Left/Right	
AGE	0	[42,86]	68
Sex	0	Male/Female	215(Male)
HYPERTENSION HX	6	None/Yes	176(None)
RENAL HX	6	Normal/Abnormal	324 (Normal)
RESPIRATORY DIS HX	14	Normal/Mild COAD/Mod COAD/sev COAD	280 (Normal)
CABG	8	No/Yes	314 (no)
ECG	16	Normal/A-Fib<90/Other	233 (normal)
HD	0	Y/N	255(N)
DIABETES	0	Diet Rx/IGT/Insulin(NIDDM)/Insulin(NIDDM)/none	306(normal)
St	0	Y/N	323(N)

Table C4: CM3aD data structure, and its summary.

- **Transformation task**: All data is required to transform to numerical values. Therefore, continuous values are rescaled to the values in the range of [0,1] by using normalisation method. Boolean values are transformed to values of 0 or 1 respectively. Categorical values are transformed as discrete-Boolean categorical values. They are then treated as Boolean transformation. The transformation summary can be seen in Table C5 below.

Attribute Name	Orginal data Type	Transformation Range	Methods
OP DURATION	Continuous	[0,1]	NewVal=(OldVal -Min)/(Max-Min)
ASA	Continuous	[0,1]	NewVal=(OldVal -Min)/(Max-Min)
EBL	Continuous	[0,1]	NewVal=(OldVal -Min)/(Max-Min)
SHUNT FOR CEA	Boolean	0/1	Yes=1; No=0
PATCH	Categorical	0/1	non-discrete
R1-A SIDE	Boolean	0/1	Right=1; Left=0
AGE	Continuous	[0,1]	NewVal=(OldVal -Min)/(Max-Min)
Sex	Boolean	0/1	Male=1; Female=0
HYPERTENSION HX	Boolean	0/1	Yes=1; None=0
RENAL HX	Boolean	0/1	Abnormal=1; normal=0
RESPIRATORY DIS HX	Categorical	0/1	Normal=0; MildCOAD=1; ModCOAD=1; Sev COAD=1
CABG	Boolean	0/1	Yes=1; No=0
ECG	Categorical	0/1	non-discrete
HD	Boolean	0/1	Y=1; N=0
DIABETES	Categorical	0/1	non-discrete
St	Boolean	0/1	Y=1; N=0

Table C5: Transformation summary for CM3aD.

- **Filtering task**: The two levels of expected outcome are calculated in the following heuristic rules. This is based on two attributes of *"PATIENT STATUS"* and *"COMBINE"* derived from (*Heart Disease, Diabetes, and Stroke)*.

Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$ *"Low risk"*

Σ(PATIENT STATUS, COMBINE) $\geq$ 1 $\rightarrow$*"High risk"*

Hence, this model contains 284 "*Low risk"* patterns, and 57 "*High risk"* patterns.

- **Step 3 (Data Mining Techniques)**: A WEKA software package (WEKA, 2005) is used. Alternative neural network classifiers as multilayer perceptron, radial basis function, and support vector machine are applied to the model. Classification results can be seen in Table C6 below. Note that number of cross-validation fold is 10; and alternative topologies are shown inside the Table C6. For example, the multilayer perceptron classifier namely "MLP_TP3" is used here with a topology of 16-0-1 (16 input nodes; 0 hidden node; and 1 output - 2 class nodes); learning rate η =0.3; and number of cycles = 100 epochs. The topology of 16-0-1 is chosen based on heuristic suggestions (about 10 cases/weight/class).

- **Step 4 (Comparison/ Evaluation):** The standard measures such as sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and mean square error are used (see detail in Table C6 below).

Model Hull_POSS.

- **Step 1 (Selection):** The data is taken from the Hull site with a selection of 22 input attributes, and 497 patients. The structure can be seen in Table C7.

- **Step 2 (Clean/Transform/Filter)**:

    - **Cleaning and transformation tasks:** The missing values are filled as the same above experiment method. The summary of filling missing data can be seen in Table C7. For example, the "Respiratory" missing values are filled by "Normal" value, because its most frequency is 431. The data is transformed to numerical values as indicated method above.

    - **Filtering task**: From the summary of data in Table C7, some input attributes can be eliminated. For example, attribute "JVP" contains 495/497 value of "N"; the attribute "PERI_SOILING" contains 497/497 values of "None". Obviously, these attributes are eliminated. Similarly, attributes as "GCS (Coma Score)"; "URGENCY"; "OPSEVERITY"; and "MALIGNANCY" can be eliminated with the same explanations. Therefore, model Hull_POSS

contains now 16 instead of 22 input attributes. The expected outcome is based on the attribute "PATIENT_STATUS". Heuristic rule is as follows:

IF PATIENT_STATUS = "Dead" ➔ *"High risk"*

Otherwise, ➔ *"Low risk"*

Therefore, the model includes 78 "High risk" patterns and 419 "Low risk" patterns.

Topologies and Parameters	**Risk**	**Confusion Matrix**		**ACC**	**Sen**	**Spec**	**PPV**	**NPV**	**MSE**
		High risk	*Low risk*						
MLP_TP1 (2H; η=0.3; 500 epochs)	*High risk*	27	30	0.88	0.47	0.96	0.73	0.90	0.09
	Low risk	10	274						
MLP_TP2 (0H; η=0.01;100 epochs)	*High risk*	0	57	0.83	0.00	1.00	N/A	0.83	0.11
	Low risk	0	284						
MLP_TP3 (0H; η=0.3; 100 epochs)	*High risk*	28	29	0.90	0.49	0.98	0.85	0.91	0.09
	Low risk	5	279						
MLP_TP4 (0H; η=0.3; 500 epochs)	*High risk*	27	30	0.90	0.47	0.98	0.84	0.90	0.09
	Low risk	5	279						
RBF_TP5 (c=0)	*High risk*	26	31	0.87	0.46	0.96	0.68	0.90	0.09
	Low risk	12	272						
RBF_TP6 (c=1)	*High risk*	27	30	0.85	0.47	0.93	0.56	0.90	0.1
	Low risk	21	263						
SVM_TP8 (poly kernel, w=1, p=1)	*High risk*	17	40	0.88	0.30	1.00	0.94	0.88	0.11
	Low risk	1	283						
SVM_TP9 (poly kernel, w=2, p=2)	*High risk*	27	30	0.89	0.47	0.98	0.82	0.90	0.1
	Low risk	6	278						
SVM_TP10 (rad-kernel w=1;δ=0.01)	*High risk*	0	57	0.83	0.00	1.00	N/A	0.83	0.16
	Low risk	0	284						

Table C6: CM3aD results with alternative classifiers and parameters.

Attribute Name	Attribute Type	Missing Values	Attribute Values	Max Freq/Mean
PhysiolScore	Continuous		[12,41]	20.37
OpSevScore	Continuous		[13,23]	14.29
AGE	Continuous		[38,93]	68
RESPIRATORY	Categorical	1	Normal/Mild COAD/Mod COAD/Severe COAD/Null	431(Normal)
WARFARIN	Boolean	2	Y/N/Null	474(N)
RESP_SYSTEM	Categorical	1	Limitting SOB/No SOB/ Null/SOB at rest/ SOB in exertion	468 (No SOB)
BP	Continuous	21	[90,220]	151.9
PULSE	Continuous	23	[42,110]	74
JVP	Boolean	2	N	495(N)
WCC	Continuous	10	[4, 24.3]	7.67
HAEMOGLOBIN (Hb)	Continuous	10	[7.7,18.2]	13.9
UREA	Continuous	8	[2.1, 17.2]	6.34
SODIUM(Na)	Continuous	11	[122, 146]	138.5
POTASSIUM(Ka)	Continuous	9	[3, 5.6]	4.3
ECG	Categorical	16	≥5 ectopics/min; Afib 60-90; Normal; Null; Other abnormal; Q waves; ST/T Wave change	338 (Normal)
GCS(Coma Score)	Continuous	1	[15]	15
URGENCY	Categorical	1	Elective; Scheduled urgent	496(Elective)
BLOOD_LOSS	Continuous	8	[100, 1800]	318
NO_PROCS	Discrete number	59	[1; 2; 3]	418 (1)
OP_SEVERITY	Categorical	0	Major/ Mayjor Plus	497
MALIGNANCY	Categorical	0	None	497(None)
PERI_SOILING	Categorical	0	None	497(None)

Table C7:Hull_POSS structure and summary.

- **Step 3 (Data Mining Techniques):** Alternative neural network classifiers as multilayer perceptron, radial basis function, and support vector machine are used with different parameters (see in Table C8) by using WEKA software package (WEKA, 2005). For example, classifier "Hull_POSS_TP1" is applied to data with multilayer perceptron of 22-2-1 topology (22 input nodes; 2 hidden nodes; and 1 output -2 class nodes); learning rate

of 0.3; and 500 training epochs. Note that the number of cross-validation folds is 10. Classification results can be seen in Table C8 below.

- **Step 4 (Comparison/ Evaluation):** Data outcomes are presented in confusion matrix for the use of standard measurements (see in Table C8).

Classifier	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV	MSE
		High risk	*Low risk*						
Hull_POSS_TP1	*High risk*	9	69						
(MLP_2H_0.3_500)	*Low risk*	18	401	0.82	0.12	0.96	0.33	0.85	0.14
Hull_POSS_TP2	*High risk*	6	72						
(MLP_0H_0.3_500)	*Low risk*	7	412	0.84	0.08	0.98	0.46	0.85	0.14
Hull_POSS_TP3	*High risk*	1	77						
(RBF_c_2)	*Low risk*	4	415	0.84	0.01	0.99	0.20	0.84	0.13
Hull_POSS_TP4	*High risk*	0	78						
(SVM_Poly_p_2)	*Low risk*	2	417	0.84	0.00	1.00	0.00	0.84	0.16

Table C8: Hull_POSS results with alternative techniques and parameters.

C.3. Case Study III

Clinical Model CM3bD

- **Step 1 (Selection):** The data is taken from the Dundee site with a selection of 16 input attributes and 341 patients (model CM3bD).

- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning and transformation tasks:** The data preparations including cleaning and transformation are similar as in the Case Study II in section C.2 (see in Table C4 and C5).
    - **Filtering task:** The expected outcomes are calculated based on two attributes of “PATIENT STATUS” and “COMBINE” as the following heuristic rules. Note that the expected outcomes here are used for comparison purpose only with the clustering results.

Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$*"Low risk"*

Σ(PATIENT STATUS, COMBINE) = 1 $\rightarrow$*"Medium risk"*

Σ(PATIENT STATUS, COMBINE) = 2 $\rightarrow$*"High risk"*

Hence, the CM3bD model contains 48 "*High risk*"; 73 "*Medium risk*"; and 220 "*Low risk" patterns*.

- **Step 3 (Data Mining Techniques)**: A SOM Toolbox clustering tool in the Matlab software package (SOM toolbox, 2000-2005) is used. Data is stored in a matrix of 341 x 16 (341 rows and 16 columns). A map with a size of [30, 16] is created. Note that value of 30 is length (*munits*); value of 16 (number of attributes) is dim of the map; and the size of [30,16] is based on a heuristic formulas for "*munits"* (Alhoniemi et al, 2005) as:

$$(5*341 \text{ (rows) } \hat{}0.54321)/4 \approx 30$$

The data is trained with the Best Matching Unit algorithm (BMU- SOM toolbox, 2000-2005). This means all the distances between input vectors and map units (map nodes) is calculated. The greatest similarity (minimum Euclidean distance) to input vectors is chosen. The node here is so called winner node. The map weight is updated; and self organizing map algorithm is continued until all the input vectors to be tested. The final map has quantization error of 0.438, and topographic error of 0.000. The visualized map of Umatrix, all components can be seen in Figure C1 below.

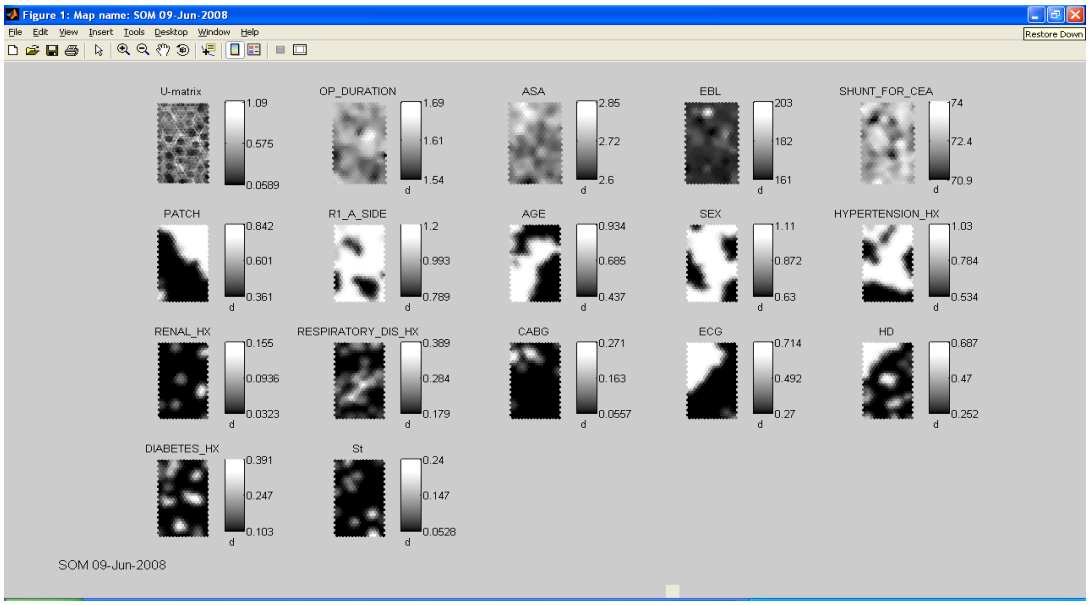

Figure C1: The U-matrix and each component plane for model CM3bD.

The data is distributed in U-matrix and individual map component plane (individual attribute). For example, "*OP_DURATION"* plane showed the continued data. To be clearer Figure C2 shows the distribution of data in only U-matrix.

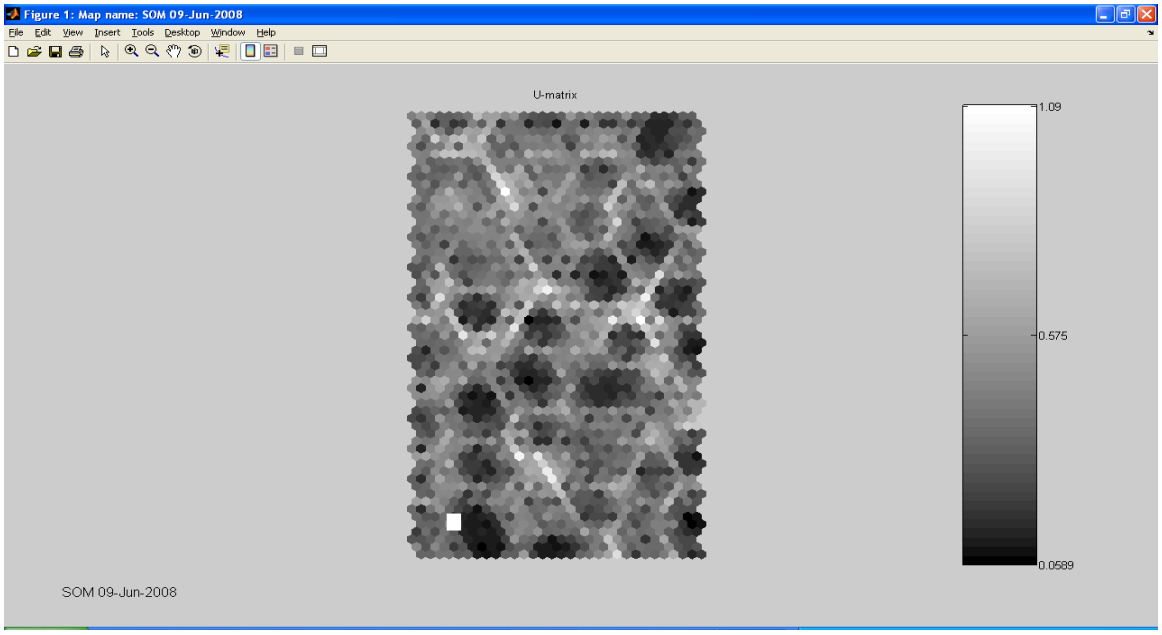

Figure C2: The U-matrix for model CM3bD.

Data is clustered by SOM Kmeans algorithm (SOM toolbox, 2005). The map cluster results and its data label distribution can be seen in Figure C3 below.

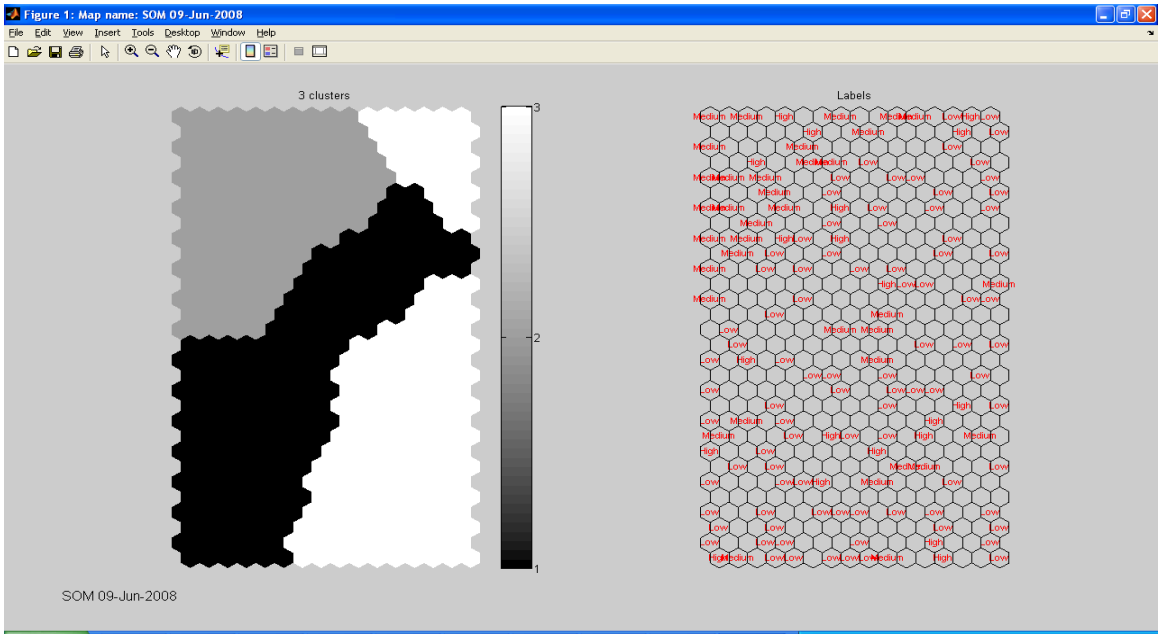

Figure C3: The clustering results of self organizing map Kmeans algorithm.

- **Step 4 (Comparison/ Evaluation):** This step is ignored in this experiment, because its objective is to illustrate the clustering data representation on the map.

C.4. Case study IV

Clinical Models CM3aD and CM3bD

- **Step 1 (Selection):** Two data sets of CM3aD and CM3bD are taken from Dundee site with 16 input attributes and 341 patient cases. Their structure can be seen in Table C4 in section C.2.

- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning task:** The strategy of filling missing values is the same as in section C.2.
    - **Transformation task:** The continuous values are purely to numerical values between the range [0,1]. Other attributes including Boolean and categorical are treated as the original attribute data types. The summary of 16 input attributes can be seen in Table C9 below.

- **Filtering task:** The heuristic rules for both models outcomes are given by:
    - **Model 1 (CM3aD)**: The two outcome levels are calculated based on two attributes of *"PATIENT STATUS"* and *"COMBINE"*.

      Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$ *"Low risk"*

      Σ(PATIENT STATUS, COMBINE) $\geq$ 1 $\rightarrow$"*High risk"*
    - **Model 2 (CM3bD):** Its outcomes are divided into three levels of ris predictions:

      Σ(PATIENT STATUS, COMBINE) = 0 $\rightarrow$*"Low risk"*

      Σ(PATIENT STATUS, COMBINE) = 1 $\rightarrow$"*Medium risk"*

      Σ(PATIENT STATUS, COMBINE) = 2 $\rightarrow$*"High risk"*

      Hence, the model CM3aD contains 57 values of "*High risk*"; and 284 values of "*Low risk*". The model CM3bD contains 48 values of "*High risk*"; 73 values of "*Medium risk*"; and 220 values of "*Low risk"*.

Attribute Name	Original Data Type	Attribute Values	Transformed Range
OP DURATION	Continuous	[0.7,3]	[0,1]
ASA	Continuous	[1,4]	[0,1]
EBL	Continuous	[0,2000]	[0,1]
SHUNT FOR CEA	Boolean	Yes/No	
PATCH	Categorical	None/Other/PTFE/Vein	
R1-A SIDE	Boolean	Left/Right	
AGE	continuous	[42,86]	[0,1]
Sex	Boolean	Male/Female	
HYPERTENSION HX	Boolean	None/Yes	
RENAL HX	Boolean	Normal/Abnormal	
RESPIRATORY DIS HX	Categorical	Normal/Mild COAD/Mod COAD/sev COAD	
CABG	Boolean	No/Yes	
ECG	Categorical	Normal/A-Fib<90/Other	
HD	Boolean	Y/N	
DIABETES	Categorical	Diet Rx/IGT/Insulin(NIDD) /Insulin(NIDDM)/none	
St	Boolean	Y/N	

Table C9: Transformation summary of 16 inputs for models CM3aD, CM3bD.

- **Step 3 (Data Mining Techniques)**: The models data is used with KMIX algorithm with alternative output clusters (k=2 and k=3) corresponding to the number of outcome classes in models CM3aD and CM3bD. The outcomes are assigned to classes "C1", and "C2" (for model CM3aD); and "C1", "C2", and "C3" (for model CM3bD). The resulted clusters of "C1" and "C2" are implicit as "Low risk" and "High risk" respectively. This is, for example, based on number of "Low risk" belong to cluster "C1" greater than that in cluster "C2". The same explanation is done for the classes in model CM3bD.

- **Step 4 (Comparison/ Evaluation):** The confusion matrix is used in this step. All standard measures are used (see Tables C10, C11).

	C2 (High risk)	C1 (Low risk)	ACC	Sen	Spec	PPV	NPV
High risk	39	18	0.48	0.68	0.44	0.20	0.88
Low risk	158	126					

Table C10: Clustering results for CM3aD model.

	C3 (High risk)	C2 (Medium risk)	C1 (Low risk)	ACC	Sen	Spec	PPV	NPV
High risk	18	17	13	0.42	0.70	0.26	0.34	0.62
Medium risk	28	22	23					
Low risk	103	59	58					

Table C11: Clustering results for CM3aD model.

- **Step 5 (Clustering Models):** A new model of CM3aDC is built based on KMIX results for the CM3aD. This new model contain includes 16 input attributes, 341 cases, and the expected outcomes ("C1L", and "C2H"). Similarly, a new model of CM3bDC is built based on the CM3bD with expected outcomes as "C1L", "C2M", and "C3H". Therefore, the CM3aDC contains 144 values of "*C1L*"; and 197 values of "*C2H*". The model CM3bDC contains 94 values of "*C1L*"; 149 values of "*C2M*"; and 98 values of "*C3H"*.

New Data Mining Process

Two models CM3aDC and CM3bDC are used with alternative supervised techniques (multilayer perceptron, radial basis function, and support vector machine). Note that the topology and its parameters here are as follows: multilayer perceptron is used with a topology of 16-0-1 (16 input nodes; 0 hidden nodes; 1 output- 2class nodes for CM3aDC), and 16-0-3 (16 input nodes; 0 hidden nodes; 3 output nodes as three level of risks) for CM3bDC, learning rate η=0.3, and number of cycles= 100 epochs; radial basis function has number of centre c of 2; and support vector machine is used with poly kernel function, and exponent parameter *p=2.* The 10-fold cross-validation is used. The results can be seen in Tables C12 and C13.

Classifiers		C2H	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3aDC-MLP (MLP16-0-1; 0.3; 100 epochs)	*C2H*	188	9	0.95	0.95	0.94	0.96	0.94	0.04
	C1L	8	136						
CM3aDC- RBF (RBF_c=1)	*C2H*	189	8	0.95	0.96	0.93	0.95	0.94	0.03
	C1L	10	134						
CM3aDC-SVM (SVM_poly_p=1)	*C2H*	185	12	0.91	0.94	0.86	0.90	0.91	0.09
	C1L	20	124						

Table C12: Neural network results of CM3aDC.

Classifier		C3H	C2M	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3bDC-MLP (MLP16-0-3; 0.3; 100 epochs)	*C3H*	97	0	1	0.96	0.98	0.89	0.96	0.98	0.02
	C2M	0	148	1						
	C1L	8	2	84						
CM3bDC-RBF (RBF_c=1)	*C3H*	81	7	10	0.92	0.94	0.87	0.95	0.85	0.06
	C2M	9	135	5						
	C1L	5	7	82						
CM3bDC-SVM (SVM_poly_p =1)	*C3H*	97	0	1	0.97	0.99	0.90	0.96	0.99	0.08
	C2M	0	149	0						
	C1L	7	2	85						

Table C13: Neural network results of CM3bDC.

C.5. Chapter 7 Experiments

C.5.1. Clinical Models CM1 and CM2

- **Step 1 (Selection):** The data structure here includes 26 attributes (24 inputs and 2 attributes for outcome heuristic calculations) and 839 patient records. They are the common attributes derived from the Hull (498 cases) and the Dundee (341 cases) data sites. The data structure can be seen in Table C14 below.

- **Step 2 (Clean/Transform/Filter):**

    - **Cleaning task:** The missing values are filled as the same method above (the mean for the continuous attributes) and the mode (for the categorical or Boolean attributes). However, a specialized heuristic transformation for the missing values in attribute "*PATCH"* is used. There were 253 missing

values in both sites where 243 missing values in the Hull site, and 10 missing values in the Dundee site. Therefore, missing values are replaced by the mode values of *PATCH* attribute in the Hull site ("*Dacron")* whereas they are replaced by *"PTFE*" in the Dundee site.

- **Transformation task**: The continuous values in data set are also purely transformed into the range [0,1] with the linear transformation method indicated in Chapter 5 (section 5.3.3 - "Data Preparation Strategy").
- **Filtering task:** The final data set contains 25 instead of 26 input attributes after removing the "empty" one (*COMP_GROUP* attribute has 650/839 missing values). The outcome for model CM1 is created by the heuristic formula based on "PATIENT_STATUS " attribute as follows:

  IF PATIENT_STATUS = "Dead" ➔ "High risk"

  Otherwise, ➔ "Low risk"

  Hence, CM1 has 126 values of "*High risk"* and 713 values of "*Low risk"* respectively. Alternatively, model CM2 outcomes are derived from by the heuristic formula corresponding to the attributes of *PATIENT_STATUS* (*attr1*), and "*30D Stroke/death)*-(*attr2*) as follows:

  IF attr1 = "Dead" or attr2="Y" ➔ *"High risk"*

  Otherwise, ➔ *"Low risk"*

  Hence, the CM2 has 139 expected outcome values of "*High risk"* and 700 values of "*Low risk"* respectively.

- **Step 3 (Data Mining Techniques)**: The neural network techniques chosen in this step are multilayer perceptron, radial basis function, and support vector machine. The classification results for the models CM1 and CM2 can be seen in Table C15 below. The same as other experiments, the data is divided into training set (90% of population), test set (10%) with a number of cross-validation folds of 10. The multilayer perceptron technique is chosen with the topology of 25-2-1 (25 input nodes; 2 hidden nodes; 1 output -2 class nodes); learning rate η is 0.3; and number of epochs is 500. Radial basis function classifier has centre parameter *c* of 2. The support vector machine uses poly kernel function with the exponent parameter p of 2.

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
AGE	Continuous	0	[38,93]	67.99
ANGINA	Boolean	12	Y/N/Null	570 (N)
ARRHYTHMIA	Categorical	8	none/A-Fib<90/min/A-Fib <90/Null/Other	792 (None)
ASPIRIN	Boolean	166	Y/N/Null	648 (Y)
ASA_GRADE	Continuous	38	[1,4]	2.24
BLOOD_LOSS	Continuous	252	[0,2000]	300.45
CABG_PLASTY	Boolean	9	Y/N/Null	778 (N)
CAROTID_DISEASE	Categorical	2	N/A	303 (TIA)
CCF	Categorical	9	<1/12/ >1/12/None/Null/Yes	803 (None)
COMP_GROUP	Categorical	605	N/A (removed)	N/A
D	Boolean	1	Y/N/Null	748 (N)
DURATION	Continuous	72	[0.7-5]	1.57
ECG	Categorical	33	Normal/Null/other abnormal/Q wave/ST/A-Fib<<90/and so on	571 (Normal)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449 (N)
Smoking	Boolean	0	Y/N	787 (Y)
PATCH	Categorical	253	PTFE/Dacron/Vein/Other Vein/Stent	171 (PTFE-170/341-Dundee site); 185 (Dacron -185/499 - Hull site)
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/Mild COAD/Mod COAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
WARFARIN	Boolean	5	Y/N/Null	809 (N)
R1-A SIDE	Boolean	0	Left/Right	458 (left)

Table C14: CM1 and CM2 data structure, and their summary.

- **Step 4 (Comparison/ Evaluation):** The results are presented in confusion matrix for the use of standard measurements of sensitivity, specificity, and so on (see Table C15).

Classifiers	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV
		High risk	**Low risk**					
	High risk	9	117					
CM1-MLP	*Low risk*	34	679	0.82	0.07	0.95	0.21	0.85
	High risk	0	126					
CM1-RBF	*Low risk*	0	713	0.85	0.00	1.00	N/A	0.85
	High risk	30	96					
CM1-SVM	*Low risk*	112	601	0.75	0.24	0.84	0.21	0.86
	High risk	6	133					
CM2-MLP	*Low risk*	27	673	0.81	0.04	0.96	0.18	0.83
	High risk	0	139					
CM2-RBF	*Low risk*	0	700	0.83	0.00	1.00	N/A	0.83
	High risk	24	115					
CM2-SVM	*Low risk*	125	575	0.71	0.17	0.82	0.16	0.83

Table C15: Using neural network techniques for model CM1 and CM2.

C.5.2. Clinical Models

Models CM3a, CM3b, CM4a, and CM4b

- **Step 1 (Selection):** Models CM3a and CM3b contain 18 attributes (16 input attributes and 2 for outcome calculations) and 839 patient records (see in Table C16 below). Models CM4a and CM4b contain 16 attributes (14 input attributes and 2 for outcome calculation) and 839 patient records (see in Table C17).

- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning and Transformation tasks:** These tasks are the same as above method used for models CM1 and CM2.

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
30D stroke/death	Boolean	0	Y/N	806 (N)
AGE	Continuous	0	[38,93]	67.99
ASA_GRADE	Continuous	38	[1,4]	2.24/0.46
BLOOD_LOSS	Continuous	252	[0,2000]	300.45
CABG_PLASTY	Boolean	9	Y/N/Null	778(N)
D	Boolean	1	Y/N/Null	748(N)
DURATION	Continuous	72	[0.7-5]	1.57
ECG	Categorical	33	Normal/Null/other abnormal/Q-wave/ ST/A-Fib<<90/; so on	571 (Normal)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449(N)
PATCH	Categorical	253	PTFE/Dacron/Vein/ OtherVein /Stent	171/341 PTFE-Dundee; 185/499-Dacron -Hull
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/MildCOAD/ ModCOAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
R1-A SIDE	Boolean	0	Left/Right	458 (left)

Table C16: CM3a and CM3b data structure and their summary.

- **Filtering task:** The outcomes for model CM3a, and CM4a are calculated the same heuristic rules in CM2 model as follows:

  IF attr1 = "Dead" or attr2="Y" ➔ *"High risk"*

  Otherwise, ➔ *"Low risk"*

  Hence, CM3a, and CM4a has 139 values of "*High risk",* and 700 values of "*Low risk"*. Alternatively, the outcomes for model CM3b, and CM4b are calculated as the following heuristic rule:

IF attr1 = "Dead" and attr2="Y" ➔ *"Very High risk"*

Otherwise IF attr1="Dead" ➔ *"High risk"*

Otherwise IF attr2="Y" ➔ *"Medium risk"*

Otherwise, ➔ *"Low risk"*

Hence, CM3b, and CM4b has 19 values of "V*ery High risk"; 107* values of "*High risk"*; 13 values of "*Medium risk"*; and 700 values of "*Low risk"* respectively.

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean/Stdev
PATIENT_STATUS	Boolean	0	Alive/Dead	713 (Alive)
30D stroke/death	Boolean	0	Y/N	806 (N)
AGE	Continuous	0	[38,93]	67.99
ASA_GRADE	Continuous	38	[1,4]	2.24/0.46
D	Boolean	1	Y/N/Null	748(N)
HD	Boolean	1	Y/N/Null	550 (N)
HYPERTENSION	Boolean	7	Y/N/Null	449(N)
PATCH	Categorical	253	PTFE/Dacron/Vein/Other Vein/Stent	171/341-PTFE - Dundee; 185/499 - Dacron - Hull site
RENAL_FAILURE	Boolean	7	Y/N/Null	820 (N)
RESPIRATORY	Categorical	16	Normal/Mild COAD/Mod COAD/Severe COAD/Null	711 (Normal)
SEX	Boolean	0	M/F	507 (M)
SHUNT	Boolean	14	Y/N/Null	501 (Y)
St	Boolean	1	Y/N/Null	565 (N)
R1-A SIDE CONS	Boolean Categorical	0 0	Left/Right 1;2;3;4;5	 383 (4)
Vascular Unit	Categorical	0	1;2	498 (2)

Table C17: CM4a and CM4b data structure and their summary.

- **Step 3 (Data Mining Techniques)**: The same neural network techniques as above experiments are used for these models. The classification results can be seen in Table C18, and C19. The detail topology and parameters as follows: multilayer perceptron technique is used with topologies of 16-2-1 (CM3a), 14-2-1 (CM4a), 16-2-4 (CM3b), and 14-2-4 (CM4b); learning rate η is 0.3; and number of epochs is 500. Radial basis function classifier has centre parameter *c* of 2; and support vector machine uses poly kernel function with the exponent parameter *p* of *2.*

- **Step 4 (Comparison/ Evaluation):** The comparisons are fulfilled based on the confusion matrix and standard rates of sensitivity, specificity, and so on.

Classifiers	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV
		High risk	Low risk					
CM3a-MLP	*High risk*	13	126	0.81	0.09	0.95	0.28	0.84
	Low risk	34	666					
CM3a-RBF	*High risk*	0	139	0.83	0.00	1.00	N/A	0.83
	Low risk	0	700					
CM3a-SVM	*High risk*	16	123	0.77	0.12	0.90	0.19	0.84
	Low risk	67	633					
CM4a-MLP	*High risk*	14	125	0.81	0.10	0.95	0.30	0.84
	Low risk	32	668					
CM4a-RBF	*High risk*	0	139	0.83	0.00	1.00	N/A	0.83
	Low risk	0	700					
CM4a-SVM	*High risk*	18	121	0.79	0.13	0.92	0.24	0.84
	Low risk	58	642					

Table C18: Experimental results of CM3a and CM4a models.

Classi-fiers	Risk	Confusion Matrix				ACC	Sen	Spec	PPV	NPV
		Very High risk	High risk	Medi-um risk	Low risk					
CM3b-MLP	*Very High risk*	2	1	0	16	0.84	0.04	0.97	0.25	0.84
	High risk	0	3	0	104					
	Medium risk	0	0	0	13					
	Low risk	5	13	0	682					
CM3b-RBF	*Very High risk*	0	0	0	19	0.85	0.00	1.00	N/A	0.83
	High risk	0	0	0	107					
	Medium risk	0	0	0	13					
	Low risk	0	0	0	700					
CM3b-SVM	*Very High risk*	0	3	0	16	0.79	0.08	0.90	0.14	0.83
	High risk	1	6	0	100					
	Medium risk	0	0	1	12					
	Low risk	13	46	8	633					
CM4b-MLP	*Very High risk*	0	1	0	18	0.83	0.07	0.96	0.27	0.84
	High risk	0	8	0	99					
	Medium risk	0	1	0	12					
	Low risk	0	27	0	673					
CM4b-RBF	*Very High risk*	0	0	0	19	0.85	0.00	1.00	N/A	0.83
	High risk	0	0	0	107					
	Medium risk	0	0	0	13					
	Low risk	0	0	0	700					
CM4b-SVM	*Very High risk*	2	4	0	13	0.8	0.14	0.90	0.21	0.84
	High risk	2	11	0	94					
	Medium risk	0	0	0	13					
	Low risk	17	45	9	629					

Table C19: Experimental results of CM3b and CM4b models.

C.5.3. Scoring Risk models

- **Step 1 (Selection):** The data is selected from the Hull site and the POSSUM and PPOSSUM results. The models are built as Mortality, Morbidity, and Death rate. Note that these models share the same structure (see in Table C20) including 499 cases and 22 input attributes.
- **Step 2 (Clean/Transform/Filter):**
    - **Cleaning task:** The method of filling missing values is the same as experiments above. This means, for example, continuous missing values are replaced by the mean of non-missing numerical values. Note that the second column in Table C20 shows number of missing values whereas the last column shows replacing values if applicable. For example, the missing values in the attribute *WCC* are replaced by the mean (7.67).
    - **Transformation task:** This task is to transform all data to numerical data type. This means the numerical data is rescaled in to the range of [0,1]. Boolean values are transformed into values of 0 or 1. The categorical data is transformed into discrete Boolean ("Normal" and "Abnormal") then they are transformed into values of 0 or 1.
    - **Filtering task**: Some attributes in these models structure can be eliminated. They are JVP, GCS (Coma Score), URGENCY, OP-SEVERITY, MALIGNANCY, and PERI_SOILING (see the summary in Table C20). For example, the attribute namely JVP contained only value of *"N"*, or MALIGNANCY contained only value of "*None"* as well. Hence, the data sets contain now 16 input attributes and 497 cases.

The outcome for three models are calculated basing on the average (mean) values of *Mortality, Morbidity,* and *Death rate* scores as follows:.

Mortality model:

IF Mortality>= mean ➔ *"High risk"*

Otherwise ➔ *"Low risk"*

Attribute name	Attribute type	Missing values	Attribute values	Max Freq/Mean
PhysiolScore	Continuous	0	[12,41]	20.37
OpSevScore	Continuous	0	[13,23]	14.29
AGE	Continuous	0	[38,93]	67.99
RESPIRATORY	Categorical	1	Normal/MildCOAD/ModCOAD/ Severe COAD/Null	431(Normal)
WARFARIN	Boolean	2	Y/N/Null	474(N)
RESP_SYSTEM	Categorical	1	Limiting SOB/No SOB/ Null/SOB at rest/ SOB in exertion	468 (No SOB)
BP	Continuous	21	[90,220]	151.9
PULSE	Continuous	23	[42,110]	74
JVP	Boolean	2	N	495(N)
WCC	Continuous	10	[4, 24.3]	7.67
HAEMOGLOBIN(Hb)	Continuous	10	[7.7,18.2]	13.9
UREA	Continuous	8	[2.1, 17.2]	6.34
SODIUM(Na)	Continuous	11	[122, 146]	138.5
POTASSIUM(Ka)	Continuous	9	[3, 5.6]	4.3
ECG	Categorical	16	≥5 ectopics/min; Afib 60-90; Normal; Null; Other abnormal; Q waves; ST/T Wave change	338 (Normal)
GCS(Coma Score)	Continuous	1	[15]	15
URGENCY	Categorical	1	Elective; Scheduled urgent	497(Elective)
BLOOD_LOSS	Continuous	8	[100, 1800]	318
NO_PROCS	Discrete number	59	[1; 2; 3]	418 (1)
OP_SEVERITY	Categorical	0	Major Plus	497 (Major Plus)
MALIGNANCY	Categorical	0	None	497 (None)
PERI_SOILING	Categorical	0	None	497 (None)

Table C20: Scoring Risk models' input structure and their summary.

Morbidity model:

IF Morbidity>= mean ➔ *"High risk"*

Otherwise ➔ *"Low risk"*

Death rate model:

IF Death rate>= mean ➔ *"High risk"*

Otherwise ➔ *"Low risk"*

Therefore, the Mortality has got 25 "*High risk*", 474 "*Low risk"*; Morbidity has got 39 "*High risk*", 460 "*Low risk"*; and Death rate contains 26 "*High risk*", 473 "*Low risk"* respectively.

- **Step 3 (Data Mining Techniques):** This step is ignored, because of the comparison purpose between the outcomes of scoring risk models and the actual risks in later step.

- **Step 4 (Comparison/ Evaluation):** The three models outcomes derived from step 2 are compared to the actual risks. Note that the "PATIENT STATUS" values as "Dead" or "Alive" are assigned as "High risk" or "Low risk" for the comparison purpose. The detailed comparisons can be seen in Table C21.

Classifiers	Risk	Confusion Matrix		ACC	Sen	Spec	PPV	NPV
		High risk	**Low risk**					
Mortality	*High risk*	10	69	0.83	0.13	0.96	0.40	0.85
	Low risk	15	405					
Morbidity	*High risk*	15	64	0.82	0.19	0.94	0.38	0.86
	Low risk	24	396					
Death rate	*High risk*	10	69	0.83	0.13	0.96	0.38	0.85
	Low risk	16	404					

Table C21: Confusion matrix for scoring risk models.

C.5.4. Clustering Models

Models CM3a and CM3b

- **Step 1 (Selection):** A selection data set for models CM3a and CM3b includes 16 input attributes and 839 patient records (see the structures and summaries in Table C16).

- **Step 2 (Clean/Transform/Filter)**:

  - **Cleaning task:** The missing values are treated as the same as models CM3a, CM3b in section C.5.2 (detailed replacements of missing values can be seen in Table C16).

- **Transformation task**: The continuous values in data set are also purely transformed into the range [0,1] as linear transformation method. Other attributes (Boolean and categorical) are ignored in this task.
- **Filtering task**: The expected outcomes for both models of CM3a and CM3b are calculated the same as in section C.5.2. Note that these outcomes are only for the comparison purpose after use of clustering algorithms.

- **Step 3 (Data Mining Techniques):** Both models of CM3a and CM3b are used with the KMIX algorithm. The number of chosen clusters is 2 or 3, according to the models CM3a and CM3b respectively. The clustering results are assigned into alternative clusters as "C2H" and "C1L" for model CM3a; and "C4VH", "C3H", "C2M", and "C1L" for model CM3b.

- **Step 4 (Comparison/ Evaluation):** Clustering results are evaluated in confusion matrix with the expected outcomes in step 2 (see in Table C22 and Table C23).

Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV
High risk	48	91	0.60	0.35	0.65	0.16	0.83
Low risk	248	452					

Table C22: The clustering results for model CM3a.

Risk	C4VH	C3H	C2M	C1L	ACC	Sen	Spec	PPV	NPV
Very High risk	7	0	6	6	0.45	0.89	0.38	0.18	0.96
High risk	43	0	33	3					
Medium risk	5	0	5	3					
Low risk	249	0	199	280					

Table C23: The clustering results for model CM3b.

- **Step 5 (Building New Models):** Two new models CM3aC and CM3bC are built based on the clustering results and the input attribute set derived from models CM3a and CM3b. Therefore, the model CM3aC contains 16 inputs and 839 cases. Its outcome set contains 403 "C2H" and 436 "C2L". The model CM3bC shares the same structure as model CM3a. Its outcome set contains 304 "C4VH", none values of "C3H", 243 "C2M", and 292 "C1L". Both models CM3aC and CM3bC are used alternative neural network

techniques. The topologies and parameters are the same as in model CM3a and CM3b experiments in section C.5.2 above. The results can be seen in Table C24 and C25.

Classifier	Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3aC-MLP (MLP_2H_0.3_500)	*C2H*	296	0	1	1	1.00	0.99	1	0
	C1L	2	541						
CM3aC-RBF (RBF_c=2)	*C2H*	230	66	0.77	0.78	0.77	0.65	0.86	0.14
	C1L	124	419						
CM3aC-SVM (SVM_poly_p=2)	*C2H*	293	3	0.99	0.99	0.99	0.99	0.99	0.01
	C1L	3	540						

Table C24: The CM3aC model results with alternative neural network classifiers.

Classifiers	Risk	C4VH	C2M	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3bC-MLP (MLP_2H_0.3_500)	*C4VH*	302	1	1	0.99	0.99	0.99	0.99	0.97	0.01
	C2M	3	233	7						
	C1L	1	3	288						
CM3bC-RBF (RBF_c=2)	*C4VH*	300	2	2	0.98	0.99	0.97	0.98	0.98	0.01
	C2M	3	235	5						
	C1L	4	5	283						
CM3bC-SVM (SVM_Poly_p=2)	*C4VH*	304	0	0	0.98	0.99	0.96	0.98	0.99	0.07
	C2M	0	240	3						
	C1L	0	12	280						

Table C25: The CM3bC model results with alternative neural network classifiers.

C.6. Case Study V

Model CM3aD and Model CM2

- **Step 1 (Selection):** A selection data set of model CM3aD includes 18 attributes (16 input attributes and 2 for outcome calculations) and 341 patient records (see in Table C4). Another selection data set of model CM2 includes 26 attributes and 839 cases data derived from the Hull and Dundee sites (see in Table C14).

- **Step 2 (Clean/Transform/Filter)**:

    - **Cleaning task:** The missing values are treated as above experiments.

- **Transformation task**: All data is transformed to appropriate categorical types. Boolean values can be seen as categorical Boolean values of 0 or 1. The categorical values are ignored in this task. The numerical data is transformed to categorical by discretization method (Venables and Ripley, 1994; Yang et al, 2001; and Tourassi et al, 2001). This means continuous data is divided into alternative *bins* with the proper number *bin* is $log_2 N + 1$, where N is number of cases. For example, the models CM3aD and CM2 have got 341 and 839 cases. Therefore, the number of *bins* is $log_2 341 + 1 = 9$ and $log_2 839 + 1 = 11$ respectively.

The summary of the transformation from numerical values into categorical values for the attributes in the models of CM3aD and CM2 can be seen in Table C26 and C27. For example, the attribute namely "OPDURATION_BIN" in Table C26 has the categorical values, which are resulted after transformation, as: "BIN1", "BIN2", to "BIN9". This attribute contains the number of cases felt into each "BIN" such as 5, 68, and so on (see detail in Table C26).

BIN	OPDURATION_BIN	EBL_BIN	AGE_BIN	ASA_BIN
1	5	313	1	4
2	68	19	9	0
3	33	6	13	0
4	172	1	50	165
5	7	1	63	37
6	48	0	78	0
7	2	0	82	129
8	5	0	35	0
9	1	1	10	6
Total	341	341	341	341

Table C26: Bins and its summary for CM3aD.

BIN	OPDURATION_BIN	EBL_BIN	AGE_BIN	ASA_BIN
1	3	4	99	155
2	6	0	620	73
3	20	0	66	403
4	42	607	31	157
5	134	38	11	29
6	171	0	7	15
7	214	0	2	0
8	169	182	0	3
9	59	0	1	3
10	20	0	1	0
11	1	8	1	1
Total	839	839	839	839

Table C27: Bins and its summary for CM2.

- **Filtering task**: This task is ignored in this experiment.

- **Step 3 (Use Mutual Information Calculations):** The data in both models of CM3aD and CM2 are used with mutual information calculations. These data also are use with the Relief algorithm with the WEKA software package (WEKA, 2005; Witten and Frank, 2005). The number of neighborhood *m* is 10.

- **Step 4 (Comparison/ Evaluation):** The comparison results can be seen in Figures C4 and C5.

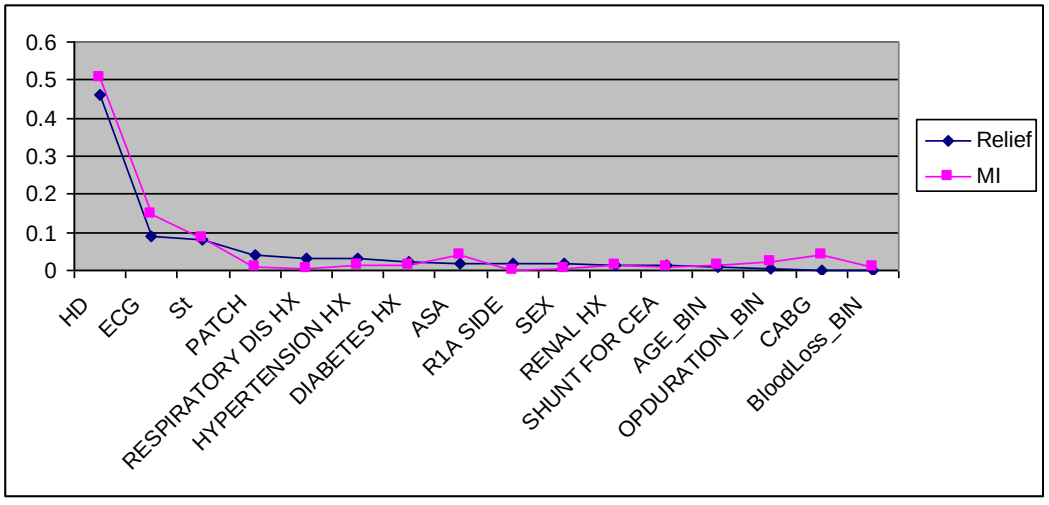

Figure C4: A comparison of MI and Relief with CM3aD model.

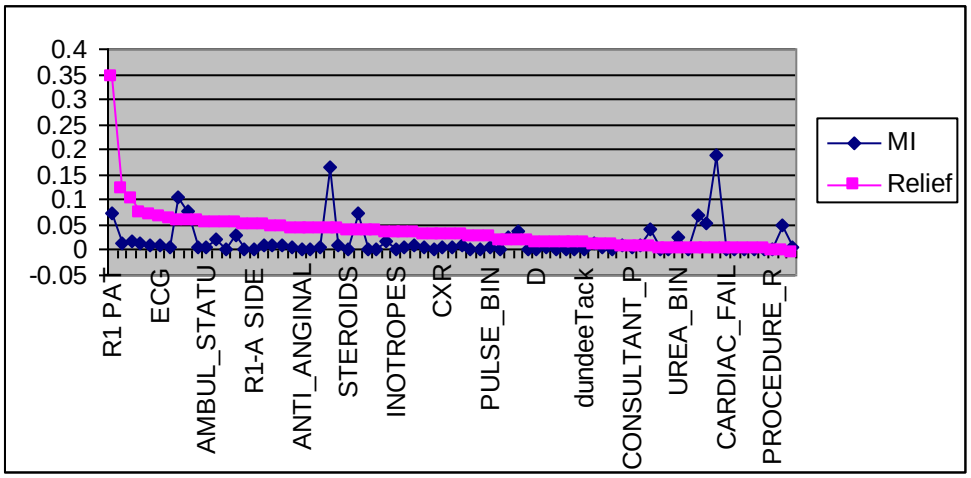

Figure C5: A comparison of MI and Relief with CM2 model.

C.7. Case Study VI

Model CM3aD

- **Step 1 (Selection):** Model CM3aD includes 16 input attributes and 341 patient records (see in Table C4).
- **Step 2 (Clean/Transform/Filter)**:
    - **Cleaning task:** The missing task is fulfilled as above experiments.
    - **Transformation task**: The data transformation is fulfilled only for numerical attributes (rescaled to the range of [0,1]), the other (Boolean and categorical) attributes are ignored.
- **Step 3 (Data Mining Techniques):** The data is used with mutual information calculations.. The detailed result can be seen in Table C28. The mutual information results are then used as attribute weights in WKMIX algorithm.

Attribute name	MI
HD	0.507
ECG	0.149
St	0.083
PATCH	0.011
RESPIRATORY DIS HX	0.005
HYPERTENSION HX	0.012
DIABETES HX	0.013
ASA	0.041
R1A SIDE	0
SEX	0.004
RENAL HX	0.013
SHUNT FOR CEA	0.01
AGE_BIN	0.013
OPDURATION_BIN	0.022
CABG	0.041
BloodLoss_BIN	0.01

Table C28: Mutual information calculation results for model CM3aD

- **Step 4 (Comparison/ Evaluation):** The comparison results can be seen in Table C29 with the use of KMIX and WKMIX for the model CM3aD over standard measurements as sensitivity, specificity, and so on.

Algorithms	Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV
WKMIX	*High risk*	30	27	0.61	0.53	0.63	0.22	0.87
	Low risk	105	179					
KMIX	*High risk*	39	18	0.48	0.68	0.44	0.20	0.88
	Low risk	158	126					

Table C29: Comparison between KMIX and WKMIX

- **Step 5 (Building New Models):** The new clustering model (CM3aDC) is built with the outcomes derived from the use WKMIX algorithm. Neural network techniques (multilayer perceptron, support vector machine, and radial basic function), and decision tree technique of J48 are used**; t**he results can be seen in Table C30.

Classifiers	Risk	C2H	C1L	ACC	Sen	Spec	PPV	NPV	MSE
CM3aDC-MLP (16-0-1; 0.3;100 epochs)	*C2H*	135	0	1	1	1	1	1	0
	C1L	0	206						
CM3aDC-RBF (c=1)	*C2H*	132	3	0.97	0.97	0.98	0.99	0.96	0.02
	C1L	6	200						
CM3aDC-SVM (poly; p=1)	*C2H*	135	0	1	1	1	1	1	0
	C1L	0	206						
CM3aDC-J48 (binary tree)	*C2H*	135	0	1	1	1	1	1	0
	C1L	0	206						

Table C30: The results of alternative techniques for CM3aDC model.