

# Feature Selection and Predicting CardioVascular Risk

T.T.T.Nguyen and D.N. Davis, Computer Science, University of Hull.

## 1 Introduction

No gold standard exists for assessing the risk of individual patients in cardiovascular medicine. The medical data used for such purposes is, itself, inconsistent over a history of patients at any one clinical site, and not always immediately useable. In this paper the clustering of data using Self Organizing Maps (SOM) is described. This method is an unsupervised neural network developed by Teuvo Kohonen [4]. The SOM is primarily used for the organization and visualization of complex high dimensional data. It produces a mapping of inputs to an output space so that similar patterns of inputs are close together on the map and relatively important inputs take up more space on the map.

The data used in this paper derived from a Dundee clinical site; one of several clinical sites being used in the study. This experimental data size is quite small and extracted from the original data domain with 341 patients and 18 features. The data is transformed from its original mixed type (categorical, Boolean and continuous) to purely numerical (0 or 1 or continuous in the range 0 to 1).

This paper also describes the use of the Mutual Information (MI) and the ReliefF algorithms in feature selection. Feature selection allows an order to be placed on the data and enables the ranking the feature set (patient attributes in the clinical data) based on the relevance to a nominated outcome. The primary purpose of feature selection is to reduce the size of feature sets, and so enable more productive classification and clustering. The data set used for this paper is small (in patient number), so the use of feature selection is likely to further the success of the project in identifying data patterns that are linked to cardiovascular risk. Mutual Information (MI) is an important information measure for identifying important feature subsets. It generates a feature selection measure, enabling high-valued features to be selected for further analysis and the low-valued features to be simply discarded. An unprincipled approach to feature selection often reserves redundant features and deletes relevant features. In this paper, a novel change to the calculation of MI for both categorical and continuous data is used and is shown to be of use in generating useful models with medical data

Another popular feature selection is used in this study. ReliefF algorithms are general and successful attribute estimators. They are able to detect conditional dependencies between attributes and provide a unified view on attribute estimation in regression analysis and classification. In addition, their quality estimates have a natural interpretation. The ReliefF algorithm used is part of the WEKA software package [12].

By using a K-Means measure in the SOM clustering algorithm, it is possible to produce clusters that identify important (natural) patterns to be found in the data domain. A new feature-based risk model can be built with the outcome from the clustering exercise. These results are then reused in (MI and ReliefF) feature selection methods, and compared with a prior (heuristic clinical) risk model. This heuristic model is based on earlier research [11]. The comparisons of before and after clustering for risk are performed using several alternative neural network and classification techniques. This end result is shown to be of help in the choosing of suitable models for predicting risk in the cardiovascular data domain.

## 2 Method

### 2.1 Self Organizing Maps (SOM)

The SOM consists of a lattice of neurons for an output layer (the dimension space is usually two dimensional) connected to an input layer. The response of each node in output space to the input is given as the difference between the input and the weights as defined by a distance measure. In this paper the distance is measured as an Euclidian distance. The categorical values in the original data are transformed to numerical values with an ordering dependent on the significant information for these attributes. This is derived from the knowledge of clinical experts. For example, one feature "RESPIRATORY DIS HX" takes the categorical values:

Normal; Mild COAD; Mod COAD; and Sev COAD. The transformed values will be 0; 1; 2; 3 respectively. Another example, Cardiac Status, has the categorical values of AF; Asympt; CVA; TIA; Stroke; and so on. Clinical knowledge of cardiovascular disease [1] suggests that all the symptoms have significant impact to the carotid patients with the exception of the Asympt value. Hence the transformed values for this attribute will be 0 (Asympt); and 1 for the rest. The winning neuron in the SOM is selected from the lattice. That winner is a neuron which is closest (most similar) to the input pattern. The neighbours of the winner are selected from output space. The weights associated with the winner and its neighbours are adjusted so that they become closer to the current input pattern. This process is repeated many times in order to produce the form of the output lattice.

In this paper the clustering technique is used with a K-Means algorithm. The data set for the algorithm is taken from Dundee site, and it is transformed to numerical values according to its type (Boolean, continuous, categorical as in the above example), or discrete values. The K-Means clustering algorithm used is from a SOM Toolbox [10]. After using the K-Means clustering algorithm the new model of risk is built with the new outcome (as given by the cluster algorithm) for each pattern. This new model is investigated for feature selection using MI and ReliefF, and subsequently using alternative neural network techniques.

## 2.2 Feature selection: Mutual Information and ReliefF.

Feature Selection is a fundamental problem in data mining. The aim is to select relevant features and remove irrelevant and redundant features from an original feature set [7]. Mutual Information (MI) is based on Shannon's definition of entropy. MI measures the general dependence of random variables without making any assumptions about the nature of their underlying relationships. Consequently, MI can potentially offer some advantages over feature selection techniques that focus only on the linear relationships of variables. The distance between each feature and outcome class is calculated based on the theory of MI. The MI selected features can be compared to those features selected by ranking. Information Theory [9] generates a measure for the entropy (or disorganisation) to be found in data. This measure  $H(x)$  is an indicator of the uncertainty of the probability distribution  $p(x)$ , and the formula represents mutual information for two random variable  $x, y$  [5,6].

The formula for measuring the MI for class  $C_i$  and the attributes  $X_j$  ( $j=1$  to  $m$ ) is given as:

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

where  $p(C_i, x_{jk})$  is the probability of finding attribute  $x_j$  in class  $C_i$  in the  $k^{\text{th}}$  state,  $p(C_i)$  is the prior probability of class  $C_i$ , and  $p(x_{jk})$  is the prior probability of finding attribute  $x_j$  in the  $k^{\text{th}}$  state ( $n$  is number of classes, and  $m$  the number of patterns). It clear that the probability of finding attribute  $x_j$  in class  $C_i$  in  $k^{\text{th}}$  state ( $p(C_i, x_{jk})$ ) is the probability of finding the number of patterns in class  $C_i$  with the attribute  $x_j$  when we consider in the  $k^{\text{th}}$  state. Based on Bayes's theorem, we have:

$$MI(C, x_j) = \sum_{i=1}^n \sum_{k=1}^s p(C_i | x_{jk}) p(x_{jk}) \log \frac{(p(C_i | x_{jk}) p(x_{jk}))}{p(C_i) p(x_{jk})} \quad (2)$$

where  $P(C_i | x_{jk})$  is the probability that vector  $x$  fell in class  $C_i$  (Class  $i^{\text{th}}$  of  $n$  different classes),  $p(x_{jk} | C_i)$  is the probability density functions, and  $p(x)$  is the unconditional probability density function, which is calculated as:

$$p(x_{jk}) = \sum_{i=1}^n p(x_{jk} | C_i) P(C_i) \quad (3)$$

It is possible to rewrite the formula for mutual information as follows:

$$MI(C, x_j) = \sum_{i=1}^n \sum_{k=1}^s p_{ijk} \log \left( \frac{p_{ijk}}{q_i \cdot r_{jk}} \right) \quad (4) \text{ where}$$

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

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

$r_{jk} = \frac{sum_{jk}}{sum}$  where  $sum_{jk}$  is the number of patterns in attribute  $x_j$ , considering state  $k$ .

The above formula (4) is also used for the continuous valued attributes. For such attributes the number of possible state ( $s$ ) is defined as the number of discrete bins placed over the continuous range, and is calculated as  $\log_2 N + 1$ , where  $N$  is the number of patients in data set.

Another method is to define the relevance and ranking of the significant aspects of a feature set using the ReliefF algorithm [2,8]. ReliefF evaluates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different classes. More over, ReliefF can operate on both discrete and continuous class data. The quality of attributes is given according to how well their values distinguish between instances that are near to each other.

### 3 Results

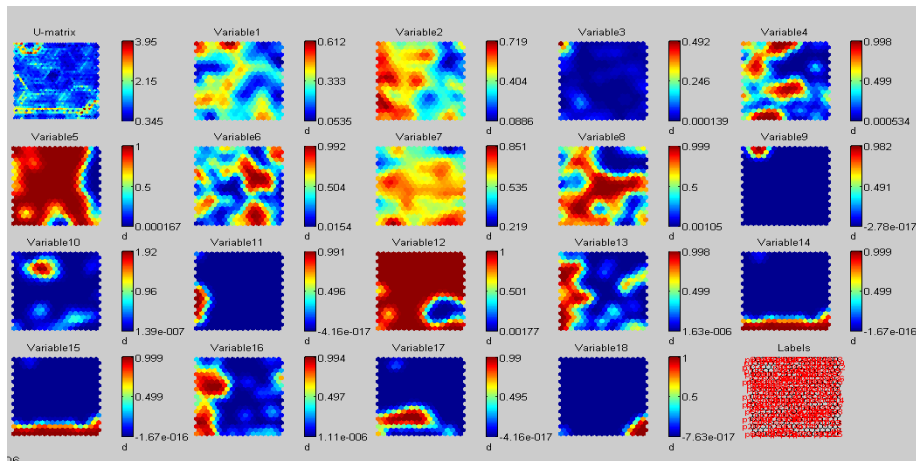


Figure 1: Representation of SOM in Umatrix and its features with quantization error 1.942; and topographic error: 0.003.

In figure 1, the data is coded to the code book vector in order to be presented to the map of a U matrix. Each feature is mapped to each variable in the figure. Each patient is a label in the map. Figure 1 shows the representation of data onto the map (U matrix) and each attribute map. However, in this study data also is input to a clustering algorithm (the K-Means\_clustering, algorithm in the SOM toolbox [10]). It generates K means for the given data set with different values of K, and is run multiple times for each K. The best of these runs is selected, based on a suitable measure (for example the sum of squared errors).

The result of clustering is used to produce the new outcome (risk) for the data set. The model with the new outcome of SOM clustering is investigated using MI and ReliefF in order to show the differences of significances before and after using the clustering method.

In figure 2, the measurement between each feature in the data set for alternative outcomes (Risk outcome, and SOM clustering outcome) is very similar except the HD attribute. This has a significant change of ordering after clustering; a change from 0.67 (MI); 0.62 (ReliefF) to 0.016 (MI), and 0.12 (ReliefF) respectively. This change may be because of the reduced significance of this attribute to the new outcome. From figure 2, we can see the results of using the novel MI formula with both discrete and continuous data. It defines the significance of each attribute in predicting the outcome, and gives results quite close to the popular ReliefF algorithm.

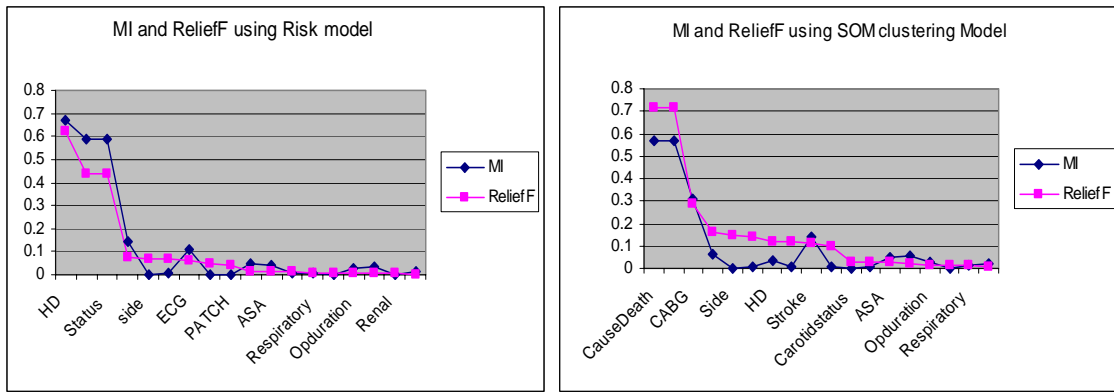


Figure 2: Results of using MI and ReliefF of Risk model and SOM clustering model.

A different evaluation of these alternative risk models can be made through the use of supervised neural network and other classification techniques. In one clinical risk model, outcome is predicted using an heuristic rule. A full investigation of this model using supervised neural network techniques is given in [11]. The new model derived from using the SOM K-Means clustering algorithm can be evaluated using the same supervised neural network techniques, and the decision tree classifier (J48). Tables 1 and 2 below show the results for the two risk models. Overall the Mean Square Error (MSE) for both models are good. MLP and J48 provide the best results with 0.001 and 0.00 respectively. For the SOM clustering method, it seems that the clusters of C1, C2, and C3 are quite close to the definition of Low, Medium, and High risk in the heuristic clinical model. It might be that the clustering of SOM produces a close match to the clinical outcome for this data set.

		Confusion Matrix			MSE
CM3b		Low	Medium	High	
					0.07
SVM	Low	220	0	0	
	Medium	1	72	0	
	High	0	0	48	
MLP	Low	220	0	0	0.002
	Medium	2	71	0	
	High	0	0	48	
RBF	Low	215	5	0	0.16
	Medium	63	10	0	
	High	44	4	0	

		Confusion Matrix			MSE
SOM		C1	C2	C3	
					0.07
SVM	C1	275	0	0	
	C2	0	18	1	
	C3	0	0	47	
MLP	C1	275	0	0	0
	C2	0	18	1	
	C3	0	0	47	
RBF	C1	275	0	0	0.14
	C2	0	19	0	
	C3	46	0	1	

Table 1 and Table 2: Alternative neural network techniques applied in risk model and SOM clustering model.

#### 4 Discussion and Further Work

By using SOM to visualize and represent the clinical data set, a measure of the distance for each data pattern (i.e. each patient) to the others is calculated. This is based on the Euclidian distance between them over the input feature set. To be suitable for use with the SOM tool box, data is transformed to numerical values. This requires the normalization of discrete, Boolean, and continuous data into the range of [0,1]. The categorical data, where possible, is transformed into ordinal data based on the knowledge of medical experts. However, this might increase the number of input feature attributes in the data set. This will impact on the consistency of the

database or the running time and effectiveness of the clustering algorithm. By looking in depth at the clusters produced using the SOM algorithm, we hope to find out the misclassification of outputs in the given patterns. By using colour coding for each unit, the clusters for the patterns are established. This initial coding needs to be subjected to further work.

The small change to the MI formula is found to be suitable for the medical data and continuous data values. From the experiments summarised in figure 2 and by comparison to the result through using ReliefF, it seems that this MI calculation can define the significance of each attribute in determining the clinical outcome. The data set used here has a quite small attribute set so its effectiveness to clarify feature selection is not fully proven. However this experiment has demonstrated its potential effectiveness for further large data sets. Further research with other clinical data sets will determine its usefulness for such data.

Future work will also investigate the confusion matrix for each classifier to see the type of errors being made. If the use of the SOM clustering model compares to the heuristic rule model (with High, Medium, Low outcomes), it may demonstrate that the definition of rules in [11] seem to be appropriate to the data domain. These results can then be applied in determining what of the original data should be used to generate a better set of classifiers of use in predicting CardioVascular risk. The methodology developed in this project can then be applied in further clinical domains.

## References

- [1] Family Practice Notebook, Cardiovascular chapter, Available at <http://www.fpnotebook.com/NEUCh7.htm>.
- [2] Kira, K. and Rendell, L.A. (1992a): "The feature selection problem: traditional methods and new algorithms". Proceedings of AAAI'92, pp. 129-134.
- [3] Kira, K. and Rendell, L.A. (1992b): "A practical approach to feature selection". Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, pp. 249-256
- [4] Kohonen, T.(2001): Self- Organizing Maps. Springer-Verlag Berlin Heidelberg:Third Edition.
- [5] Kullback S. & Leibler. R. A. (1951) On Information and Sufficiency. Annals of Math. Stats. 22 pp79-86.
- [6] Kullback, S. (1959) Information Theory and Statistics. Wiley.
- [7] Liu H., Motoda H. Feature Selection for Knowledge Discovery and Data Mining. Norwell, MA USA: Kluwer Academic; 1998.
- [8] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics. Edited by Lucien M. Le Cam and Jerzy Neyman. University of California Press.
- [9] Shannon, C. E. A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.
- [10] SOM Tool Box in Matlab, free down load at <http://www.cis.hut.fi/projects/somtoolbox/download/>.
- [11] Thuy,N.T.T, Davis, N.D. (2006), Predicting Cardiovascular using POSSUM, PPOSSUM and Neural network techniques. International Conference on Enterprise Information System (ICEIS2006), Cyprus.
- [12] WEKA software (University of Waikato, New Zealand, version 3.4.5). Download free at : <http://www.cs.waikato.ac.nz/~ml/weka/index.html>