A NOVEL NEURAL NETWORK TECHNIQUE FOR EXPLORATORY DATA ANALYSIS*

David Lowe and Michael Tipping

Neural Computing Research Group
Aston University, Birmingham B4 7ET, United Kingdom.
Email: `lowed@aston.ac.uk` `tippinme@aston.ac.uk`
Tel: +44 (121) 359 3621 x4671
Fax: +44 (121) 333 6215

Abstract: *This paper surveys the context of feature extraction by neural network approaches, and compares and contrasts their behaviour as prospective data visualisation tools in a real world problem. We also introduce and discuss a hybrid approach which allows us to control the degree of discriminatory and topographic information in the extracted feature space.*

FEATURE EXTRACTION AND NEURAL NETWORKS

Amongst the pattern processing abilities of neural networks, perhaps the most generic ability is that of feature extraction. Feature extraction is at the centre of developments in neural networks for two main reasons: one is the significant advantage to be gained by dimensionality reduction which often accompanies feature extraction, and the other is the necessary ability to construct *nonlinear* representations of data which may then be exploited for information analysis. Feature extraction is also an important area as it provides a common ground for exploring the relationships between neural network approaches and other more traditional techniques.

Typically, feature extraction techniques can be characterised by whether they are unsupervised or supervised (a hybrid is discussed shortly). The standard classical unsupervised feature extraction technique is principal components analysis, which is optimal from the perspective of preserving maximum variance in the transformed data. The links between associative networks and linear principal components analysis have been explored (Oja, 1989) and extended (Saund, 1989) to consider a form of nonlinear principal components related to the method of principal curves (Hastie and Stuetzle, 1989). More recently (Webb, 1995; Lowe, 1993; Mao and Jain, 1995) a different class of feature extraction techniques has been explored which relates networks to a nonlinear projection method known as a *Sammon mapping* (Sammon Jr, 1969), and also to the statistical technique of *multidimensional scaling* (MDS) (Davison, 1983).

These are techniques based on unsupervised training: situations in which we have no extra knowledge concerning possible class membership. If we have extra class information, then it is possible to construct superior feature extraction spaces which are more appropriate for subsequent discrimination of patterns. In this context it has been shown (Gallinari et al., 1991) that the hidden layer space of a *linear* feed forward network trained to perform a classification task is optimised in such a way as to create a linear discriminant transformation which maximises the clustering criterion $|\mathbf{S}_B|/|\mathbf{S}_W|$, the ratio of the determinants of the between-class and within-class scatter matrices of the transformed data. This linear discriminant space relates directly to the classical Fisher's Discriminant (Fukunaga, 1990).

*This work was partly supported by EPSRC grant no. J75425, "Novel Developments in Learning Theory for Neural Networks".

More generally, a *nonlinear* neural network, with linear output units trained to achieve classification, exploits a very specific optimised internal representation (Webb and Lowe, 1990). This feature space is constructed to maximise the *network discriminant function* $Tr[\mathbf{S}_B\mathbf{S}_T^+]$, which is related to standard techniques for dimensionality reduction and feature extraction in statistical pattern recognition (Fukunaga, 1990).

These are examples which represent extremes of behaviour from unsupervised to fully supervised, and from linear to nonlinear extensions.

Another, and less well explored, characteristic of feature extraction techniques is their 'topographic' nature – i.e. whether the transformed feature space preserves neighbourhood relations between patterns in the original space. The Kohonen network is the archetypal neural network which attempts to retain topographic information, but has several inherent disadvantages. An alternative is the Sammon mapping, which attempts to preserve pairwise inter-pattern distances in the transformed space, and hence patterns which were close in the original space are likely to be close in the projected space. This is not necessarily the case in other feature extraction approaches, such as generalised discriminant analysis. The work in (Webb, 1995; Lowe, 1993; Mao and Jain, 1995) relates Sammon mappings and MDS to feed-forward neural networks trained to optimise similar structure preservation criteria. Hence there is a class of network topologies, which may be exploited to construct topographic feature extraction spaces, that is still relatively unexplored and unknown to most researchers.

Although generally the relationships between networks and traditional methods does not produce any novel development, other than increased insight and understanding of the behaviour of network methods, the use of networks to perform Sammon mappings and MDS is an exception. The two main novel aspects are the ability of a network to construct a *transformation*, and the advantage that the computational complexity is not dictated by the amount of data, but rather by its complexity. The transformation characteristic is important as it permits the subsequent projection of novel data into the discovered feature space, whereas the original technique only allows a reconfiguration of the fixed training set. In addition to these advantages, the neural network approach to 'dissimilarity matching' permits a hybrid model which allows a smooth transition between unsupervised and supervised nonlinear topographic feature extraction, which we now discuss.

SAMMON MAPPINGS, NEURAL NETWORKS AND A HYBRID MODEL

The Sammon mapping is a topographic transformation that seeks to retain the neighbourhood relationships between data points. Given a prior choice of dimension, the Sammon map is generated by minimisation of the *Sammon Stress* error term:

$$SS = \frac{1}{\sum_{ij} d_{ij}^*} \sum_{ij} \frac{(d_{ij} - d_{ij}^*)^2}{d_{ij}^*} \qquad (1)$$

Where d_{ij}^* is the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the data space, and d_{ij} is the distance between their corresponding images $\mathbf{y}_i$ and $\mathbf{y}_j$ in the feature space. Since $d_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||$, the mapping can be determined by adjusting the points $\mathbf{y}$ iteratively (by a gradient descent method, for example).

In order to permit generalisation to new data, the points $\mathbf{y}_i$ may be defined as a parameterised function of the original data points, $\mathbf{x}_i$. That is, $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i; \mathbf{W})$, and the parameters $\mathbf{W}$ can be estimated in order to minimise the Sammon Stress.

The Sammon mapping as described is purely unsupervised in that it takes no account of class information. However, there may well exist additional subjective knowledge, in terms of a measure of *dissimilarity* between classes, that can be exploited in the transformation. In this paper, this knowledge is incorporated in a parameterised transformation effected by a radial basis function network. We introduce a modification to the distance matrix that allows us to smoothly interpolate between unsupervised and supervised feature extraction by replacing the distances d^*_{ij} in equation (1) by the dissimilarities:

$$\delta_{ij} = (1 - \alpha).d^*_{ij} + \alpha.c_{ij}, \tag{2}$$

Where c_{ij} is the assigned dissimilarity between the classes to which patterns $\mathbf{x}_i$ and $\mathbf{x}_j$ belong. The parameter α can thus be considered as an interpolating parameter between unsupervised and supervised transformation, and is a more general method of including class information than that utilised in (Cox and Ferry, 1993). With $\alpha = 0$, the network is effecting a parameterised Sammon mapping. With $\alpha = 1$, the transformation is no longer dependent on the distribution of the data explicitly, but is determined by the assigned class dissimilarities. As such, the network effectively performs a discrimination task. It is only the *relative* positioning of the transformed data points that is relevant. Although there is no specific target value for the network to aim for during training, there is more information available than in a completely unsupervised process, and so we refer to this optimisation as 'relative supervision'.

EXAMPLE: THE 1992 RAE DATABASE

In 1992, the Universities Funding Council undertook a Research Assessment Exercise (RAE) of 72 separate subject areas in all higher education institutions in the United Kingdom. Institutions supplied numerous quantitative indicators of their respective research activity, such as the number of active researchers, postgraduate students, the values of grants awarded and various numbers of publications. These indicators were used as a guide to the assignment of a *research rating*, on a scale of 1 to 5, to each subject area at each institution.

This data has been made available and it is naturally of interest to analyse it, particularly with the aim of classifying the data with respect to the research rating. Some statistical analyses have already been published, for example (Taylor, 1994).

The data presents a challenge due to both its high dimensionality, which is of the order of 150 in its raw form, and due to the "noise" present in the values of the research rating class labels due to their subjective assignment.

Although there are over 4000 records in the database for all subjects, the distribution of the explanatory variables and the correllation between them and the research rating varies widely between subject areas. We chose three related and well correllated subjects for our analysis – Physics, Chemistry and Biological Sciences. In addition, we retained data for Applied Maths in order to assess generalisation.

We undertook some pre-processing of the data in order to generate a data set of 217 examples each with 80 explanatory variables, and the following techniques were applied for the analysis of the data: Principal Components Analysis, Generalised Linear Discriminant Analysis, Sammon Mapping, Kohonen Mapping, Neural Network Classifier and the RBF topographic transformation ('NeuroScale') for values of $\alpha = 0.5$ and 1.0. For the purposes of visualisation, all projections were down to two dimensions, and the feature spaces thus extracted are illustrated in figures 1 to 8 below.

Figure 1: Principal Components Analysis

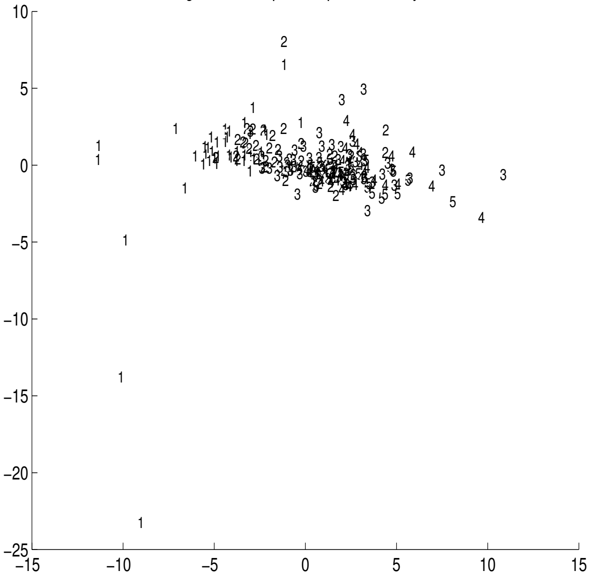


Figure 2: Sammon Mapping

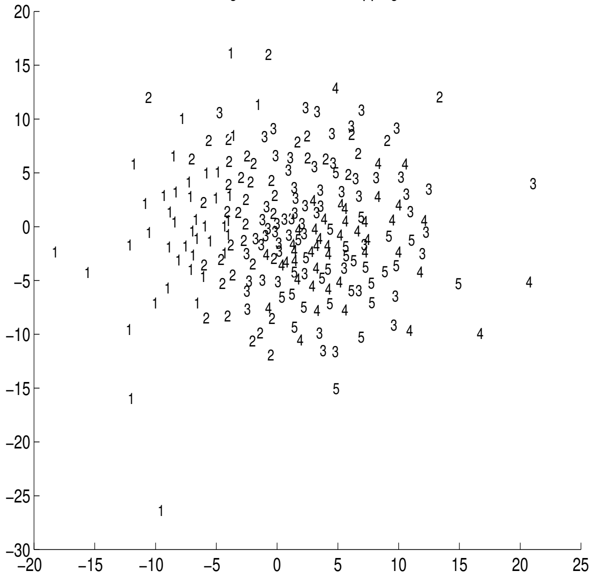


Figure 3: Linear Discriminant Analysis

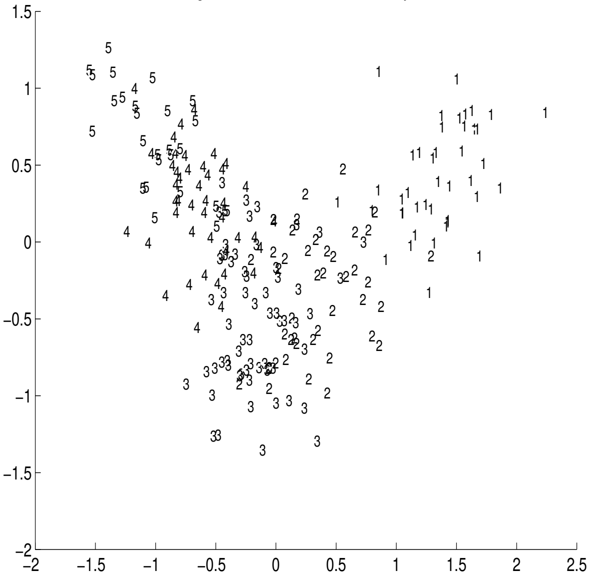


Figure 4: Test Projection – Linear Discriminant Analysis

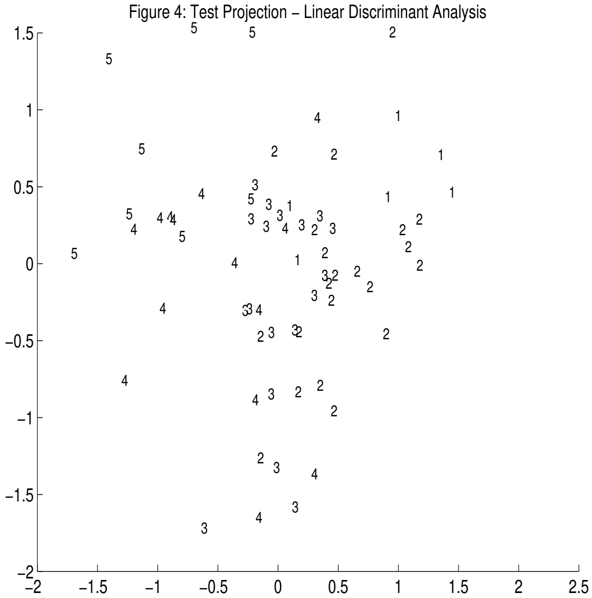


Figure 5: NeuroScale, alpha=0.5

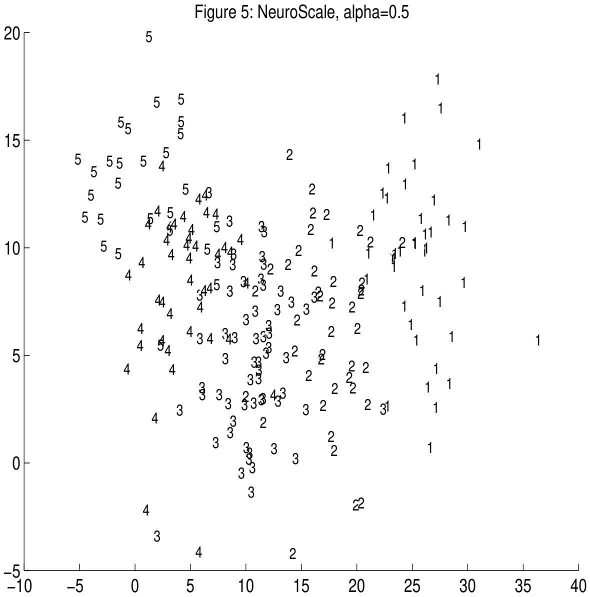


Figure 6: Test projection – NeuroScale, alpha=0.5

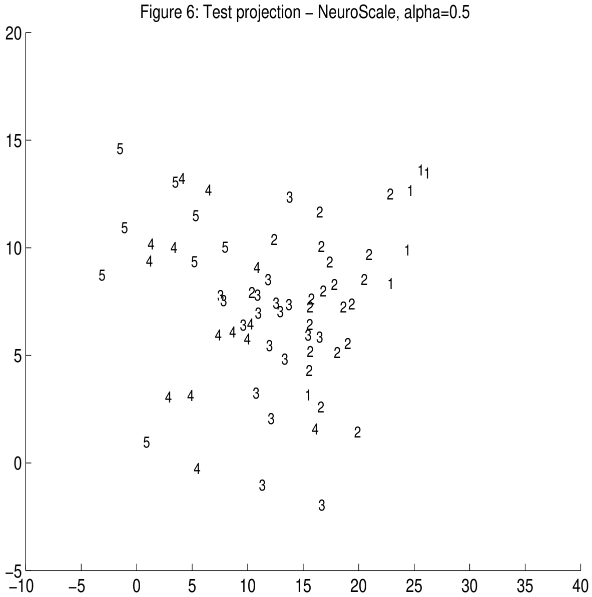


Figure 7: NeuroScale, alpha=1

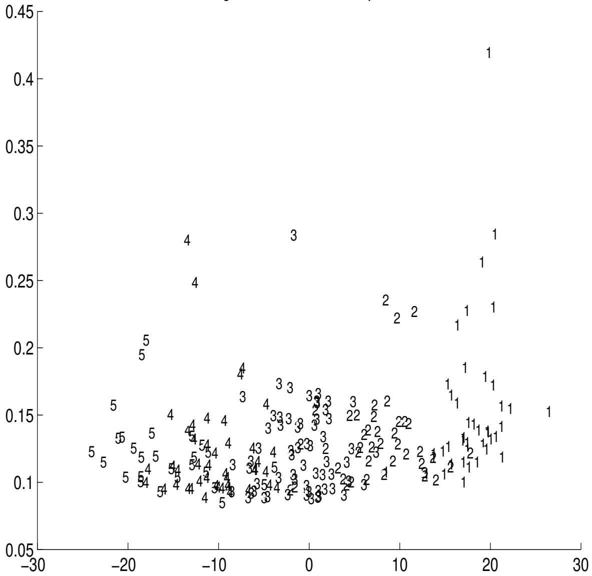
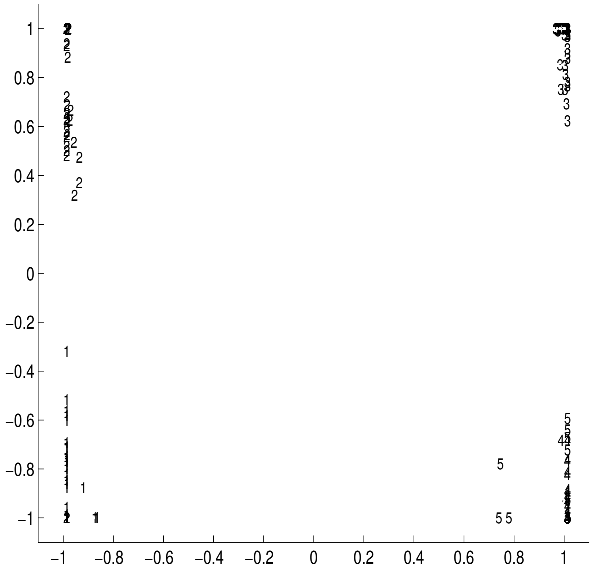


Figure 8: Hidden Unit Space of Neural Network Classifier



DISCUSSION

The figures illustrate a range of features extracted under assumptions of unsupervised to supervised, and linear to nonlinear. The RBF topographic transformation is a nonlinear process which incorporates varying degrees of extra subjective knowledge, which in this example is specific class information. Although a Kohonen network was also applied to this database, the characteristics of the data render the Kohonen representation ineffective and hence the results were not presented. From figure 1 it is clear that the features extracted by the PCA are not likely to be of any exploratory use, since no significant structure could be extracted. However by exploiting explicit class information, the LDA space (figure 3) manages to separate the patterns approximately into class clusters. The same is true for the hidden unit space of the multilayer perceptron (figure 8), except that the data separation is more severe (and two of the classes, 4 and 5, are ‘confused’ by the network as there are only 2 hidden nodes). The use of the LDA to provide a transformation rather than just a mapping is exhibited by figure 4 which demonstrates the effect of projecting a test set of data (‘Applied Mathematics’) into the LDA space.

The Sammon mapping (Figure 2) depicts the *nonlinear* but *unsupervised* feature space. Despite not exploiting any class information, this representation produces a set of features which allow patterns in similar classes to remain in close proximity. Neighbourhood relations in the data space are approximately preserved in the feature space. This is the same feature space that the ‘NeuroScale’ network would produce with $\alpha = 0$. Figures 5 and 7 show the feature spaces produced by the NeuroScale network which incorporate progressively more subjective ‘class’ knowledge into the transformation. Similarly, it is observed that the neighbourhood relations in the data associated with their class labels are approximately preserved in the feature space transformation.

The main advantage of the NeuroScale approach to a Sammon mapping is the ability to generalise. Figure 6 shows the projection of the test set of data (‘Applied Mathematics’) into the derived NeuroScale feature space for the case of $\alpha = 0.5$, exhibiting a ‘sensible’ projection which could be used for subsequent decision making or inference.

In these figures, we have illustrated the advantage of nonlinear feature extraction techniques which exploit aditional ‘subjective’ information. In our example, extra information takes the form of class labels, for which we may also apply other methods such as LDA. However, in general we may not know explicit class labels, but we may have a more general knowledge of ‘dissimilarity’ than is normally provided in the raw data iself. In this case, the proposed NeuroScale approach is one of the few accessible techniques available.

References

Cox, T. F. and Ferry, G. (1993). Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153.

Davison, M. L. (1983). *Multidimensional Scaling*. John Wiley & Sons, New York.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, second edition.

Gallinari, P., Thiria, S., Badran, F., and Fogelman-Soulié, F. (1991). On the relations between discriminant analysis and multilayer perceptrons. *Neural Networks*, 4:349–360.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84:502–516.

Lowe, D. (1993). Novel 'topographic' nonlinear feature extraction using radial basis functions for concentration coding in the 'artificial nose'. In *3rd IEE International Conference on Artificial Neural Networks*. IEE, London.

Mao, J. and Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68.

Sammon Jr, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Saund, E. (1989). Dimensionality-reduction using connectionist networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(3):304–315.

Taylor, J. (1994). Measuring research performance in business and management studies in the United Kingdom: The 1992 research assessment exercise. *British Journal of Management*, 5:275–288.

Webb, A. R. (1995). Multidimensional scaling by iterative majorisation using radial basis functions. *Pattern Recognition*, 28(5):753–759.

Webb, A. R. and Lowe, D. (1990). The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3(4):367–375.