

Generating and Verifying Risk Prediction Models Using Data Mining (A Case Study from Cardiovascular Medicine)

Darryl N. Davis and Thuy T.T. Nguyen

Darryl N. Davis, University of Hull

Department of Computer Science, University of Hull, Cottingham Road, Hull, HU6 7RX, UK

e-mail D.N.Davis@hull.ac.uk

Telephone +44(0)1482466469

Fax +44(0)1482466666

Thuy T.T. Nguyen, University of Hull

Department of Computer Science, University of Hull, Cottingham Road, Hull, HU6 7RX, UK

e-mail: T.T.Nguyen@dcs.hull.ac.uk

Telephone +44(0)1482465253

Fax +44(0)1482466666

Generating and Verifying Risk Prediction Models Using Data Mining (A Case Study from Cardiovascular Medicine)

Abstract

Risk prediction models are of great interest to clinicians. They offer an explicit and repeatable means to aide the selection, from a general medical population, those patients that require a referral to medical consultants and specialists. In many medical domains, including cardiovascular medicine, no gold standard exists for selecting referral patients. Where evidential selection is required using patient data, heuristics backed up by poorly adapted more general risk prediction models are pressed into action, with less than perfect results. In this study existing clinical risk prediction models are examined and matched to the patient data to which they may be applied using classification and data mining techniques, such as neural nets. Novel risk prediction models are derived using unsupervised cluster analysis algorithms. All existing and derived models are verified as to their usefulness in medical decision support on the basis of their effectiveness on patient data from two UK sites.

Introduction

Risk prediction models are of great interest to clinicians. They offer the means to aide the selection of those patients that need referral, to medical consultants and specialists, from a general medical population. In many medical domains, including cardiovascular medicine, no gold standard exists for selecting referral patients. Existing practice relies on clinical heuristics backed up by poorly adapted generic risk prediction models. In this study existing clinical risk prediction models are examined and matched to the patient data to which they may be applied using classification and data mining techniques, such as neural nets. The evidence from earlier research suggests that there are benefits to be gained in the utilization of neural nets for medical diagnosis (Janet, 1997; Silipo & Marchesi, 1998; Lisboa, 2002).

In this chapter, the cardiovascular domain is used as an exemplar. The problems associated with identifying high risk patients, (i.e. patients at risk of a stroke, cardiac arrest or similar life threatening event), are symptomatic of other clinical domains where no gold standard exists for such purposes. In routine clinical practice, where domain specific clinical experts are unavailable to all patients, the patient's clinical record is used to identify which patient's are most likely to benefit from referral to a consulting clinician. The clinical record typically, although not always, contains generic patient data (for example age, gender etc.), a patient history of events related to the disease (for example, past strokes, cardiovascular related medical operations), and a profile of measurements, and observations from medical examinations, that characterize the nature of the patient's cardiovascular system. The general practitioner may use a risk prediction model, together with observations from medical examinations, as an aid in determining whether to refer the patient to a consultant (Gunning & Rowan, 1999). Currently, any such risk prediction model will be based on a general clinical risk prediction system, such as APACHE (Knaus et al., 1985; Knaus et al., 1991; Rowan et al., 1994) or POSSUM (Copeland et al., 1991; Copeland, 2002; Yii & Ng, 2002),

which generate a score for patients. Clinicians expert in the disease may well use further risk prediction models, based on their own research and expertise. Such risk prediction models are described in more detail in the second section. The strengths and flaws of the available models for the current clinical domain are explored in the third section, where they are used in conjunction with supervised neural nets. It should be noted, that although this chapter predominantly reports on the use of supervised neural nets and unsupervised clustering in predicting risk in patients, a wide range of other techniques, including decision trees, logistic regression, Bayesian classifiers (Bishop, 2006; Witten & Eibe, 2005) have been tried. The results from applying these other techniques are not given, but typically are similar to or worse than the results presented here. The fourth and fifth sections present an alternative to the coercion of outcome labels, arising from current risk prediction models, with the use of unsupervised clustering techniques. The results from these sections are discussed in the sixth section. The problems associated with making available to clinicians, risk prediction models that arise from the application of data mining techniques, are discussed in that and the concluding section.

Risk Prediction Models

In this section, two forms of risk prediction model, as used in routine clinical practice, are introduced. The first, POSSUM, typifies the application of generic models to specific medical disciplines. The second set reflect the clinical heuristics regularly used in medicine. The data used throughout this case study is from two UK clinical sites. The attributes are a mixture of real number, integer, Boolean and categorical values. The data records typically contain many default and missing values. For both sites there is typically too high a data value space (i.e. the space of all possible values for all attributes in the raw data) for the data volume (i.e. the number of records) to perform naïve data mining, and some form of data preprocessing is required before using any classifier if meaningful results are to be obtained. Furthermore, as can be seen in the tabulated results, the data once labeled is class-imbalanced; with low risk patients heavily outnumbering high risk patients.

The main characteristics of the cardiovascular data from Clinical Site One (98 attributes and 499 patient records) are:

- Many redundant attributes such as date or time attributes with mostly null values, or explanatory attributes. These attributes bear little relevance to the risk prediction models and experiments. For example, the attribute labelled as “*THEATRE_SESSION_DATE*” shows the date of a patient’s operation. This is ignored in these experiments, and so this feature can be removed. Other examples are “empty features”, containing mostly null values. For example, the attribute labelled as “*LOWEST_BP*” is an attribute representing the lowest blood pressure of the patients during an operation. All of its values are null except for four patient entries. Such attributes are best removed.
- Data has 7018 out of 42914 cells (16%) with missing values after removing the type of redundant attributes described above; leaving 86 "meaningful" attributes.

- Noisy and inconsistent data such as abbreviations in categorical attributes and outlier values in some numerical attributes; these are replaced with “meaningful” values.
- Data includes the scored values (*PS* and *OSS* explained below) for the POSSUM and PPOSSUM risk prediction systems.

The data from Clinical Site Two includes 431 patient records with 57 attributes. The nature and structure of the data has similar characteristics to the first site, with many redundant and noisy attributes, missing values, etc., as follows:

- The redundant attributes have the same characteristics as above. For example, the attribute “*ADMISSION_DATE*” shows the patient’s date of operation - it can be removed. Two attributes labelled as “*Surgeon name1*” and “*Surgeon name2*” represent names of operating doctors. Their values might be helpful in a more general evaluation, but offer negligible relevance to the specific purposes of this study.
- The data includes 1912 out of 12311 cells with missing values (16%) after deletion of the redundant attributes (leaving 36 "meaningful" attributes). This is the same ratio as for the first site.
- 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 will reduce the accuracy in any transformation process. Such outlier values will be replaced with more meaningful maximum and minimum values.
- The site does not include the scored values (*PS* and *OSS*) for the POSSUM and PPOSSUM systems. Moreover, the information in the data is insufficient to generate these scored values and so cannot be used with these risk prediction systems.

In summary the clinical data is noisy, contains many null values and is problematic for use with standard risk prediction models (Kuhan et al., 2001). This is typical of the challenges faced by researchers and workers who want to apply data mining techniques to clinical data. The naïve application of data mining techniques to raw data of this type typically provides poor results. Indeed clinicians availing themselves of data mining packages reported a disillusion with data mining on initially applying these packages to the data used in this study. The same may well be true for other domains.

Generic Clinical Risk Prediction Models

The Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM) (Copeland et al., 1991) and the Portsmouth POSSUM (P-POSSUM) (Prytherch et al., 2001) are generic clinical risk assessment systems which are widely used in the UK. When used over a database of patient records, they produce a classification of patients ranked in percentile groups. Neither can be used with a single patient record. This is problematic where clinicians want to compare a new patient with an existing group, or where only new patient data is available to the referring clinician.

Copeland et al. (1991) originally assessed 48 physiological factors plus 14 operative and postoperative factors for each patient. Using multivariate analysis techniques these were reduced to 12 physiological and 6 operative factors based on the following key factors:

- Physiological status of the patient
- Disease process that requires surgical intervention
- Nature of operation
- Pre and post-operative support

POSSUM is a two stage process, where the 12 physiological factors give rise to a physiological score (PS), and the 6 operative factors are used to generate an operative severity score (OSS). These can then be used to generate a Mortality and Morbidity rate.

Mortality rate: $R_1 = \frac{1}{1 + e^{-x}}$ (Equation 1)

where $x = (0.16 * \text{physiological score}) + (0.19 * \text{operative score}) - 5.91$

Morbidity rate: $R_2 = \frac{1}{1 + e^{-y}}$ (Equation 2)

where $y = (0.13 * \text{physiological score}) + (0.16 * \text{operative score}) - 7.04$

The scalars in the above formulae were found using regression analysis in the work by Copeland et al., and are medical domain independent. POSSUM was evaluated by Prytherch et al. (2001) using 10,000 patients over a two year period, and found that POSSUM over predicted mortality. In an effort to counteract the perceived shortcoming of conventional POSSUM, Whitley et al. (1996) devised the similar PPOSSUM (Portsmouth Predictor) equation for mortality. The PPOSSUM equation uses the same physiological and operative severity scores to provide a risk-adjusted operative mortality rates, but generates a Predicted Death Rate.

Predicted Death Rate: $R_3 = \frac{1}{1 + e^{-z}}$ (Equation 3)

where $z = (0.1692 * \text{PS}) + (0.150 * \text{OS}) - 9.065$

Both POSSUM and PPOSSUM give a numeric measure of risk for groups of patients. A relatively naïve clinical model for assigning individual patients as High or Low risk, based on these numbers for Mortality, Morbidity or Death rate, is possible using the following heuristics:

Mortality Prediction: IF “Mortality Score” \geq *Threshold* THEN Patient = “High risk”
 Otherwise Patient = “Low risk”

Morbidity Prediction: IF “Morbidity Score” \geq *Threshold* THEN *Patient* = “High risk”
Otherwise *Patient* = “Low risk”

Death Rate Prediction: IF “Death Rate Score” \geq *Threshold* THEN *Patient* = “High risk”
Otherwise *Patient* = “Low risk”

The threshold used in these prediction heuristics will vary depending on the intended use of the prediction, but typically the *Threshold* is set to the mean score for the predictor in question.

Clinical Heuristic Models

A different form of risk prediction model are those that make use of the clinical expertise of practitioners in the domain. Such models arise from the daily use of data for clinical purposes, or result from research, which can involve practitioners from other fields (e.g. statisticians and computer scientists). Typically, the indicative attributes used in such models are those to be found in the POSSUM models, or deemed to be important by the clinicians based on their experience or found to significant when matched against outcomes in logistic regression studies. Many of these heuristic models use an aggregation of attributes from clinical records to form the outcome labels. The following models (or risk prediction rules) arose from exactly such studies. They either make use of the attributes used in the POSSUM (and PPOSSUM) models, attributes found to be highly related to the given output from logistic regression studies, or deemed to be important, by experienced clinicians, in attributing risk (Kuhan et al., 2001).

One clinical model in this study (heuristic model CM1) uses patient death within 30 days of an operation as the “High Risk” outcome, with other patients are labeled as “Low Risk”. A further model (CM2) uses patient death or severe cardiovascular event (for example *Stroke* or *Myocardial Relapse or Cardio Vascular Arrest*) within 30 days of an operation as the “High Risk” outcome; other patients are labeled as “Low Risk”. Both the CM1 and CM2 models use all attributes from the “cleaned” patient records, other than those aggregated to form the output labels, as inputs. Further models use only a limited set of attributes. Heuristic model CM3a, for example, uses 16 input attributes derived from the CM1 and CM2 models. These attributes are thought to be significant by experienced clinicians. Again the outcome label is based on an aggregation of attributes as described for CM2. Heuristic model CM3b, a variation on CM3a, uses four outcome labels; a variation on the four attributes used in the outcome labels of CM1, CM2 and CM3a. Hence, given *Aggregate2* is *Stroke* or *Myocardial Relapse or Cardio Vascular Arrest* (all Boolean valued attributes), outcome for CM3b is determined by the following aggregation rules:

IF *Status* = “Dead” AND *Aggregate2* = “TRUE” THEN *Outcome* = “Very High risk”
Else IF *Status* = “Dead” THEN *Outcome* = “High risk”
Else IF *Aggregate2* = “TRUE” THEN *Outcome* = “Medium risk”
Otherwise *Outcome* = “Low risk”

The remaining two models (CM4a and CM4b) have similar outcome variations to CM3a and CM3b (respectively) but with 15 input attributes, derived from a combination of clinical knowledge and those marked as significant in a logistic regression exercise using the complete data.

Like the heuristic rules used with the POSSUM models, patients can be assigned a risk prediction outcome that allows a grading of patients from low to high risk. However, unlike the POSSUM models, these heuristic models can be used to provide a qualitative labeling of patients, based on individual records and do not require a complete data set of cardiovascular patients to be available. This is an important point. The opening paragraph of this chapter stated that these risk prediction models should offer the means to aide selection from a general medical population, those patients that need referral to medical consultants and specialists. In many cases the referring clinician does not have a population of cardiovascular patients with which to compare a new patient. Indeed, in many cases, patient records are represented as single entry databases. In such situations the POSSUM type models are of limited use.

Testing Existing Risk Prediction Models

The clinical and heuristic model outcomes described in the previous section were evaluated against data using a variety of supervised classifiers. A variety of performance measures were used; in particular, the Mean Square Error (*MSE*), the confusion matrix, sensitivity (*sen*) and specificity (*spec*) rates, and true positive (*tpr*) and false positive (*fpr*) rates as used in ROC (Receiver Operating Characteristics) graphs (Dunham, 2002; Kononenko & Kukar, 2007).

Assume that the data domain has n input patterns x_i ($i=1, 2, \dots, n$), each with a target pattern Y_i , for which the classifier produces the output y_i . The *MSE* is the averaged square of the error between the predicted and the target output, given by:

$$MSE = \sum_i (y_i - Y_i)^2 / n \quad (\text{Equation 4})$$

Assume that the cardiovascular classifier output set includes two outcomes {High risk; Low risk}. Each pattern x_i ($i=1, 2, \dots, n$) is allocated to one label from the set $\{P, N\}$, *Positive* (P) = *High Risk* or *Negative* (N) = *Low Risk*. Hence, each input pattern might be mapped into one of four possible classifier outcomes: *TP* (*True Positive- True High Risk*); *FP* (*False Positive- False High Risk*); *TN* (*True Negative- True Low Risk*); or *FN* (*False Negative- False Low Risk*). The set $\{P, N\}$ and the predicted risk set are used to build a confusion matrix (Kohavi & Provost, 1998; Domingos, 1998).

From the confusion matrix in table 1 the number of correct or incorrect (misclassified) patterns is generated as a measure of classifier Accuracy, where

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{Equation 5})$$

Table 1. A Confusion Matrix.

		Predicted classes	
		P (High risk)	N (Low risk)
Desired classes	P (High risk)	<i>TP</i>	<i>FP</i>
	N (Low risk)	<i>FN</i>	<i>TN</i>

Some related concepts, such as true positive rate (*tpr*), false positive rate (*fpr*), sensitivity (*sen*) and specificity (*spec*), can all be built from the confusion matrix. True positive rate (*tpr*) is the rate of the correct “*High risk*” number in the total correct prediction pattern number (including “*High risk*” and “*Low risk*”):

$$tpr = \frac{TP}{TP + TN} \quad (\text{Equation 6})$$

Conversely, false positive rate (*fpr*) is the rate of the incorrect “*High risk*” number in the total incorrect prediction pattern number:

$$fpr = \frac{FP}{FP + FN} \quad (\text{Equation 7})$$

The sensitivity rate (*sen*) indicates the effectiveness at identifying the true positive cases, and is calculated as:

$$sen = \frac{TP}{TP + FN} \quad (\text{Equation 8})$$

The specificity rate (*spec*) is the rate of correct “*Low risk*” number (*TN*) in the total of the *TN* and *FP* number:

$$spec = \frac{TN}{TN + FP} \quad (\text{Equation 9})$$

ROC graphs (Swets, 1988; Tom, 2006) are two-dimensional graphs in which the true positive rate is plotted on the Y axis and the false positive rate is plotted on the X axis. An ROC graph depicts relative tradeoffs between benefits (*tpr*) and costs (*fpr*). Figure 1 shows an example ROC graph. The pair of (*fpr*, *tpr*) for each classifier is represented as a point in the ROC space. It shows the benefit and the cost results of 5 classifiers labeled as A, B, C, D, and E. The diagonal line in the ROC graph presents the strategy of randomly guessing a class. Any classifier (E) that appears in the lower right triangle performs worse than random; conversely good classifiers (A, B, D) appear above the diagonal. The classifier labeled as D is *perfect* because of its *tpr*=1 and *fpr*=0. This means all input patterns are classified correctly; i.e. the prediction risks are the same as the prior classified outcomes. The classifier labeled as C yields the point (0.7; 0.7) in the ROC graph, meaning it guesses “*High risk*” (*tpr*=70%) at the same rate as its “*incorrect High risk*” (*fpr*=70%); in effect, equivalent to random.

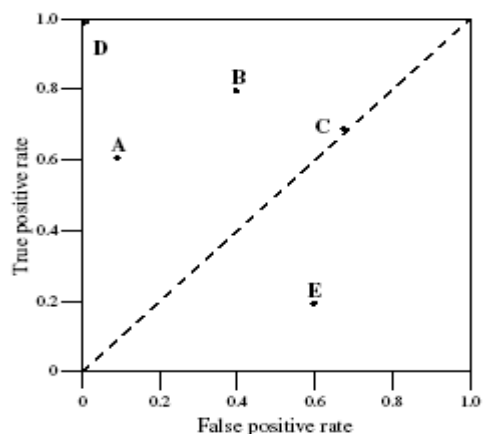


Figure 1. An example ROC graph.

POSSUM and PPOSSUM

Table 2 shows the confusion matrix, and classifier measures, for the two POSSUM risk scoring models, with the Mortality, Morbidity, and Death Rate Predictions from the models evaluated against the actuality from the patient data for individual patients. Note that the model data are derived from only one clinical site, and makes use of 499 records.

Table 2. Confusion matrix and classifier measures for POSSUM and PPOSSUM.

Model	Risk	High risk	Low risk	Total	Sen	Spec	tp rate	fp rate
<i>Mortality</i>	<i>High risk</i>	10	69	79	0.4	0.85	0.13	0.16
	<i>Low risk</i>	15	405	420				
	<i>Total</i>	25	474	499				
<i>Morbidity</i>	<i>High risk</i>	15	64	79	0.38	0.86	0.19	0.15
	<i>Low risk</i>	24	396	420				
	<i>Total</i>	39	460	499				
<i>Death rate</i>	<i>High risk</i>	10	69	79	0.38	0.85	0.13	0.16
	<i>Low risk</i>	16	404	420				
	<i>Total</i>	26	473	499				

There is very little difference between the three models and overall, all the POSSUM model predictors give poor results, with the exception of the specificity rate; demonstrating their limited reliability as predictors of high risk patients.

Heuristic Models and Neural Nets

In this section three supervised neural net techniques, Multi-Layer Perceptrons (MLP), Radial Basis Functions (RBF), and Support Vector Machines (SVM) (Haykin, 1999; Bishop, 2006) are used with the heuristic models described above (CM1, CM2, CM3a, CM3b, CM4a and CM4b). The available data was prepared according to the attribute input list and output labeling as given by the

various models. This resulted in six labeled data sets. Various training strategies were investigated, with varying parameters to identify the parameters that would result in the optimal neural classifier for these models. This experimentation determined that splitting the data into training, validation and test sets and then using bespoke neural net software (SNNS, 1995) gave no benefit over using 10-fold cross validation within a more general data mining package, WEKA (Witten & Eibe, 2005). Other data mining techniques, available within WEKA, were also investigated. These other classifiers, for example J48 Decision Tree or Naïve Bayesian classifiers (WEKA, 2007), offered no improvement on the supervised classifiers described in the rest of this chapter.

The data for CM1 and CM2 models includes 26 attributes with 839 patient records. Table 3 shows a confusion matrix for both models with the three supervised neural techniques. Overall, all classifiers give poor results especially in regard to the important *sensitivity* rates, with values typically less than 0.2. The classifier CM2-SVM has the poorest *sensitivity* rate (0.13), while CM1-SVM has the highest (0.27). Interestingly, all classifiers display very similar *specificity* rates (about 0.85). Specificity rates represent the outcomes labeled as "Low Risk", the negative risk prediction outcome in this domain. Both models, irrespective of the neural net used, give poorer Sensitivity but similar Specificity to the POSSUM models.

Table 3. Confusion matrix for CM1 and CM2 models with NN techniques.

Classifiers	Risk	High risk	Low risk	Total	Sensitivity	Specificity
CM1-MLP	High risk	19	107	126	0.15	0.86
	Low risk	45	668	713		
CM1-RBF	High risk	23	103	126	0.18	0.86
	Low risk	33	680	713		
CM1-SVM	High risk	34	92	126	0.27	0.86
	Low risk	99	614	713		
CM2-MLP	High risk	24	111	135	0.18	0.85
	Low risk	32	672	704		
CM2-RBF	High risk	24	111	135	0.18	0.85
	Low risk	32	672	704		
CM2-SVM	High risk	17	118	135	0.13	0.85
	Low risk	23	681	704		

The clinical risk prediction models CM3a and CM4a are two outcome models based on 16 (CM3a) and 14 (CM4a) input attributes. Table 4 shows the use of the three neural techniques (MLP, RBF, and SVM) on the same patient records as for CM1 and CM2 in table 3. Overall, all classifiers display identical *specificity* rates (about 0.85) with poor *sensitivity* rates (on average 0.18). The best of these classifiers (CM3a-SVM) gives poorer *sensitivity* performance than the POSSUM models. Again, similarly to the CM1 and CM2 results, these classifiers and models offer poor prediction results for identifying high risk patients.

Table 4. Confusion matrix for CM3a and CM4a models with NN techniques.

Classifiers	Risk	<i>High risk</i>	<i>Low risk</i>	<i>Total</i>	Sen	Spec
<i>CM3a-MLP</i>	<i>High risk</i>	26	109	135	0.19	0.85
	<i>Low risk</i>	79	625	704		
<i>CM3a-RBF</i>	<i>High risk</i>	20	115	135	0.15	0.85
	<i>Low risk</i>	24	680	704		
<i>CM3a-SVM</i>	<i>High risk</i>	32	103	135	0.24	0.85
	<i>Low risk</i>	111	593	704		
<i>CM4a-MLP</i>	<i>High risk</i>	28	107	135	0.20	0.85
	<i>Low risk</i>	69	635	704		
<i>CM4a-RBF</i>	<i>High risk</i>	19	116	135	0.14	0.85
	<i>Low risk</i>	17	687	704		
<i>CM4a-SVM</i>	<i>High risk</i>	20	115	135	0.15	0.85
	<i>Low risk</i>	49	655	704		

Table 5 shows the confusion matrix results from applying the neural techniques to clinical models CM3b and CM4b, using the same 839 patient records but with the expanded outcome scales as described above. It is clear that for many records the outcome labeled as "*Medium risk*" is misclassified as "*Low risk*". This might be due to the close distance between the classes labeled as "*Medium risk*" and "*Low risk*" in the prediction risk scales defined in the heuristic formulae.

Overall none of the existing risk prediction models, when combined with the supervised neural net techniques, offer sufficiently high sensitivity rates to be considered useful for regular clinical practice or investigated further in trial studies for that purpose. Experimentation with other supervised techniques, such as the many Decision Tree or Naïve Bayesian classifiers within WEKA, showed no improvement; and in many cases poorer performance.

An issue that came to light when discussing the results with the cardiovascular clinicians is an inherent bias in the given data. This bias arises from an interpretation of raw data from handwritten patient records as it was entered into the database. Judgments on interpreting values were made at that point. Unfortunately access to the raw, uninterpreted data is not possible. Problems associated with the gathering of the data are discussed further in the final sections of this chapter.

Table 5. Confusion matrix for CM3b and CM4b models with NN techniques.

Classifiers	Risk	<i>Very High risk</i>	<i>High risk</i>	<i>Medium risk</i>	<i>Low risk</i>
<i>CM3b-MLP</i>	<i>Very High risk</i>	3	3	0	14
	<i>High risk</i>	2	13	0	91
	<i>Medium risk</i>	0	1	0	12
	<i>Low risk</i>	6	53	5	636
<i>CM3b-RBF</i>	<i>Very High risk</i>	1	3	0	16
	<i>High risk</i>	3	6	2	95
	<i>Medium risk</i>	0	0	0	13
	<i>Low risk</i>	3	19	5	673
<i>CM3b-SVM</i>	<i>Very High risk</i>	0	2	0	18
	<i>High risk</i>	1	18	0	87
	<i>Medium risk</i>	0	0	0	13
	<i>Low risk</i>	9	98	10	583
<i>CM4b-MLP</i>	<i>Very High risk</i>	1	3	0	16
	<i>High risk</i>	2	18	0	86
	<i>Medium risk</i>	0	1	0	12
	<i>Low risk</i>	7	52	2	639
<i>CM4b-RBF</i>	<i>Very High risk</i>	1	4	1	14
	<i>High risk</i>	3	7	2	94
	<i>Medium risk</i>	0	1	0	12
	<i>Low risk</i>	3	11	2	684
<i>CM4b-SVM</i>	<i>Very High risk</i>	2	5	0	13
	<i>High risk</i>	4	12	0	90
	<i>Medium risk</i>	0	0	0	13
	<i>Low risk</i>	13	49	10	628

The use of otherwise reliable neural net (and other supervised classifier) techniques suggest that the input attribute set does not match well to the outcome label for any of these models. It is argued that this failure is due, in part, to the poor outcome labeling rules in the heuristic model rules and the poor transfer from general clinical practice to this specific domain for the POSSUM and PPOSSUM models. It is suggested that the input attribute set when used with no coerced outcome label may be sufficient to produce more reliable predictions for clinical risk. To do this requires the use of unsupervised classifier or machine learning techniques, where the inherent structure of the data is used to produce more natural outcomes (or clusters).

The Self Organizing Map (Kohonen, 1995) is one of the most popular unsupervised neural network models (Haykin, 1999). It is a competitive unsupervised learning network that produces a low dimensional output space where the input topological properties remain. The SOM provides a topology preserving mapping from the high dimensional input space onto the output map units. The topology preserving property means the mapping preserves the relative distance between points near each other in the input space, and nearby map units in the SOM. Hence, SOM can serve

as a cluster analyzing tool of high-dimensional data, with the ability to recognize and characterize the input data. The input attributes from model CM3b were used to create a SOM within Matlab (SOM toolbox, 2005), with a final U-matrix quantization error of 1.723, and topographic error of 0.021. The U-matrix was then sectioned into 4 (predetermined) clusters; each defined using final quantization error and topographic error. While this provided what appeared to be visually appropriate groupings, SOM offers an uncertainty as a clustering algorithm, as alternative runs might offer alternative clustering results. Furthermore specific cases could not be easily identified so negating the use of SOM for individual patient risk assessment. So alternative means of performing unsupervised learning (or clustering) were investigated. The following sections detail this.

Clustering with KMIX

Partitioning is a fundamental operation in data mining for dividing a set of objects into homogeneous clusters. Clustering is a popular partitioning approach. A set of objects are placed into clusters such that objects in the same cluster are more similar to each other than objects in other clusters according to some defined criteria. The K-means algorithm (Kanungo et al., 2002) is well used for implementing this operation because of its efficiency in clustering large data sets. However, K-means only works on continuous values. This limits its use in medical domains where data sets often contain Boolean, categorical, and continuous data. The traditional approach to convert categorical data into numeric values does not necessarily produce meaningful results where categorical attributes are not ordered. KMIX, and later WKMIX, are improved from K-means in order to cluster mixed numerical and categorical data values. In the KMIX algorithm, a dissimilarity measure is defined that takes into account both numeric and categorical attributes via the Euclidean distance for numerical features and the number of mismatches of categorical values for discrete features. For example, assume that $d^N(X,Y)$ is the squared Euclidean distance between two objects X and Y over continuous features; and $d^C(X,Y)$ is the dissimilarity measure on categorical features in X, Y . The dissimilarity between two objects X, Y is given by the distance

$$d(X, Y) = d^N(X, Y) + d^C(X, Y) \quad (\text{Equation 10})$$

The clustering process of the KMIX algorithm is similar to the K-means algorithm except that a new method is used to update the categorical attribute values of cluster. The motivation for proposing KMIX based on K-means is that KMIX can be used for large data sets, where hierarchical clustering methods are not efficient.

Clustering and similarity measures

Cluster analysis provides the means for the organization of a collection of patterns into clusters based on the similarity of these patterns, where each pattern is represented as a vector in a multidimensional space. Assume that X is a pattern (an observation or sample from a data set). X typically consists of m components, represented in multidimensional space as:

$$X = (x_1, x_2, \dots, x_m) = (x_j) \quad j = 1, \dots, m$$

Each component in multidimensional space is called a feature (attribute). A data set includes n patterns X_i where $i \in [1, n]$ and $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$; forming a $n \otimes m$ pattern matrix. Note that the m features here may include continuous and discrete valued features. Due to the variety of feature types and scales, the distance measure (or measures) must be chosen carefully (Jain, 1999; Bishop, 2006). It is most common to calculate the dissimilarity between two patterns using a distance measure defined on the feature space.

A similarity measurement is the strength of the relationship between two patterns in the same multidimensional space. It can be represented as some function of their observed values such as

$$sim_{i,j} = sim(x_i, x_j) \quad i, j \in [1, n]$$

Similarity is regarded as a symmetric relationship requiring $sim(x_i, x_j) = sim(x_j, x_i)$ (Gower, 1988). However, the dissimilarity measure of patterns has been introduced as the complement of similarity measures. A list of dissimilarity measures can be seen in (Gower, 1985). For continuous features, the most common used measure is the Euclidean distance between two patterns. This is very dependent upon the particular scales chosen for the variables (Everitt, 1994). Typically all (numeric) features are transformed to the range $[0, 1]$, so as to avoid feature bias.

The dissimilarity measure of two “continuous” patterns using Euclidean distance is given as:

$$dissim(x_i, x_j) = [D(x_i, x_j)]^2 = \sum_{k=1}^m (x_{ik} - x_{jk})^2, i, j \in [1, n_1], n_1 \leq n \quad (\text{Equation 11})$$

where D is the Euclidean distance between x_i and x_j .

This means the dissimilarity of two patterns x_i and x_j is the sum of the square of the feature distance between them.

For discrete features, the similarity measure between two patterns depends on the number of similar values in each categorical feature (Kaufman & Rousseeuw, 1990). This means the dissimilarity will be the number of different values between two patterns for each categorical feature. We can represent this dissimilarity in the following formula:

$$dissim(x_i, x_j) = d(x_i, x_j) = \sum_{k=1}^m \theta(x_{ik}, x_{jk}) \quad i, j \in [1, n_2], n_2 \leq n \quad (\text{Equation 12})$$

$$\text{where } \theta(x_{ik}, x_{jk}) = \begin{cases} 0 & \text{if } x_{ik} = x_{jk} \\ 1 & \text{if } x_{ik} \neq x_{jk} \end{cases}, \quad k = 1, 2, \dots, m; i, j \in [1, n_2]$$

For binary features, the dissimilarity measures are calculated as for either discrete (categorical) features or continuous (numeric) valued attributes dependent on the interpretation of the provided binary data.

Clusters and centre vectors

For each cluster to be produced, the cluster defining vector is referred to as the centre vector. Here, the centre vector will include two groups of continuous and discrete components as the data feature set includes both continuous and discrete features; binary features are treated as continuous or discrete dependent upon the interpretation of each binary-valued attribute. Assume that the data feature set includes m features, where the p first features are continuous features and the $m-p$ remaining features are discrete. This means each pattern X_i in the space can be seen as

$$X_i = (x_{i1}, x_{i2}, \dots, x_{ip}, x_{ip+1}, x_{ip+2}, \dots, x_{im}).$$

Assume that Q_j is a centre vector for the data set cluster C (C is a sub set of whole data set). So Q_j can be represented as

$$Q_j = (q_{j1}, q_{j2}, \dots, q_{jp}, q_{jp+1}, q_{jp+2}, \dots, q_{jm}).$$

The task is to find p *continuous* component values, and $m-p$ *discrete* component values for vector Q_j . For *continuous* component values, $\{q_{jk}\}_{k=1, \dots, p}$ defines the means of the k^{th} feature in C (Hand, 1981). For *discrete* component values, $\{q_{jk}\}_{k=p+1, \dots, m}$ defines the set of $mode_k$, where $mode_k$ is the mode of the k^{th} feature.

Definition 1: A vector Q is a mode vector of a data set

$$C = (X_1, X_2, \dots, X_c), c \leq n$$

if the distance from each vector X_i ($i \in [1, c]$) is minimized. This means

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

is minimized. Huan (1998) proved that this distance will be minimized only if the frequency of value $q_j \in Q$ is maximized. This means the frequency of each value q_j in data set C , considered in terms of feature j , needs to be greater or equal to the frequency of all different $x_{i,j}$ such that $x_{i,j} \neq q_j$ for the same feature ($j \in [1, m]$). Hence we can choose the mode vector for the $m-p$ categorical components where each component value is the mode of that feature or the value which has biggest frequency value in that feature using:

$$\{q_{jk}\}_{k=p+1, \dots, m} = mode_k = \{\max \text{freq}(\text{Val}_{ck})\}.$$

The K-MIX algorithm

The K-MIX algorithm is a four step process:

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

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

$$Q_j = (q_{j1}^N, q_{j2}^N, \dots, q_{jp}^N, q_{jp+1}^C, \dots, q_{jm}^C), j \in \{1, 2, \dots, K\}$$

where $\{q_{jk}^N\} \in \{1, p\} \Rightarrow \{\text{mean}_{jk}^N\}$ (mean of k^{th} feature in cluster j);

and $\{q_{jk}^C\} \in \{p+1, m\} \Rightarrow \{\text{mode}_{jk}^C\}$ (mode of k^{th} feature in cluster j);

Step 3: Update clusters:

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

$$d(X_i, Q_j) = d^N(X_i, Q_j) + d^C(X_i, Q_j); j = 1, 2, \dots, K \text{ (Equation 13)}$$

where $d^N(X_i, Q_j)$ is calculated according to (Equation 11),

and $d^C(X_i, Q_j)$ is calculated according to (Equation 12).

Allocate X_i into the nearest cluster such that $d(X_i, Q_j)$ is least.

Do this for whole data set, and save them to the new interpretation of the data set with K new centre vectors.

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

Experiments with standard data sets

Before running on the target data domain, many experiments were run with data derived from the UCI repository of databases as used by the machine learning community for the empirical analysis of machine learning algorithms (Merz & Murphy, 1996). The clustering accuracy for measuring the clustering results was computed as follows. Given the final number of clusters, K , clustering accuracy r was defined as:

$$r = \sum_{i=1}^K a_i / n$$

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. Consequently, the clustering error is defined as $e = 1 - r$. The lower value of e suggests the better clustering result.

The experimental data sets are Small Soybean data set (Michalski & Chilausky, 1980) with 47 samples and 35 attributes, in 4 class distributions, Votes data set (Jeff, 1987) containing 16 key attributes with all categorical data types in 435 records (included meaningful missing value records "?"), and 2 output classes labelled to 168 republicans and 267 democrats. This algorithm is also

used for experiments with Zoo small data set (Merz & Murphy, 1996). It has 101 records distributed in 7 categories with 18 attributes (included 15 Boolean, 2 numerical, and 1 unique attribute(s)). The fourth experiment for KMIX is with the Wisconsin Breast Cancer data set (Merz & Murphy, 1996). It contains 683 records by removing 16 missing value records. The data set includes 9 numerical attributes divided into 2 class label of “2” or “4”. The comparison results can be seen in Table 6.

Table 6. Comparison of KMIX to publication results on standard data sets (see text for explanation of labels).

Data set	Publication results	KMIX results
Soy Bean	0.11 ¹ ~	0.07
Votes	0.132 ^{2,3}	0.141
Zoo small	0.166 ²	0.151
Wisconsin Breast Cancer	0.03 ⁴ ; 0.132 ²	0.03

From Table 6, KMIX performs as well as other published results for the Soy Bean¹ [Ohn et al., 2004]; Votes² (Shehroz & Shri, 2007), Votes³ (Zengyou et al., 2005), and Wisconsin Breast Cancer⁴ (Camastra & Verri, 2005) data sets. For the latter, the KMIX result of 0.03 compares favourably compared to 0.132² (Shehroz & Shri, 2007). Further more, this algorithm solves problems associated with the mixture of categorical and discrete data; a common feature of medical data domains.

Experiments in the Clinical Domain

The research project requires that a comparative audit of the data for different outcomes to be investigated. Patient parameters such as “*Patient Status*”, and the combination of other risk outcomes, such as “*Heart Disease*” (HD), “*Diabetes*” (D), and “*Stroke*” (St) may all be used as outcome indicators for individual patients. Subsequently a new summary output attribute (*Risk*) is built based on the value for the combination of the main disease symptoms. For alternative outcomes the appropriate models are built based on different heuristic rules:

- Model 1 (CM32): Two outcome levels are defined as:

$$\Sigma(\text{Status}, \text{Combine}) = 0 \rightarrow \text{Risk} = \text{Low}$$

$$\Sigma(\text{Status}, \text{Combine}) > 0 \rightarrow \text{Risk} = \text{High}$$

- Model 2 (CM33): Similar to CM32 but divided into three levels of risk:

$$\Sigma(\text{Status}, \text{Combine}) = 0 \rightarrow \text{Risk} = \text{Low}$$

$$\Sigma(\text{Status}, \text{Combine}) = 1 \rightarrow \text{Risk} = \text{Medium}$$

$$\Sigma(\text{Status}, \text{Combine}) > 1 \rightarrow \text{Risk} = \text{High}$$

These outcomes will be used to provide meaningful names to the clusters generated using the KMIX algorithm, and its variation WKMIX.

The next experiment used the K-means algorithm for comparison with KMIX on the cardiovascular data. Inputs were defined as the attribute set from the heuristic model CM3a; the best performing of the existing models. Hence these experiments were run using 16 input attributes, with 341 patient records. The results, in Table 7, show the derived clusters compared to the existing outcomes, and so indicate the extent of agreement between the unsupervised clustering algorithms and the existing clinical outcomes. The sensitivity rates for the two algorithms are small (0.15, 0.25 for K-means, and KMIX), with high specificity rates. However Table 7 clearly shows the advantage of KMIX over K-means, with it also improving over the best of the supervised classifiers (*CM3a-SVM*) in Table 4.

Table 7. Clustering results of K-means and KMIX compared to Model CM3a Outcomes.

Algorithm	Risk	C1 (High)	C2 (Low)	Sensitivity	Specificity
K-means	High	36	21	0.15	0.82
	Low	168	116		
KMIX	High	35	22	0.25	0.89
	Low	107	177		

Table 8. Supervised NN results for KMIX generated CM32 outcomes.

Classifier	Cluster	C1 (High)	C2 (Low)	Sensitivity	Specificity
MLP	C1H	121	21	0.90	0.90
	C2L	13	186		
SVM	C1H	120	22	0.85	0.89
	C2L	22	177		

Table 9. Supervised NN results for KMIX generated CM33 outcomes.

Classifier	Cluster	C1 (Medium)	C2 (High)	C3 (Low)	Sensitivity	Specificity
MLP	C1M	75	4	6	0.98	0.96
	C2H	5	112	0		
	C3L	4	0	135		
SVM	C1M	71	0	14	0.98	0.91
	C2H	8	109	0		
	C3L	3	1	135		

It is possible using KMIX to define the number of clusters required. Again, the input attribute set from clinical model CM3 was used, with the algorithm allowed to generate a two cluster output (model CM32) and a three cluster output (model CM33). The records associated with each cluster were cross-referenced to the Low, High (for CM32) and Low, Medium, High outcomes (for model CM33) associated with the clinical models. Supervised neural network techniques, Support Vector Machine (SVM), and Multilayer Perceptron (MLP), were then trained on these machine generated outcomes. Tables 8 and Table 9 show the results.

From Tables 8 and 9 it can be deduced that the boundary for each cluster may be ambiguous. However in Table 9 no High Risk patients were placed in the cluster most closely associated with Low Risk patients (C3). Fortunately in the cardiovascular domain, and the given data mining task, the clinician's interest is primarily with the identification of both medium and high risk patients. These cases can be reported in terms of the sensitivity rate. In both table 8 and 9 both sensitivity and specificity rates are over 0.90 except for the CM32-SVM classifier (0.85; 0.89 respectively). The clinicians have an accepted rate of 0.80 for the use of neural classifiers; all classifiers achieve this. This suggests that the KMIX clustering results show some promise for the identification of risk for individual patients in the cardiovascular data domain. To better these results, an improved version was developed as explained in the next section.

Clustering with WKMIX

WKMIX (Weighted KMIX), an extension to the KMIX algorithm described above, makes use of an entropy based measure to weight the input attributes. In the previous section, KMIX was described working with inputs where the attribute set was defined using the existing, acknowledged as flawed, risk prediction models. An alternative approach is to take the complete dataset and determine what are the more appropriate attributes to use as classifier and clustering inputs.

Feature selection is a very important step in classification (Liu & Motoda, 1998). It can reduce the irrelevant and redundant features, which often degrade the performance of classification algorithms in both speed and prediction accuracy. Most feature selection methods use certain evaluation functions and search procedures to achieve their targets. The evaluation functions measure how good a specific feature subset is in discriminating between the classes (Dash & Liu, 1997). Here a variation on Mutual Information is used. In this work, features determined to be irrelevant, using MI, are automatically given a low (near zero) coefficient; so reducing their impact on the clustering dissimilarity measure. Experimentation has shown that this technique is as reliable as the more common Relief feature selection techniques (Kira & Rendell, 1992; Kononenko, 2001).

Mutual Information

Mutual Information (MI) measures the arbitrary dependencies between random variables. It is suitable for assessing the "information content" of the attribute contributions to the outcome classes in data domain. In information theory, Shannon (1948) defined entropy or information entropy as a measure of the uncertainty associated with a discrete random variable. Equivalently, entropy is a measure of the average information content the recipient is missing when they do not know the value of the random variable. Mathematically, entropy can be written as:

$$H(X) = -c \sum_{i=1}^n p_i(x) \log p_i(x) \text{ (Equation 14)}$$

where $p_i(x)$ is probabilities of occurrence in a set of possible events X, and c is a positive constant (usually c=1).

Suppose there are two events X and Y . $H(X,Y)$ is a joint entropy of two discrete variables X and Y with a joint distribution $p(x,y)$, and $H(X,Y)$ can be written as the follows:

$$H(X,Y) = -c \sum_{i=1}^n \sum_{j=1}^m p_{i,j}(x,y) \log p_{i,j}(x,y) \quad (\text{Equation 15})$$

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

The Mutual Information (MI) between two 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. Whereas entropy is a measure of the uncertainty in its distribution (Shannon, 1948), the relative entropy (Equation 16) is a measure of the statistical distance between two distributions. Originally introduced by Kullback and Leibler (1951), it is known as the Kullback Leibler distance, or Kullback Leibler divergence (Cover & Thomas, 1991).

$$K(p,q) = \sum_{x \in A} p(x) \log \left(\frac{p(x)}{q(x)} \right) \quad (\text{Equation 16})$$

where $p(x)$, $q(x)$ are probability density functions of x .

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

$$MI(X,Y) = H(X) - H(X|Y) = \sum_i \sum_j p_{i,j}(x,y) \log [p_{i,j}(x,y) | p_i(x) p_j(y)] \quad (\text{Equation 17})$$

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

The concepts of entropy can be extended to the continuous random variables (Shannon, 1948; Cover & Thomas, 1991). So $MI(X,Y)$ for the continuous valued case can be written as:

$$MI(X,Y) = H(X) - H(X|Y) = \int p_{i,j}(x,y) \log \left(\frac{p_{i,j}(x,y)}{p_i(x) p_j(y)} \right) dx dy \quad (\text{Equation 18})$$

Based on the Kullback Leibler divergence, in equation 16, equations 17 and 18 can be rewritten as

$$MI(x,y) = K(P(x), P(x).P(y)) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \cdot \log \frac{P(x,y)}{P(x).P(y)} \quad (\text{Equation 19})$$

and

$$MI(x,y) = K(P(x), P(x).P(y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x,y) \cdot \log \frac{P(x,y)}{P(x).P(y)} dx dy \quad (\text{Equation 20})$$

The Weighted KMIX Algorithm (WKMIX)

This algorithm is improved from KMIX algorithm by putting the weight for each attribute during the cluster process in KMIX algorithm. The steps of the algorithm are the same as the KMIX algorithm described above, and equation 13 is rewritten as:

$$d(X_i, Q_j) = W_{iN}d^N(X_i, Q_j) + W_{iC}d^C(X_i, Q_j); j=1,2,..k \text{ (Equation 21)}$$

where W_{iN} , W_{iC} are the MI of each numerical, or categorical attribute, and all the i^{th} , j^{th} , and k^{th} remain unchanged in their meaning.

The MI for Equation 20 is calculated as in Equation 22.

$$MI(C, x_j) = \sum_{i=1}^c \sum_{k=1}^s p_{ijk} \log \left(\frac{p_{ijk}}{q_i \cdot r_{jk}} \right) \text{ (Equation 22)}$$

where $p_{ijk} = \frac{\text{sum}_{ijk}}{\text{sum}}$ with sum_{ijk} the number of patterns in class C_i with attribute x_j in state k .

$$q_i = \frac{\text{sum}_i}{\text{sum}} \text{ where } \text{sum}_i \text{ is number of patterns belong to class } C_i$$

$$r_{jk} = \frac{\text{sum}_{jk}}{\text{sum}} \text{ where } \text{sum}_{jk} \text{ is number of patterns with attribute } x_j \text{ in the } k^{\text{th}} \text{ state.}$$

Hence Equation 21 is finally rewritten as:

$$d(X_i, Q_j) = MI_{iN}d^N(X_i, Q_j) + MI_{iC}d^C(X_i, Q_j); j=1,2,..k \text{ (Equation 23)}$$

The experimental data used here is identical to that used in comparing the K-Means and KMIX algorithms, using the inputs attributes from risk prediction model CM3a. The results from the weighted KMIX algorithm (WKMIX) are compared to the KMIX algorithm from Table 7 (for the CM3a model), and given in Table 10. The WKMIX outcome models are then derived from the clustering results. These models are then investigated with the use of supervised NN techniques.

General speaking, table 10 shows that the experimental performance of using an Information Theory based weight in clustering is higher than without it. In particular, the sensitivity and specificity rates for using weights (WKMIX) are the highest of any model and classifier combination in this study (1, 0.96). They better the use of the clustering algorithm without attribute weight (KMIX: 0.25, 0.89), and higher than the rate from clinical expert's advice (about 80%).

Table 10. The results of alternative weights for CM3a model.

Algorithm	Output	High risk	Low risk	Sensitivity	Specificity
WK MIX	High risk	48	9	1	0.96
	Low risk	0	284		
KMIX	High risk	35	22	0.25	0.89
	Low risk	107	177		

The new clustering model (CM3aC) is built with new outcomes derived from the clusters produced using the WK MIX algorithm. The NN techniques applied here are Support Vector Machine (SVM); Radial Basis Function (RBF); and MultLayer Perceptron (MLP). More over the decision tree technique J48 is also applied to this new model. The results can be seen in table 11. This shows that there are no errors in the classification process (MSE =0.00) except with the use of the RBF technique (0.08). Moreover, all classifiers achieve very high *sensitivity* and *specificity* rates.

Table 11. The results of alternative techniques for the CM3aC clustering model.

Classifier	Risk	High risk	Low risk	MSE	Sens	Spec	Accuracy
SVM	High risk	48	0	0.00	1	1	100%
	Low risk	0	293				
RBF	High risk	46	2	0.08	0.97	0.99	99%
	Low risk	1	292				
MLP	High risk	48	0	0.00	1	1	100%
	Low risk	0	293				
J48	High risk	48	0	0.00	1	1	100%
	Low risk	0	293				

Looking at the confusion matrix in table 11, it is clear that all techniques classified the risks the same as the clustering results except RBF with 2 mis-classed "*High risk*" and 1 mis-classed "*Low risk*" respectively. Interestingly, from table 11 the sensitivity and specificity rates achieve ideal results except for the Radial Basis Function net (0.97 and 0.99 respectively). This means the neural techniques (SVM and MLP) and the decision tree method (J48) achieve 100% accuracy.

Although this study achieves ideal results from a data mining perspective, its acceptance as a clinical risk model is problematic. The use of MI values as algorithm weights achieves a high performance in the case study, albeit on attributes associated with an existing clinical heuristic model. An extended investigation with an expanded data set is required, allowing an increased number of attributes and patterns. Such a study is perhaps best undertaken in closer collaboration with the consultant clinicians within the aegis of a clinical study to derive a more reliable, medically acceptable, risk prediction model.

Discussion

The paucity of the current risk prediction models, in cardiovascular medicine, are made evident in this case study, with none of the existing individual patient risk prediction models being able to

better 27% sensitivity (Support Vector Machine classifier with clinical model CM1 in Table 3), although many of the models present 85% Specificity. The POSSUM and PPOSSUM models better this but with the disadvantage of losing single patient risk prediction capability. The clinicians working in this field of medicine are well aware of these limitations in their current models (Kuhan et al, 2001). However the clustering algorithms which use the input attribute sets from these models (and WKMIX in particular) manage to produce very much higher Sensitivity and Specificity. This suggests that the outcome models in the existing clinical risk prediction models are inadequate. Furthermore it suggests that much better risk prediction models are possible and that the present attributes offer sufficient inherent patterns, as detected using these clustering algorithms, to support much better risk prediction rules.

The proposed algorithm KMIX compares favourably to publicised results on standard machine learning data sets. Furthermore, because this algorithm was developed for use with a specific medical data domain, it meets the requirement to work with data that contains a mixture of numerical, categorical and Boolean data types. By using the clustering algorithm, new outcomes are generated for input attributes associated with the heuristic CM32 and CM33 risk models. These new models are evaluated through the use of the neural network techniques, such as MLP and SVM. From Table 8 and 9 we can see that the boundary of each cluster is not clear. For example, in Table 8, a number of high risk patients (C1H) fell in clusters C1 and C2. However no high risk patients were placed in the cluster most closely associated with low risk patients (C3 in Table 9). Fortunately in the cardiovascular domain, the clinician's interest is in the rate of potential risk of patients (medium and high risk). Further work on improving this algorithm allowed weights to be applied in order to indicate the level of importance for attributes in the feature set for the data domain. This algorithm gives near ideal classifier performance on the given clinical data. However, alternative data domains need to be investigated to determine the usefulness of this algorithm for medical data domains in general.

The history of expert and knowledge base systems contains many examples of intelligent decision systems used in medical domains (Davis et al., 1977; Davies & Owen, 1990; Miller et al., 1982; Shortliffe, 1990). More recently decision support tools that make use of data mining techniques have been developed for medicine (Lavrač et al., 2000; Groselj, 2002; Lavrač & Zupan, 2005). The combination of clinical knowledge and the approach adopted in this case study via fuzzy set theory might produce a realistically more intelligent tool for the risk prediction process (Gorzalczany & Piasta, 1999; Negnevitsky, 2004). Moreover, the use of neuro-fuzzy classification techniques might enhance the results; enabling multiple models to be used to suit the needs of the clinician, and make for more reliable risk prediction in a clinical situation. For this to happen, a number of important and time-consuming hurdles, relating to clinical trials, have to be met. Most of these hurdles are the constraints imposed by a highly regulated health system. Some are the results of ill-informed clinical data practices. The latter are particularly problematic where paper based data collection (i.e. patient records) are prevalent. In an effort to simplify subsequent human decision making, much data is interpreted as it is transferred to computer record. This means that instead of access to raw data (and many meaningful data values), many clinical attributes are reduced to a small number of clinical interpretations. The interpretation can vary from one region to another and between clinicians, according to their school of thought on best medical practice. This form of interpretation introduces a bias into the data model that cannot be overcome by any data mining

technique, other than collection of the original raw values. Such a bias in the data will cause problems for even the most intelligent of decision systems.

Conclusion

This case study showed the advantages of using Pattern Recognition techniques in producing risk models. However, these techniques are poor in providing a visualization of the data domain. Although the Self Organising Feature Map (SOM) showed its ability to visualize the data, its clustering was limited. Further research with a combination of KMIX or WKMIX and SOM might provide meaningful insights into the nature of the data, cluster models and medical outcomes.

The concept of MI is introduced and discussed in depth. From there, a combination of Bayes theorem and MI produced a new formula to calculate the MI between the attributes and the outcome classes (Equation 22). This idea has been used for alternative data mining purposes (water quality ascription) as a measurement for SOM clustering (O'Connor & Walley, 2000). Here it is used for ranking the significance of attributes in the cardiovascular data domain. The rewriting of the KMIX algorithm by using the MI weights provides a new clustering algorithm. Although its results show a very high performance level, more investigation was needed. The use of MI values as the weights in the clustering algorithm (WKMIX) has given better initial results. However, more experiments with alternative sources of data need to be investigated. The building of an independent feature selection algorithm using MI is another further task. This will be one of the required tools for the classification process, in particular with the current data domain.

This research has shown how Pattern Recognition and Data Mining techniques (both supervised and unsupervised) can better the current risk prediction models in cardiovascular data. The research can be seen as setting the base knowledge, with a set of practical decision metrics, for a complete decision support tool in the cardiovascular area. The research limitations for its use require a more complete ontology for the medical domain, involving general medical knowledge. Furthermore, it cannot be used in trials for clinical diagnosis, without ethical clearance. So, the supervised and unsupervised results in these case studies can not, as yet, be used to support the clinicians in their decision making for individual cardiovascular patients. There needs to be further collaborative research, between clinicians and computer scientists, to build a more complete decision support tool for the cardiovascular area. A complete decision tool is one that encompasses every stage of the decision process from the collection of raw data to feedback on predictions. Making the clinicians aware of their data gathering practices, and the effect it has on subsequent computer based decision making is an issue that will have a great impact on the feasibility of such collaborations. Hopefully, the move to collecting raw data directly onto computer databases rather than relying on the interpretation of data in paper based systems may help to reduce clinical bias. The same is probably true for many other clinical domains. Given this proviso, the future for data mining domain improving risk prediction in medical domains looks very promising.

Acknowledgment

The majority of the work reported here is undertaken in conjunction with PhD thesis research at the Computer Science Department in The University of Hull. Funding for the PhD studentship to Ms

Nguyen is from the Government of Vietnam, with further travel bursaries from The Clinical Biosciences Institute, University of Hull. We are grateful for the clinical input from cardiovascular consultants, Dr Ganesh Kuhan and Professor Peter McCollum of the Hull Royal Infirmary.

References

- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Camastra, F. & Verri, A. (2005). A novel Kernel Method for Clustering. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 27(5), 801-805.
- Copeland, G.P. (2002). The POSSUM system of surgical audit. *Archives of Surgery*, 137, 15-19.
- Copeland, G.P., Jones, D. & Walters, M. (1991). POSSUM: a scoring system for surgical audit. *British Journal of Surgery*, 78, 355-360.
- Cover, T. & Thomas, J. (1991). *The Elements of Information Theory*. New York: Plenum Press.
- Dash, M. & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(3).
- Davies, M. & Owen, K. (1990). Complex uncertain decisions: medical diagnosis. Case Study 10 in *Expert System Opportunities from the DTI's Research Technology Initiative*, HMSO.
- Davis, R., Buchanan, B.G. & Shortliffe, E.H. (1977). Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence*, 8, 15-45.
- Domingos, P. (1998). How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications. *Proceedings of the AAAI-98/ICML-98 Workshop on the Methodology of Applying Machine Learning* (pp. 1-7), Madison, WI: AAAI Press.
- Dunham, M.H. (2002). *Data Mining: Introductory and Advance Topics*. Upper Saddle River, NJ : Prentice Hall/Pearson Education.
- Everitt, B.S. (1994). *Cluster Analysis*, 3rd ed. John Wiley & Son, New York.
- Gorzalczany, M.B. & Piasta, Z. (1999). Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support. *Information Sciences*, 120(1), 45-68.
- Gower, J.C. (1985). Measure of similarity, dissimilarity and distance. In: *Encyclopedia of Statistical Sciences*, Vol. 5. John Wiley & Son, New York.
- Gower, J.C. (1988). Classification, geometry and data analysis. In H.H. Bock, (Ed.), *Classification and Related Methods of Data Analysis*. Elsevier, North-Holland, Amsterdam.
- Groselj, C. (2002). Data Mining Problems in Medicine. *15th IEEE Symposium on Computer-Based Medical Systems (CBMS'02)*. Maribor, Slovenia.

- Gunning, K. & Rowan, K.M. (1999). ABC of intensive care: outcome data and scoring systems. *British Medical Journal*, 319, 241-244.
- Hand, D.J. (1981). *Discrimination and Classification*. John Wiley & Sons.
- Haykin, S. (1999). *Neural networks: A comprehensive foundation*, 2/e, Macmillan College Publishing Company, Inc.
- Huan, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3). Kluwer Academic Publishers
- Jain, A.K. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3).
- Janet, F. (1997). Artificial Neural Networks Improve Diagnosis of Acute Myocardial Infarction. *Lancet*, 350(9082), 935.
- Jeff, S. (1987). *Concept acquisition through representational adjustment*. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA
- Kanungo, T., Mount, D.M., Netanyahu, N., Piatko, C., Silverman, R. & Wu, A.Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24, 881-892.
- Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data—An Introduction to Cluster Analysis*. Wiley.
- Kira, K. & Rendell, L.A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *AAAI-92: Proceedings of the Ninth National Conference on Artificial Intelligence*, 129-134. AAAI Press.
- Knaus, W.A., Draper, E.A., Wagner, D.P. & Zimmerman, J.E. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine*, 13, 818-829.
- Knaus, W.A., Wagner, E.A., Draper, J.E., Zimmerman, M., Bergner, P.G., Bastos, C.A., Sirio, D.J., Murphy, D.J., Lotring, T. & Damiano, A. (1991). The APACHE III prognostic system: risk prediction of hospital mortality for critically hospitalised adults. *Chest*, 100, 1619-1636.
- Kohavi, R. & Provost, F. (1998). Glossary of Terms. *Machine Learning*, 30(2/3), 271-274.
- Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.
- Kononenko, I. (2001). Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artificial Intelligent Medicine*. 23(1), 89-109.

- Kononenko, I. & Kukar, M. (2007). *Machine Learning and Data Mining*. Horwood Publishing Ltd.
- Kuhan, G., Gadiner, E.D., Abidia, A.F., Chetter, I.C., Renwick, P.M., Johnson, B.F., Wilkinson, A.R. & McCollum, P.T. (2001). Risk Modelling study for carotid endarterectomy. *British Journal of Surgery*, 88, 1590-1594.
- Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*. 22, 79-86.
- Lavrač, N., Keravnou, E. & Zupan, B. (2000). Intelligent Data Analysis in Medicine. In A. Kent et al., (Eds.). *Encyclopedia of Computer Science and Technology*, Vol.42, 113-157, Dekker, New York.
- Lavrač, N. & Zupan, B. (2005). Data Mining in Medicine. In O. Maimon & L. Rokach (Eds.). *Data Mining and Knowledge Discovery Handbook*, Springer US.
- Lisboa, P.J.G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Network*, 15, 11-39.
- Liu, H. & Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic, Norwell, MA USA.
- Merz, C.J. & Murphy, P. (1996). UCI Repository of Machine Learning Database. Available: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- Michalski, R.S. & 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*, 4(2), 125-161.
- Miller, R.A., Pople Jr., H.E. & Myers, J.D. (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine*, 307, 468-676.
- Negnevitsky, M. (2004). Design of a hybrid neuro-fuzzy decision-support system with a heterogeneous structure. *Proceedings IEEE International Conference on Fuzzy Systems, 2004*. Vol. 2, 1049-1052.

- O'Connor, M.A. & Walley, W.J. (2000). An information theoretic self-organising map with disaggregation of output classes. *2nd Int. Conf. on Environmental Information systems, Stafford, UK*. 108-115. ISBN 9 72980 501 6
- Ohn, M.S., Van-Nam, H. & Yoshiteru, N. (2004). An alternative extension of the K-means algorithm for clustering categorical data. *International Journal Mathematic Computer Science*, 14(2), 241-247.
- Prytherch, D.R., Sutton, G.L. & Boyle, J.R. (2001). Portsmouth POSSUM models for abdominal aortic aneurysm surgery. *British Journal of Surgery*, 88(7), 958-63.
- Rowan, K.M., Kerr, J.H., Major, E., McPherson, K., Short, A. & Vessey, M.P. (1994). Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. *Critical Care Medicine*, 22, 1392-1401.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, 379-423 and 623-656.
- Shehroz, S.K. & Shri, K. (2007). Computation of initial modes for K-modes clustering algorithm using evidence accumulation. *20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, India.
- Shortliffe, E.H. (1990). Clinical decision-support systems. In Shortliffe, E.H., Perreault, L.E., Wiederhold, G. & Fagan, L.M. (Eds.). *Medical informatics - Computer Applications in Health Care*, Addison-Wesley, Reading, M.A.
- Silipo, R. & Marchesi, C. (1998). Artificial Neural Networks for automatic ECG analysis. *IEEE Transactions on Signal Processing*, 46(5), 1417-1425.
- SNNS (1995). *Stuttgart Neural Network Simulator (Version 4.1)*, University of Stuttgart, Germany. Download free at <http://www.nada.kth.se/~orre/snns-manual>. Last accessed November 2007.
- SOM Tool Box (2005). *A function package for Matlab 5 implementing the Self-Organizing Map in Matlab*, Adaptive Informatics Research Centre, Helsinki University of Technology. Finland. Download free at <http://www.cis.hut.fi/projects/somtoolbox/>. Last accessed November 2007.
- Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.

- Tom, F. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27. Science Direct, Elsevier.
- WEKA (2007). *Weka Machine Learning Project*, Weka Software (Version 3.4.5), University of Waikato, New Zealand. Download free at <http://www.cs.waikato.ac.nz/ml/weka/>. Last accessed November 2007.
- Whiteley, M.S., PRYTHERCH, D.R., Higgins, B., Weaver, P.C. & Prout, W.G. (1996). An evaluation of the POSSUM surgical scoring system. *British Journal of Surgery*, 83, 812-815.
- Witten, I.H. & Eibe, F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2/e.
- Yii, M.K. & Ng, K.J. (2002). Risk-adjusted surgical audit with the POSSUM scoring system in a developing country. *British Journal of Surgery*, 89, 110-113.
- Zengyou, H., Xiaofei, X. & Shengchun, D. (2005). TCSOM: Clustering Transactions Using Self-Organizing Map. *Neural Processing Letters*, 22, 249–262.

Key terms

Centre vectors: For each cluster to be produced, the cluster defining vector (the idealized data entry for that cluster) is referred to as the centre vector.

Clustering: Clustering is a popular partitioning approach, whereby a set of objects are placed into clusters such that objects in the same cluster are more similar to each other than objects in other clusters according to some defined criteria.

Confusion matrix: A confusion matrix or table details the number of correct and incorrect (or misclassified) patterns generated by a classifier. It gives rise to measures such as accuracy, true positive rate, false positive rate, sensitivity and specificity.

Mutual Information: Mutual Information measures the arbitrary dependencies between random variables. It arises from information theory, where Shannon (1948) defined entropy or information entropy as a measure of the uncertainty associated with a discrete random variable.

Partitioning: Partitioning is a fundamental operation in data mining for dividing a set of objects into homogeneous clusters, or sets of data that are more similar to each other than other data. A partition can be hierarchical in nature.

Risk Prediction: Risk prediction models offer the means to aide selection from a general medical population, those patients that need referral to medical consultants and specialists.

Sensitivity: The sensitivity rate indicates the effectiveness of a classifier at identifying the true positive cases, i.e. those cases that are of most interest.

Specificity: The specificity rate indicates the effectiveness of a classifier at identifying the true negative cases, i.e. those cases that are the complement of those categories of most interest.

Supervised Classifier: A classifier that learns to associate patterns in data with a-priori (given) labels for each pattern. Examples include multi-layer perceptrons and decision trees.