Pattern Analysis & Applications (2001)4:227–234
© 2001 Springer-Verlag London Limited

MLP Based Linear Feature Extraction for Nonlinearly Separable Data

A. Raudys and J. A. Long

School of Computing Information Systems and Mathematics, South Bank University, London, UK

Abstract: A novel approach to linear feature extraction is presented. Most supervised feature extraction algorithms use mean square error or other measures based on the difference between expected and actual output values as a performance criterion. The novel approach presented here uses data visualisation together with an empirical classification error (percentage of cases classified incorrectly) as performance criterion. To find the optimal data transformation weights, the Multilayer Perceptron cost function with a special regularisation term is applied. The technique proposed is verified and compared with five competing mapping techniques with respect to visualisation and different classification error criteria. For comparison, two artificial and 12 real world data sets are used.

Keywords: Feature extraction; Linear; Mapping; Multilayer perceptron; Principal components; Sammon; Transformation; Visualisation

1. INTRODUCTION

A number of research papers are devoted to the development and comparison of new feature extraction algorithms [1,2]. John *et al.* [3] compared PC (Principal Components) and linear discriminant analyses on several small handwritten character data sets. Backer *et al.* [4] studied four unsupervised nonlinear feature extraction methods. Other research work on this topic can be found elsewhere [3,5–9].

Two important aspects of feature extraction can be formulated. One is mapping accuracy measured in terms of classification or prediction error. The second is *data visualisation*. Good data visualisation is necessary in data mining and knowledge discovery applications where there is a need to present a solution in a very simple, easily understandable way. It is also important for optimal initialisation of neural networks [10]. Researchers have offered some modifications on existing feature extraction algorithms and created new ones. A dimensionality reduction technique that seeks directions emphasising multimodality was presented by Intrator [11]. In Raudys [10] a feature extraction technique for active neural network initialisation was proposed. In Pao and Shein [12], the authors proposed a novel neural network feature extraction algorithm based on variance conservation. The authors also noted the importance of 2D feature extraction for visualisation purposes. A number of feature extraction algorithms are based on neural networks. All of them incorporate nonlinear functions to transform and reduce the dimensionality of the data.

In spite of the abundance of mapping algorithms, there is a shortage of simple linear methods which reveal a data structure with respect to multiple pattern classes and, at the same time, allowing data visualisation. The novel feature extraction method presented in this paper uses linear transformation. It utilises classification error, the number of cases classified incorrectly, as the feature mapping criterion. To minimise the classification error, a modified conventional MLP cost function is suggested, paying special attention to the magnitudes of the output layer weights.

Received: 6 March 2000
Received in revised form: 12 September 2000
Accepted: 11 October 2000

2. TAXONOMY OF FEATURE MAPPING ALGORITHMS

Feature Extraction (FE), also known as mapping, variance conservation or feature compression, can be divided into several groups (see Fig. 1). Feature extraction is determined by two major factors:

1. The type of *transformation* determining how the new features are made from the initial (the original) ones.
2. The *criterion* for extracting features (a cost function is

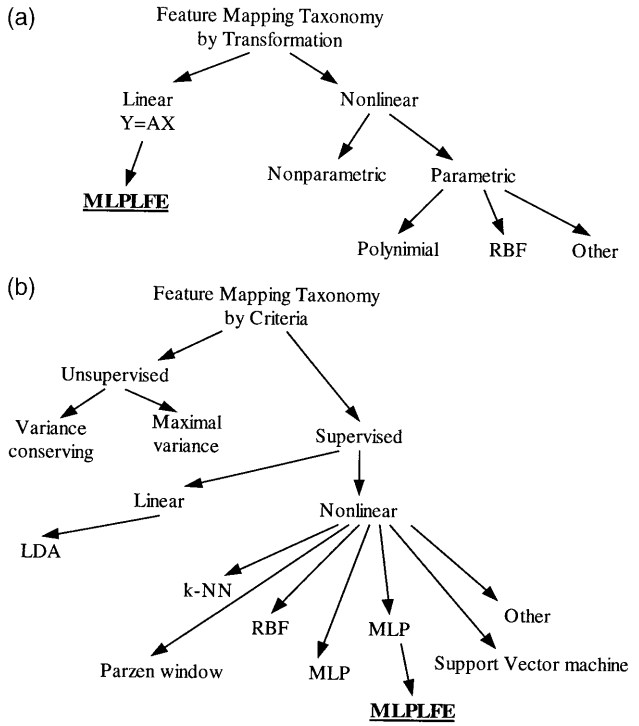

Fig. 1. Taxonomy of feature mapping algorithms. (a) According to type of the transfromation function; (b) according to a type criterion (of the cost function), and according to the method used to evaluate mapping performance used to find parameters of the transformation function.

used to find optimal coefficients that determine the transformation).

Sometimes researchers confuse the type of transformation and FE criterion. For example, Sammon mapping [13] is sometimes called nonlinear. However, it utilises linear transformation and a nonlinear criterion (vector inter-distance preservation).

The transformation. Often researchers use linear transformations $\mathbf{Y}=\mathbf{V}\boldsymbol{X}$, where $\mathbf{V}$ is a transformation matrix, in performing FE. Depending on the transformation matrix, linear transformation can be orthogonal or not orthogonal. Examples of linear transformation feature extraction techniques are Principal Component analysis [14], Sammon [13] and Foley-Sammon [14] methods. For nonlinear transformation techniques, polynomial functions, radial basis functions and outputs of the MLP can be employed. An example of nonlinear transformation feature extraction is illustrated by the auto associative neural network [15].

The criterion. The criterion helps to evaluate the mapping accuracy, i.e. how good the extracted features are. This criterion is minimised during feature extraction training. For example, in the Sammon method, the criterion is vector inter-distance. During training, the algorithm is trying to preserve vector inter-distances in the new mapped feature space. In pattern classification problems, the criterions can be subdivided into:

1. *Unsupervised* (a criterion does not utilise class indexes).
2. *Supervised* (the class indexes of the data are taken into account).

The aim of supervised FE is to extract features where classes (in the extracted features) can be separated in the best possible way. Supervised FE can be subdivided into two groups, linear and nonlinear. Another way to subdivide the supervised mapping algorithms is based on their relation to the classification error. In the literature, only one single method which minimises classification error directly [16] can be found. The assumption that data is distributed in a multivariate Gaussian way is made, and the probability of misclassification is minimised in the new reduced feature space. Unfortunately, this method becomes useless if the data is complex, multimodal or asymmetric (non-Gaussian). Other known methods minimise criteria that are only approximately related to the classification error (e.g. mean square error). None of the other FE methods utilise classification error directly, that is, the difference between correctly and incorrectly classified cases. Supervised criteria can be applied to linear (classification error is evaluated by a linear algorithm – Foley-Sammon) and to nonlinear (classification error is evaluated by a nonlinear algorithm – complex MLP, *k*-nearest neighbour (*k*-NN) algorithm, Parzen window algorithm, etc.) mapping methods. Below several popular FE methods used in benchmarking are presented.

Principal Component Analysis [14], also known as the Karhunen–Loeve transformation, is the most popular FE technique. It is an unsupervised linear transformation algorithm. The central idea of Principal Components is to find features with maximal rate of decrease of variance. Extracted features are orthogonal between each other.

Foley–Sammon. The widely known Foley–Sammon feature mapping algorithm (sometimes called linear discriminant analysis feature extraction) [17], belongs to the group of supervised methods. The Fisher linear classifier is used by Foley-Sammon to create a linear decision boundary [14] which helps to extract the first new feature. Distance to a discrimination hyperplane serves as a new feature. This feature extraction procedure is repeated on the rest of the orthogonal features until there are no features left. The first extracted feature is a direction where pattern classes are best separated linearly. The second feature is the direction where classes are separated slightly worse, and so on. The disadvantage of this algorithm is its linearity and limitation for two pattern classes only.

Auto Associative Neuronal Network (AANN) is also known as nonlinear, neural network based Principal Component analysis [15,18]. It is an unsupervised feature extraction algorithm. It has two groups of layers, one for encoding the data, and another one for decoding (see Fig. 2). Each group of layers can contain from one to several layers. The number of encoding group outputs is equal to the number of decoding group inputs and equal to the number of extracted features. The number of encoding inputs is equal to the dimensionality of the input vector $\mathbf{X}$ and equal to the decoding group outputs. The encoding layer compresses the original features to fewer features, and the other group

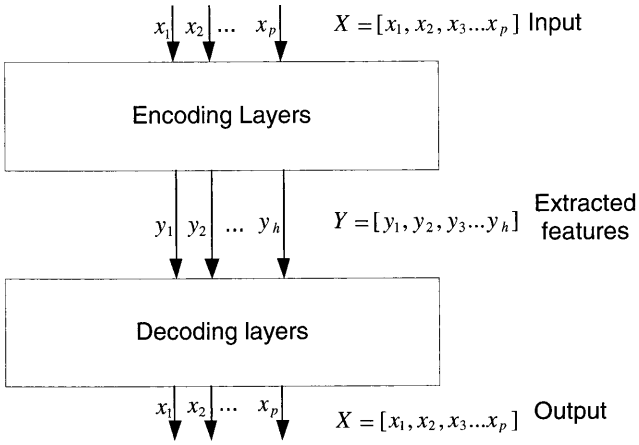

Fig. 2. Auto associative neural network model.

decodes the features back to the original (input) space. There are two major architectures of the AANN. The first of them uses one layer for encoding and one layer for decoding (3 layer MLP). The second one uses two layers for encoding and two layers for decoding (5 layer MLP). In 5 layer AANN the *p-h-r-h-p* (the letters indicate the number of neurons in each layer) architecture is utilised, where *p* is a number of inputs, *r* is a number of extracted features and *h* is a number of neurons in encoding and decoding hidden layers. Instead of the nonlinear feature extraction architecture *p-h-r-h-p*, a linear 3 layer architecture *p-r-p* can be used.

After the network is trained (the outputs of the network are approximately equal to the inputs), the decoding group can be removed and only the encoding layers must be used to extract features.

Multilayer Perceptron feature extraction. In this technique, the conventional MLP (MultiLayer Perceptron) based classifier is trained, and the outputs are taken as the new extracted features. It has been used in some comparative studies [7]. The limitation of this algorithm is that the number of classes is equal to the number of extracted features. There is no freedom in selecting the desired number of extracted features. In the case of two pattern classes, only two features can be extracted. In the case of three classes, three features can be extracted, and so on. The geometry of the data is lost as well. In the new feature space of the two pattern class problem, the pattern vectors are concentrated in opposite corners of a ((0;0) and (1;1)) square (see Fig. 3).

3. A NOVEL MULTILAYER PERCEPTRON BASED, LINEAR FEATURE EXTRACTION TECHNIQUE

The previous literature review has shown that, up to now, there has been no linear feature mapping method able to work with multiple pattern classes using classification error as the criterion for FE. The reason is simple. While calculating the classification error, a threshold function should be

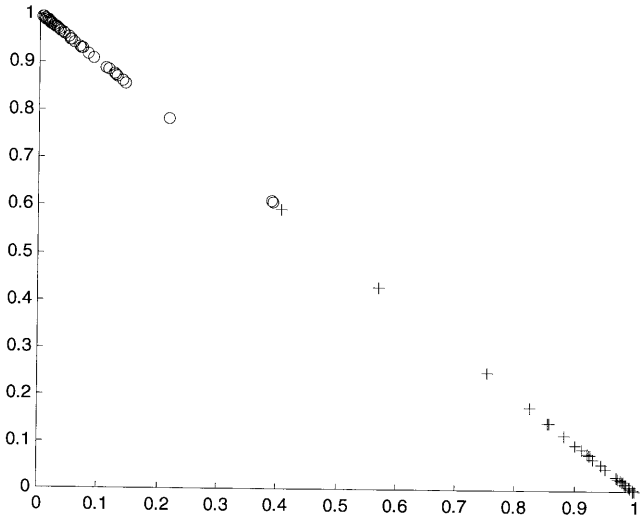

Fig. 3. Two features extracted using the MLP technique from ionosphere data. Horizontal axis represent MLP first output, and vertical second output.

used. This function is non-differentiable and cannot be minimised using conventional gradient descent optimisation methods. In principle, to overcome this numerical difficulty, genetic optimisation methods can be utilised. These methods, however, have a shortcoming in that their convergence is rather slow.

To solve the criterion problem, an observation in Raudys [19] is used. This is the fact that with an increase in the magnitude of the weights, the cost function of a nonlinear single layer perceptron begins to minimise the classification error. Indeed, let a cost function of a Single Layer Perceptron (SLP) be:

$$\text{cost}_t = \frac{1}{N} \sum_{k=1}^{L} \sum_{f=1}^{N_i} (t_j^{(i)} - f(W_k^T X_j^{(i)} + w_{k0}))^2 \qquad (1)$$

where $tj^{(i)}$ is a desired output (a target) for $\mathbf{X}_j^{(i)}$, *j*th training set observation from class ω_i, L is a number of pattern classes, $\mathbf{W}_k$, $\mathbf{w}_{k0}$ are weights, and $f(c)$ is a nonlinear activation function, e.g. the sigmoid function: $f(c)=1/(1+\exp(-c))$. When the weights are small, the activation function acts almost as a linear function. When the weights are large, the weighted sums $c_{ij} = W_k^T \mathbf{X}_j^{(i)} + w_{k0}$ becomes large and, for all training vectors $\mathbf{X}_j^{(i)}$, outputs of SLP $f(W_k^T \mathbf{X}_j^{(i)} + w_{k0})$ approach either 1 or 0. In such a case, separate terms $(t_j^{(i)} - f(W_k^T \mathbf{X}_j^{(i)} + v_{k0}))^2$ in Eq. (1) are very close either to 0 or 1. It means the cost function begins to minimise the frequency of misclassifications. Consequently, in principle, the cost function (1) can be utilised to minimise the empirical classification error.

To ensure convergence, the initial weights should be small. The cost function will then be smooth and differentiable. During training, the weights should increase. In back propagation training, typically the starting weight values are chosen as small randomly determined values. These values increase during training up to optimal values. At the end

of the training process, the magnitudes of the weighs are determined by the classification error as expressed by the degree of overlap of the training sets of the opposite pattern classes. Therefore, in high empirical error cases, in order to ensure the weights grow and thereby minimising the empirical classification error, it is recommended to add a special anti-regularisation term $+ \lambda \sum_{k=1}^{L} (\mathbf{W}_k^T \mathbf{W}_k - c^2)^2$ to the cost function, where λ and c^2 are parameters of the penalty (regularisation) terms. Parameter c^2 controls the magnitudes of the weights vectors $\mathbf{W}_k$ ($k = 1, 2, \ldots, L$).

To apply this useful peculiarity of the nonlinear SLP training, the outputs $\mathbf{Y}=\mathbf{V}\mathbf{X}=(y_1,y_2)^T$ of the linear transformation have to be used as the inputs to the nonlinear SLP (here $(y_1,y_2, \ldots$ are components of the vector $\mathbf{Y}$). The weights of SLP can be found by minimising the cost function (1) by the conventional back propagation algorithm. Unfortunately, SLP forms only the linear decision boundary. The classification error is evaluated only by the linear algorithm.

To overcome this difficulty, *nonlinearly transformed outputs* $f(y_j)$ are used as inputs to the nonlinear SLP in an upper part of the information processing schema (see Fig. 4). Analysis of this calculation schema shows that actually, the MLP with one hidden layer is used. To force the MLP classifier to minimise the number of training set errors, the cost (1) with the additional anti-regularisation term $+ \lambda \sum_{k=1}^{L} (\mathbf{W}_k^T \mathbf{W}_k - c^2)^2$ is utilised, where $\mathbf{W}_k$ is the weight vector of the kth output layer neuron. In practice, training is stopped when the minimal number of classification errors is obtained. The parameters λ and c^2 are evaluated in a trial. Thus, in reality, to find the transformation matrix $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]^T$, the MLP classifier is trained. In the conventional MLP FE, both the hidden and output layers are used for FE and minimisation of the classification error. In the novel FE technique, the linear part of the hidden layer is used for FE and the rest is used for performance evaluation. Thus, the novel approach differs from MLP in two aspects:

1. Instead of outputs of the hidden neurons as the new features the weighted sums $W_1^T\mathbf{X} + w_{1,0}$, $\mathbf{W}_2^T \mathbf{X} + w_{2,0}$, ... are used.
2. The modified cost function is applied as the criterion in order to minimise the empirical classification error.

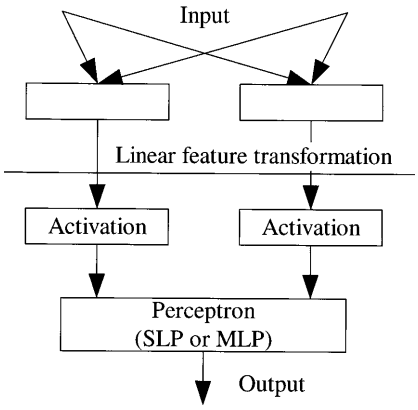

Fig. 4. Principal MLPLFE technique flow diagram.

In the new extracted feature space, the proposed linear mapping method with nonlinear SLP in its output stage can be limited complexity. In the case of the two new features, the maximum complexity shape is this: ⌴⌜. In higher feature cases, it can be more complex. In principle, instead of the SLP the MLP can be utilised. To ensure convergence of the MLP with two hidden layers applied to complex nonlinear data sets, adequate MLP initialisation and training strategies should be developed.

An important aspect of the novel approach is the utilisation of the anti-regularisation term. This term is used in order to ensure the minimisation of the frequency of misclassifications (empirical error). The taxonomy of the feature extraction algorithms discussed above shows that the novel algorithm is linear (Fig. 1a). It also belongs to the group of the supervised-nonlinear (Fig. 1b) algorithms, and uses the classification error as the performance measure. Minimisation of the empirical classification error does not rely on assuming the distributions of the pattern classes to be normal. Thus, as can be seen from taxonomy diagram, there is only one algorithm that satisfies this description.

4. DATA

Several data sets were used for comparative experiments. First, the new method was tested on specially created artificial data sets. After analysis of peculiarities of the new algorithm with the artificial data, a number of real world data sets were utilised to evaluate the usefulness of the new approach. Table 1 contains brief information about the number of classes, features and size of each data set.

The *3Gauss* artificial data set is a mixture of three non-overlapping multivariate Gaussian components with means located on one line. The left and the right distributions components belong to the first class, the middle one belongs to the second class. Data was generated in a two-dimensional space, and later to make task more difficult, to each two dimensional vector, six-variate random vectors were added to form a final eight-dimensional vector. Finally, a random, orthogonal space transformation was made in order to hide the first two features between the noise features.

The *Palm* data set has a palm shaped decision boundary in a two-variate subspace. The first pattern class is inside the 'palm' shape and the second class is outside. As in the first artificial data set, two *small* noise features were added to original two-dimensional data, and later to hide the first features between the noise features, the four-variate data was orthogonally transformed. The distribution of the two most informative features is widest. This was done on purpose in order to help the Principal Component analysis to extract features in the best way. The results in Table 2 demonstrate this.

A short description of the real world data sets is presented in Table 1. More details about the data sets can be found elsewhere [19–21].

Table 1. List of data sets used in experiments

Nr	Name	Features	Classes	Patterns	Short description
1	3 Gauss	6	2	200 + 200	3 not overlapped Gaussian distributions spread sequentially. The first and last distributions belong to first class, the middle one–to the second class.
2	Palm	3	2	500 + 500	The first class is inside the palm shape and the second class is outside.
3	Ionosphere	33	2	126 + 225	Radar signals from the ionosphere. Arguments of an auto correlation function.
4	Stock	69	2	697 + 683	4 years daily history of stock market closing prices. Classes are growth vs fall of the index.
5	Chromosome	30	2	499 + 501	Features describe the geometry of the chromosome.
6	Wave data	12	6	104 + 101 + 133 + 100 104 + 108	Represents six different human pronounced phonemes.
7	Dow	64	2	134 + 251	Machine vibration data classified into 'good' and 'bad', velocity and acceleration characteristics.
8	Vowel	28	2	400 + 400	Vowels pronounced by 20 speakers, 28 spectral and cepstral features.
9	Thyr	18	2	93 + 191	Healthy and hypothyroid patients.
10	Sonar	60	2	111 + 97	Sonar signals patterns, features characterise energy within particular frequency.
11	Satim	9*4	6	961 + 415 470 + 1038	Vectors representing satellite images characterised by nine pixel values in four spectral bands.
12	Musk	166	2	207 + 29	Musk and non musk molecules conformations. Features represent parameters of the shape.
13	Mammo	18 + 47	2	57 + 29	Benign and malignant mammograms. Features as: number, shape, size, a texture, histogram statistics, Gabor wavelet response, etc.
14	Call	8*3	2	134 + 231	Phone call intensity. Classes–number of calls will increase vs. decrease tomorrow.

Table 2. Nearest neighbour classifier, misclassification error (in %), on two extracted features

Nr	Classes	Data Set Name	Original	PC	Fol.Sam.	MLPLFE	AANN3	AANN5	MLP9
1	2	3 gauss (synthet)	8	39	46.5	2.2	**2**	35.7	0.25
2	2	Palm (synthet)	*1.7*	**0.5**	8.7	5.5	4.4	1.6	*3.3*
3	2	Ionosphere	*15.9*	26.7	8.5	**1.1**	19.3	13.6	*0*
4	2	Stock	*61.3*	55.2	**14.1**	28.0	48.6	46.6	*19.4*
5	2	Chromosomes	*2.3*	18	9.6	**1.6**	11.2	12.4	*0.3*
6	6	Wavedata	*15.5*	51.2	—	**27.8**	56.1	51.6	—
7	2	Dow	*8.0*	44.9	3.8	**0.7**	29.8	27.7	*0*
8	2	Vowel	*0.1*	15	1.8	**0.3**	14.7	11.3	*0.1*
9	2	Thyr	6.6	4.5	1.7	**0.3**	9.1	7.7	*0*
10	2	Sonar	*12.0*	42.3	15.8	**4.8**	32.2	28.8	*0*
11	2	Satim	*9.4*	21.1	—	**17.5**	22.5	30.3	—
12	2	Musk	*14.0*	35.0	7.5	**3.7**	28.3	33.6	*0.2*
13	2	Mamo	*34.8*	33.7	0	**0**	24.4	25.5	*0*
14	2	Call	*38.9*	40.2	36.4	**35.3**	40	36.1	*14.5*

5. EXPERIMENTS

In the experiments, the two best features were extracted to enable visualisation in two-dimensional space. The dimensionality reduction sometimes decreases classification accuracy (if data cannot be mapped on a two-dimensional plane), but it allows visualisation. This makes it easier to understand the data structure.

To compare results numerically, each method's two best extracted features were evaluated. The evaluation was made using the Nearest Neighbour (NN) classifier. This classifier used Euclidean distance measure, and the leave-one-out classification error estimation technique [14]. At first, all methods under investigation were compared on artificial data, and later, on all the real world data sets. In the experiments, four evaluations were skipped (see Table 2). The Folley-Sammon algorithm can process only two class data, so the six class *Wavedata* data set and six class *Satim* data set was not processed. The conventional MLP FE algorithm can extract the same number of features as the number of classes. So, these experiments were skipped as well.

6. RESULTS

Results of the experiments are presented in Table 2. The abbreviation AANN3 stands for Auto Associative Neural Network with 3 layers, one layer for encoding and one for decoding the data. AANN5 is the same as AANN3 but contains five layers and two are for encoding and other two are for decoding. MLP9 stands for the multilayer perception with nine hidden units and two outputs used as new features. Winning algorithms are presented in bold type. The results of incomparable algorithms are in *italics*.

In principle, it is not possible to make a numerical comparison between the feature set extracted by MLP9 and the other methods, because MLP9 destroys the geometry of the data. Therefore, it extracts the best features for classification. In the new MLPLFE method, accuracy is sacrificed in order to obtain good visualisation. The *Palm* data set can be classified correctly with only a minimum of nine hidden neurons. MLP9 has nine hidden neurons, while the MLPLFE has only two hidden neurons. Thus, in principle, the MLPLFE algorithm with the single layer perception in its output and two new features extracted *cannot perform well* on such complicated data as the *Palm* data set. For complicated data, in the output stage of the MLPLFE algorithm, the MLP with a sufficient number of hidden neurons should be utilised. In most real world problems, however, the complexity of simple MLPLFE with SLP in its output was sufficient to reveal the data structure.

On the *3Gauss* artificial data set, the AANN3 produced the best results, and the MLPLFE algorithm was ranked second. As predicted by the theoretical considerations, the synthetic *Palm* data set was too complicated for the simple MLPLFE algorithm (see Figs 5 and 6). The MLPLFE results using the *Palm* data set were similar to those produced by the Foley-Sammon and AANN3 algorithms. However, the

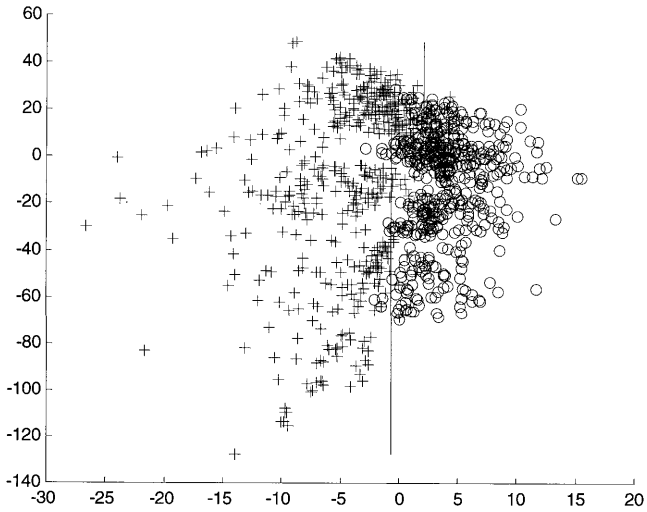

Fig. 5. Two features extracted using the MLPLFE technique from synthetic palm data. Straight line represents decision boundary formed by two hidden neurons.

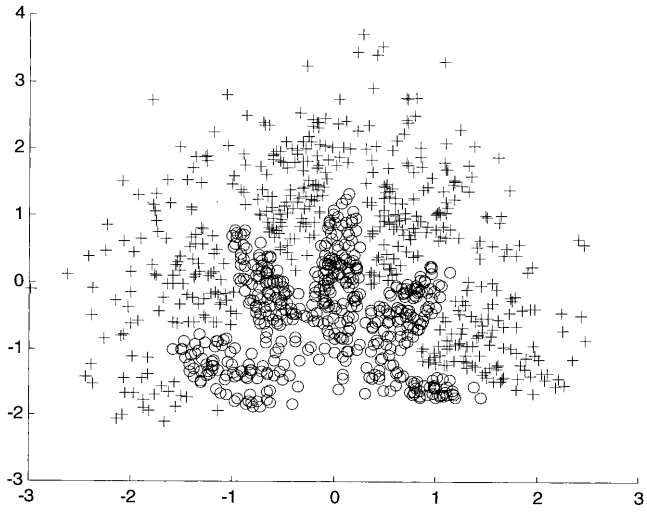

Fig. 6. Original palm data.

Principal Components algorithm produced the best results here. It is because the distribution of data was widest in informative features, and the PC method extracts features with the widest distribution.

The first two features extracted from the *Ionosphere* data are visualised in Figs 7, 8, 9 and 3. As can be seen (circles-first class, pluses-second) in Fig. 9, the MLPLFE extracted features are the best ones, i.e. the pattern classes could be separated in the best possible way. The MLP9 algorithm (see Fig. 3) gave the best separation. However, it destroyed the geometry of the data completely.

The novel method applied to the *Stock* data set showed lower accuracy than the Foley-Samon algorithm. This result may be explained by a fact that actually the *Stock* data set is linearly separable and the Foley-Samon algorithm is designed to process linearly separable data.

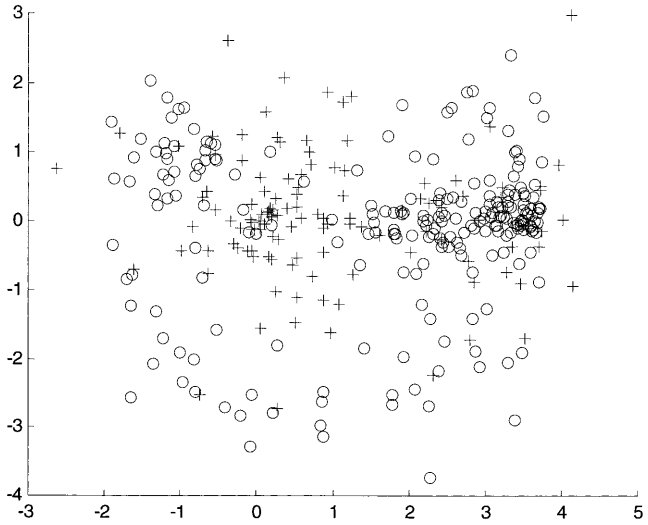

Fig. 7. First two principal components extracted from ionosphere data.

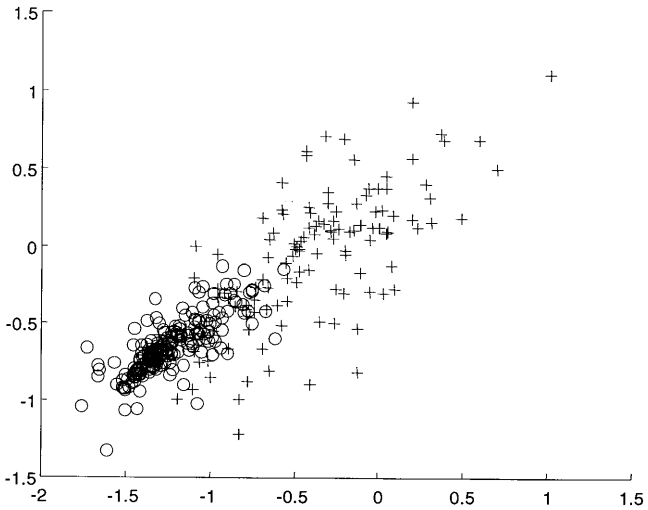

Fig. 8. First two Sammon features extracted from ionosphere data.

An interesting observation can be made regarding the *Musk* data set. The new MLPLFE algorithm was more computationally efficient than the Foley-Sammon algorithm on this data set. This is because of the large number (166) of features in the data set. The Foley-Sammon algorithm is based on LDA and requires estimation and inversion (or pseudo inversion) of the covariance matrix. The 166×166 matrix takes a relatively long time to invert. The MLPLFE scales up linearly, whereas the Foley-Sammon method scales up quadratically with the number of input dimensions.

The MLP classifier based FE (MLP9) cannot be included in the comparison table because it destroys the geometry of the data (see Fig. 3). This fact makes the Multilayer perceptron FE absolutely useless for visualisation. However, it actually produces the best classification results.

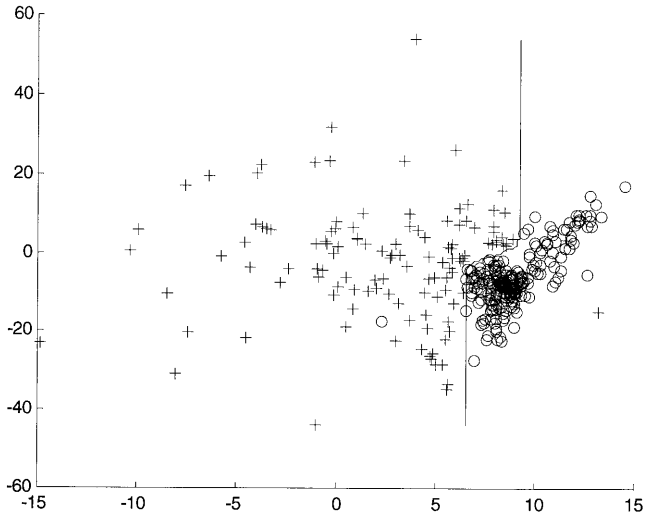

Fig. 9. Two features extracted using the MLPLFE technique from ionosphere data. Straight line represents decision boundary formed by two hidden neurons.

7. CONCLUSIONS

In the new extracted feature space, the proposed linear mapping method with nonlinear SLP in its output stage generates a nonlinear decision boundary of limited complexity (⌐¬ shaped in the two new feature case) and minimises the empirical classification error. It is good at analysing simple nonlinearly separable data sets. The experiments showed that the novel MLPLFE runs well on the majority of the two synthetic and 12 real world data sets from different domains. In many cases, the method proposed outperformed the Principal Components, Foley-Sammon, Auto Associative Neural Networks (with 3 and 5 layers) and MLP based feature extractor techniques. It was found, however, that there is no single mapping technique which is good in all situations.

Several conclusions can be drawn.

Advantages

- The MLPLFE with nonlinear SLP classifier in its output stage works well with simple nonlinear data, however, it fails with highly complex nonlinear data sets.
- The method is faster than the Foley-Sammon method where the data set has high dimensionality.
- The method uses linear data transformation and a nonlinear performance criterion. Hence, it is easier to understand how the data is distributed in its original space.
- Transformation back to the original space can be performed if required (some nonlinear transformations do not allow this).

Disadvantages

- The method is slower than the Principal Components or the Foley-Sammon (in low dimensionality cases) methods.
- The method with the SLP in its output stage may not be as good as the MLP FE algorithm in overall accuracy,

however, this may be offset by the possibility of superior visualisation.

Some suggestions can be made for future research. The method reported above can be enhanced to work with much more complex nonlinear data. For this, in the upper part of the data processing schema (used for the performance evaluation), instead of the nonlinear SLP, the MLP with sufficiently large number hidden nodes must be used. To ensure convergence of the MLP with two hidden layers applied to complex nonlinear data sets, adequate MLP initialisation and training strategies should be developed. In the present paper, data sets were not divided into training and testing samples. It may be well to investigate doing so in future research.

Acknowledgements

Authors acknowledge the anonymous referees for valuable suggestions.

References

1. Mao J, Jain A. Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans Neural Networks 1995; 6: 296–317
2. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE Trans Pattern Analysis and Machine Intelligence 2000; 22(1): 4–37
3. Jahn A, Gloger JM, Franke J. A Comparison of Linear Dimension Reduction Algorithms using Different Feature Extraction Methods. STIPR'97, 1997; 79–84
4. Backer SD, Naud A, Scheunders P. Non-linear dimensionality reduction techniques for unsupervised feature extraction. Pattern Recognition Letters 1998; 19: 711–720
5. Gelsma ES, Eden G. Mapping algorithms in ISPAHAN. Pattern Recognition 1980; 3: 127–136
6. Jain AK, Dubes R. Feature definition in pattern recognition with small sample size. Pattern Recognition 1978; 10: 85–97
7. Lerner B, Guterman H, Aladjem M, Dinstein I. A comparative study of neural network based feature extraction paradigms. Pattern Recognition Letters 1999; 20(1): 7–14
8. Lerner B, Guterman H, Aladjem M, Dinstein I, Romen Y. On pattern classification with Sammon's nonlinear mapping – an experimental study. Pattern Recognition 1998; 31(4): 371–381
9. Siadliecki W, Siadlecka K, Sklansky J. An overview of mapping techniques for exploratory pattern analysis. Pattern Recognition 1988; 21(5): 411–429
10. Raudys A. A non-parametric data mapping technique for active initialization of the multilayer perceptron. Joint IAPR Int Workshops/SSPR'98 and SPR'98, 1998; 989–996
11. Intrator N. Feature extraction using unsupervised neural network. Neural Computation 1992; 4(1): 98–107
12. Pao Yh, Shein Cy. Visualization of pattern data through learning of non-linear variance-conserving dimension-reduction mapping. Pattern Recognition 1997; 30(10): 1705–1717
13. Sammon JW. A nonlinear mapping for data structure analysis. IEEE Trans Computers 1969; 18: 401–409
14. Fukunaga K. Introduction to Statistical Pattern Recognition. 2nd ed. Computer Science and Scientific Computing. Academic Press, 1990; 591
14. Sammon JW. An optimal discriminant plane. IEEE Trans Computers 1970; C-19: 826–829
15. Bourland H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biol Cybernet 1988; 59: 291–294
16. Tubbs JD, Coberley WA, Young DM. Linear dimension reduction and Bayes classification with unknown parameters. Pattern Recognition 1982; 14(3): 167–172
17. Foley DH, Sammon JWJ. An optimal set of discriminant vectors. IEEE Trans Computers 1975; C-24: 281–289
18. Baladi P, Hornik K. Neural networks and principal component analysis. Learning from examples without local minima. Neural Networks 1989; 2: 53–58
19. Raudys S. Evolution and generalization of a single neurone: I. Single-layer perceptron as seven statistical classifiers. Neural Networks 1998; 11(2): 283–296
20. Raudys A, Mockus J. ARMA and perceptron comparison on economical time series. Informatica 1999; 10(2): 231–244
21. Saudargiene A. Structurization of the covariance matrix by process type and block diagonal models in the classifier design. Informatica 1999; 10(2): 245–269

James Allen Long received a BS, in mathematics at an American university, and a MSc in computer science and a PhD in logic from Bristol University, UK. He has authored a number of conference/journal papers in pattern matching and AI applications in business.

Aistis Raudys received a BS, in computer science from Vilnius University, Lithuania in 1996 and an MS computer science degree from Kaunas University of Technology, Lithuania in 1999. From 1996 to 1998, he was a research assistant at the Data Analysis Department of the Institute of Mathematics and Informatics, Vilnius, Lithuania. Presently he is a PhD student. Since 1999 he has been at the Department of Computing, South Bank University, London, UK. His research interests are in artificial neural networks, feature extraction, and computationally efficient and large-scale data mining.

Correspondence and offprint requests to: A. Raudys, School of Computing Information Systems and Mathematics, South Bank University, 103 Borough Road, London SE1 0AA, UK. Email: raudysa@sbu.ac.uk