

Feature Extraction by Neural Network Nonlinear Mapping for Pattern Classification

B. Lerner, H. Guterman, M. Aladjem, and I. Dinstein

Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel

Abstract

Feature extraction has been always mutually studied for exploratory data projection and for classification. Feature extraction for exploratory data projection aims for data visualization by a projection of a high-dimensional space onto two or three-dimensional space, while feature extraction for classification generally requires more than two or three features. Therefore, feature extraction paradigms for exploratory data projection are not commonly employed for classification and *vice versa*. We study extraction of more than three features, using neural network (NN) implementation of Sammon's nonlinear mapping to be applied for classification. Comparative classification experiments reveal that Sammon's method, which is primarily an exploratory data projection technique, has a remarkable classification capability. The classification performance of (the unsupervised) Sammon's mapping is highly comparable with the performance of the principal component analysis (PCA) based feature extractor and is slightly inferior to the performance of the (supervised) multilayer perceptron (MLP) feature extractor. The paper thoroughly investigates a random and a non-random initializations of Sammon's mapping. Only one experiment of Sammon's mapping is required when the eigenvectors corresponding to the largest eigenvalues of the sample covariance matrix are used to initialize the projection. This approach tremendously reduces the computational load and substantially raises the classification performance of Sammon's mapping using only very few eigenvectors.

The 13th International Conference on Pattern Recognition, ICPR13, Vienna, vol. 4, 320-324, 1996. Corresponding author: Boaz Lerner, University of Cambridge Computer Laboratory, New Museums Site, Cambridge CB2 3QG, UK. Email: boaz.lerner@cl.cam.ac.uk

1. Introduction

Feature extraction is the process of mapping the original features (measurements) into fewer features which include the main information of the data structure. A large variety of feature extraction methods based on statistical pattern recognition or on artificial neural networks appears in the literature [1]-[9]. In all the methods, a mapping f transforms a pattern \underline{y} of a d -dimensional feature space to a pattern \underline{x} of an m -dimensional projected space, $m < d$, i.e.,

$$\underline{x} = f(\underline{y}), \quad (1)$$

such that a criterion J is optimized. The mapping $f(\underline{y})$ is determined amongst all the transformations $g(\underline{y})$, as the one that satisfies [9],

$$J\{f(\underline{y})\} = \max_g J\{g(\underline{y})\}. \quad (2)$$

The mappings differ by the functional forms of $g(x)$ and by the criteria they have to optimize.

Feature extraction methods can be grouped into four categories [4] based on *a priori* knowledge used for the computation of J : supervised *versus* unsupervised, and by the functional form of $g(x)$: linear *versus* nonlinear. In cases where the target class of the patterns is unknown, unsupervised methods are the only way to perform feature extraction. In other cases, supervised paradigms are preferable. Linear methods are simpler and are often based on an analytical solution but they are inferior to nonlinear methods when the classification task requires complex hypersurfaces. Widespread unsupervised methods for feature extraction are PCA [3], [9] (a linear mapping) and Sammon's nonlinear mapping [6]. The PCA attempts to preserve the variance of the projected data, whereas Sammon's mapping tries to preserve the interpattern distances. The MLP when acting as a feature extractor provides a supervised nonlinear mapping of the input space into its hidden layer(s).

Feature extraction for exploratory data projection enables high-dimensional data visualization for better data structure understanding and for cluster analysis. In feature extraction for classification, it is desirable to extract high discriminative reduced-dimensionality features which reduce the classification computational requirements. However, feature extraction criteria for exploratory data projection regularly aim to minimize an error function, such as the mean square

error or the interpattern distance difference whereas feature extraction criteria for classification aim to increase class separability as possible. Hence, the optimum extracted features (regarding a specific criterion) calculated for exploratory data projection are not necessarily the optimum features regarding class separability and *vice versa*. In particular, two or more classes may have principal features that are similar. Moreover, feature extraction for exploratory data projection is used for two or three-dimensional data visualization, whereas classification usually needs more than two or three features. Consequently, feature extraction paradigms for exploratory data projection are not generally used for classification and *vice versa*.

This paper studies the application of feature extraction paradigms for exploratory data projection to be also employed for classification. It uses Sammon's nonlinear mapping which is primarily an exploratory data projection technique. The classification accuracy of a NN implementation of Sammon's mapping for more than three features is compared with the accuracy of the PCA based and the MLP feature extractors which are usually employed for classification. In addition, the paper extensively compares and investigates the trade-offs between a random and a nonrandom initializations of Sammon's mapping.

2. Paradigms of feature extraction for exploratory data projection and classification

Sammon [6] proposed a feature extraction method for exploratory data projection. This method is an unsupervised nonlinear paradigm that attempts to maximally preserve all the interpattern distances. We extend in this study the domain of the method to be applicable for classification purposes. The classification capability using Sammon's mapping is compared to two well-known feature extraction paradigms for classification. The first is the PCA which is an unsupervised linear paradigm and the second is the MLP feature extractor which is a supervised nonlinear paradigm. The outline of the experiments is shown in Fig. 1.

A. Sammon's mapping

The criterion to minimize in Sammon's mapping is Sammon's stress (error), defined as:

$$E = \frac{1}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n d^*(i,j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{[d^*(i,j) - d(i,j)]^2}{d^*(i,j)} \quad (3)$$

where $d^*(i,j)$ and $d(i,j)$ are the distances between pattern i and pattern j in the input space and in the projected space, respectively. The Euclidean distance is frequently used. Sammon's stress is a measure of how well the interpattern distances are preserved when the patterns are projected from a high-dimensional space to a lower dimension space. The minimum of Sammon's stress is achieved by carrying out a steepest-descent procedure. As in steepest-descent based approaches, local minima in the error surface is often unavoidable. This implies that a repetitive number of experiments with different random initializations have to be performed before the initialization with the lowest stress is obtained. However, several methods which make use of some knowledge of the feature data may be more effective. For example, the initialization could be based on the first norms of the feature vectors [2] or on the projections of the data onto the space spanned by the principal axes of the data [2], [4]. The second drawback of Sammon's mapping is its computational load which is $O(n^2)$. In each iteration $n(n-1)/2$ distances, along with the error derivatives, must be calculated. As the number of vectors (n) increases, the computational requirements (time and storage) grow quadratically.

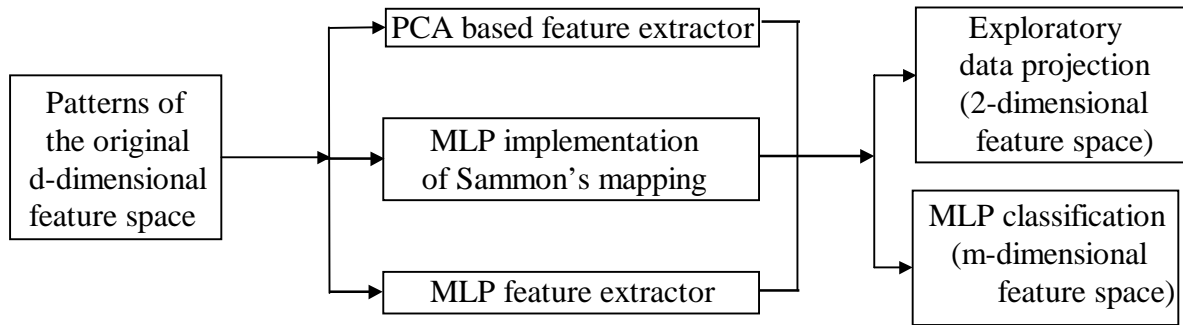


Fig. 1. The experiments' layout.

Mao and Jain [4] have suggested a NN implementation of Sammon's mapping. Fig. 2 shows the NN architecture they have used in their paper. It is a two layer feedforward network whereas the number of input units is set to be the feature space dimension, d , and the number of output

units is specified as the extracted feature space dimension, m . No rule for determining the number of hidden layers and the number of hidden units in each hidden layer is suggested. They derived a weight updating rule for the multilayer feedforward network that minimizes Sammon's stress based on the gradient descent method. The general updating rule for all the hidden layers, $l=1, \dots, L-1$ and for the output layer ($l=L$) is:

$$\begin{aligned} \Delta \omega_{jk}^{(l)} &= -\eta \frac{\partial E_{\mu\nu}}{\partial \omega_{jk}^{(l)}} = \\ &= -\eta (\Delta_{jk}^{(l)}(\mu) y_j^{(l-1)}(\mu) - \Delta_{jk}^{(l)}(\nu) y_j^{(l-1)}(\nu)) \end{aligned} \quad (4)$$

where ω_{jk} is the weight between unit j in layer $l-1$ and unit k in layer l , η is the learning rate, $y_j^{(l)}$ is the output of the j th unit in layer l and μ and ν are two patterns. The $\Delta_{jk}^{(l)}$'s are the errors accumulated in each layer and backpropagated to a preceding layer, similarly to the standard backpropagation. However, in the NN implementation of Sammon's mapping the errors in the output layer are functions of the interpattern distances.

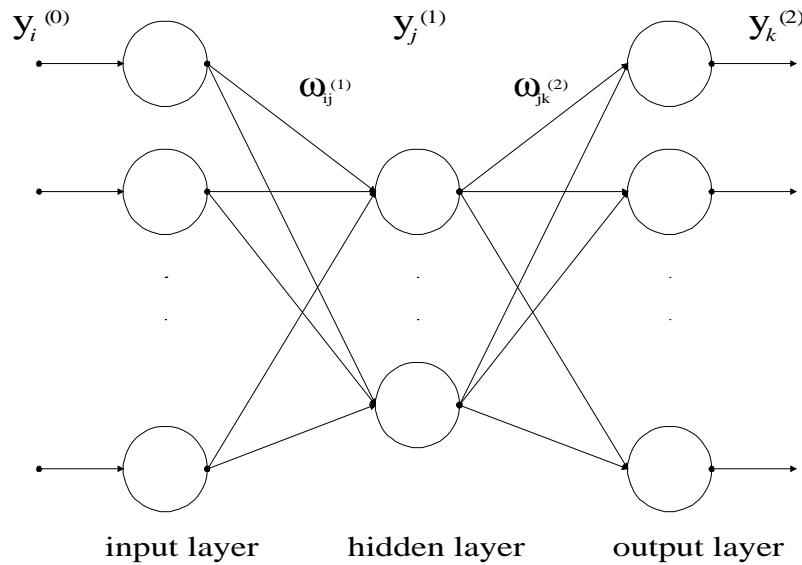


Fig. 2. A two-layer perceptron NN for Mao and Jain's implementation of Sammon's mapping and for the MLP feature extractor.

In Mao and Jain's implementation the network is able to project new patterns after training, a property Sammon's mapping does not have. Mao and Jain suggested to use data projections along

the PCA axes as an initialization to Sammon's mapping. They employed a two stage training phase using the standard backpropagation algorithm for the first stage and their modified unsupervised backpropagation algorithm for a refinement in the second stage. Our Sammon's mapping study has been stimulated by Mao and Jain's research. The NN based Sammon's mapping implementation we use is similar to the implementation suggested by Mao and Jain but it is simpler. Only one training stage using Mao and Jain's unsupervised backpropagation algorithm (their second stage) is used. In addition, Mao and Jain in their research employed a PCA based initialization for Sammon's mapping whereas we employed and compared both random and PCA based initializations.

B. The PCA based feature extractor

Among the unsupervised linear projection methods the PCA is probably the most widely used. The PCA, also known as the Karhunen-Loe've expansion, attempts to reduce the dimensionality of the feature space by creating new features that are linear combinations of the original features. The procedure begins with a rotation of the original data space followed by ranking the transformed features and picking out few projected features. This procedure finds the subspace in which the original sample vectors may be approximated with the least mean square error for a given dimensionality.

Let $\underline{x} = f(\underline{y})$ be a linear mapping of a random feature vector \underline{y} , $\underline{y} \in \mathbb{R}^d$, $\underline{x} \in \mathbb{R}^m$ and $m < d$. The approximation $\hat{\underline{y}}$,

$$\hat{\underline{y}} = \sum_{j=1}^m \underline{x}_j \underline{u}_j \quad (5)$$

with the minimum mean square error,

$$\epsilon = \mathbf{E} \left\{ (\underline{y} - \hat{\underline{y}})^t (\underline{y} - \hat{\underline{y}}) \right\} \quad (6)$$

is obtained when \underline{u}_j ($\forall j=1, m$) are the eigenvectors associated with the m largest eigenvalues λ_j of the covariance matrix Ψ of the mixture density ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq \dots \geq \lambda_d$). The expansion coefficient x_j associated with \underline{u}_j is the j th PCA feature of x ,

$$\underline{x}_j = \underline{u}_j^t \underline{y}. \quad (7)$$

C. The MLP feature extractor

When acting as a classifier, the MLP hidden unit outputs can be used as an implementation of a nonlinear projection of high-dimensional input (feature) space to a much simpler (abstract) feature space [10]. Patterns represented in this space are more easily separated by the network output layer. Furthermore, visualization of the last hidden internal representations may supply an insight to the data structure, hence, to play as a mean of data projection. Using this approach, the classifier acts ideally as feature extractor and as exploratory data projector. Although not acting as a classifier, the MLP feature extractor training is based on class label information, hence it is supervised. The number of input units (Fig. 2) is specified to be the number of features and the number of output units to be the number of pattern classes. The hidden layer dimension is set according to the task, either exploratory data projection or a classification.

3. The experiments

A. The data set

The data set was derived from chromosome images which were gathered in Soroka Medical Center, Beer-Sheva, Israel. The chromosome images were acquired and segmented in a process described elsewhere [11]. The experiments were held with 300 patterns from three types of chromosomes (types "13", "19" and "x"), 100 patterns from each type. The chromosome patterns were represented in feature space by 64 density profile (d.p.) features (integral intensities along sections perpendicular to the medial axis of the chromosome) [11].

B. The classifier

A two layer feedforward NN trained by the standard backpropagation learning algorithm was chosen to be used as a classifier. The number m , of input units was set by the projected space dimension and the number of output units was determined by the number of classes (three classes in our case). Higher complex architectures were not considered as candidates for the classifier because only low-dimensional extracted features were employed as the classifier input. The classifier parameters which were adapted for the chromosome data in a previous investigation [12] were: learning rate of 0.1, momentum constant of 0.95, 10 hidden units and a training period

of 500 epochs. Each experiment with the classifier was repeated ten times with different randomly chosen initial weight matrices and the results were averaged. Although only one experiment of the classifier is sufficient to compare the feature extraction paradigms, averaging over several classifier initializations yields more objective results. Exactly the same ten classifier initializations were used for examining all the feature extraction paradigms.

C. The methodology

C1. General

The general scheme of the experiments was outlined in Fig. 1. As Fig. 1 indicates, the paradigms extract features from the 64-dimensional chromosome patterns. The outputs of the three feature extraction paradigms are used to project the samples into two-dimensional maps and to train and test the MLP classifier. The two-dimensional projection maps are visually analyzed and compared to the two-dimensional scatter plots of two of the original features. The probability of correct classification of the test set is evaluated for one to seven extracted features and compared to these probabilities based on the first 10 and all the 64 d.p. features. The first 10 d.p. features which are extracted from the upper tip of the chromosome, provide the cytotechnician an enhanced discriminative capability. In addition, they were ranked by a feature selection algorithm among the best d.p. features [11].

Twenty-one randomly chosen training and test sets were derived from the entire chromosome data set for the classification experiments. Each training set contained randomly selected 90% of the data set while the remainder patterns were reserved for the test (the holdout method [3]). Each feature extraction paradigm was applied to these data sets. Classification results were averaged over the twenty-one data sets and the ten classifier initializations (see Sec. 3B).

C2. Sammon's mapping

In a preliminary study, ten random generator seeds were tested to initialize Sammon's mapping. The seed which was responsible for the highest classification performance was chosen to initialize the weight matrices of the random initialization. The second initialization of Sammon's mapping was based on all the eigenvectors of the sample covariance matrix estimated from the training data set. In the exploratory data analysis experiments, the two Sammon's

projections were obtained by setting the network output dimension to 2. For the classification experiments the network output dimension was changed in the [1,7] range. Sammon's mapping parameters for both initializations were: learning rate of 1, momentum constant of 0.5, 20 hidden units and a training period of 40 epochs. This set of parameters yielded the ultimate performances in a preliminary study.

C3. The PCA based feature extractor

In this study we use the classical implementation of PCA. The eigenfeatures of the training set were sorted by a descending order corresponding to the eigenvalue magnitudes. The first one to seven eigenfeatures were used for classification and the first two eigenfeatures were used to plot the two-dimensional projection map.

C4. The MLP feature extractor

A two layer perceptron NN trained by the backpropagation algorithm was employed as a feature extractor. The input layer was 64-dimensional and the output was 3-dimensional (3 classes). The number of the hidden layer units was set to be 2 in the exploratory data projection experiments and it was changed from 1 to 7 in the classification experiments.

D. Results

D1. Projection maps

A comparison is made between the two-dimensional projection maps of the chromosome feature set projected by the three feature extraction paradigms. The evaluation of the projection maps is only based on visual judgment which is, to our opinion, the most qualitatively way to evaluate these maps, except for complex psychophysical experiments. Furthermore, a quantitative evaluation of the feature extraction paradigms for projection purposes is appeared to be inherently biased toward one of the paradigms. For example, Sammon's stress, when was used to evaluate projection methods, ranked Sammon's mapping as the best projection method [4]. To our knowledge, there is no criterion to judge objectively the projection methods.

In Fig. 3, the two-dimensional projection maps of the three paradigms are given. For a comparison, a scatter plot of the first two original d.p. features is given in Fig. 3a. These two

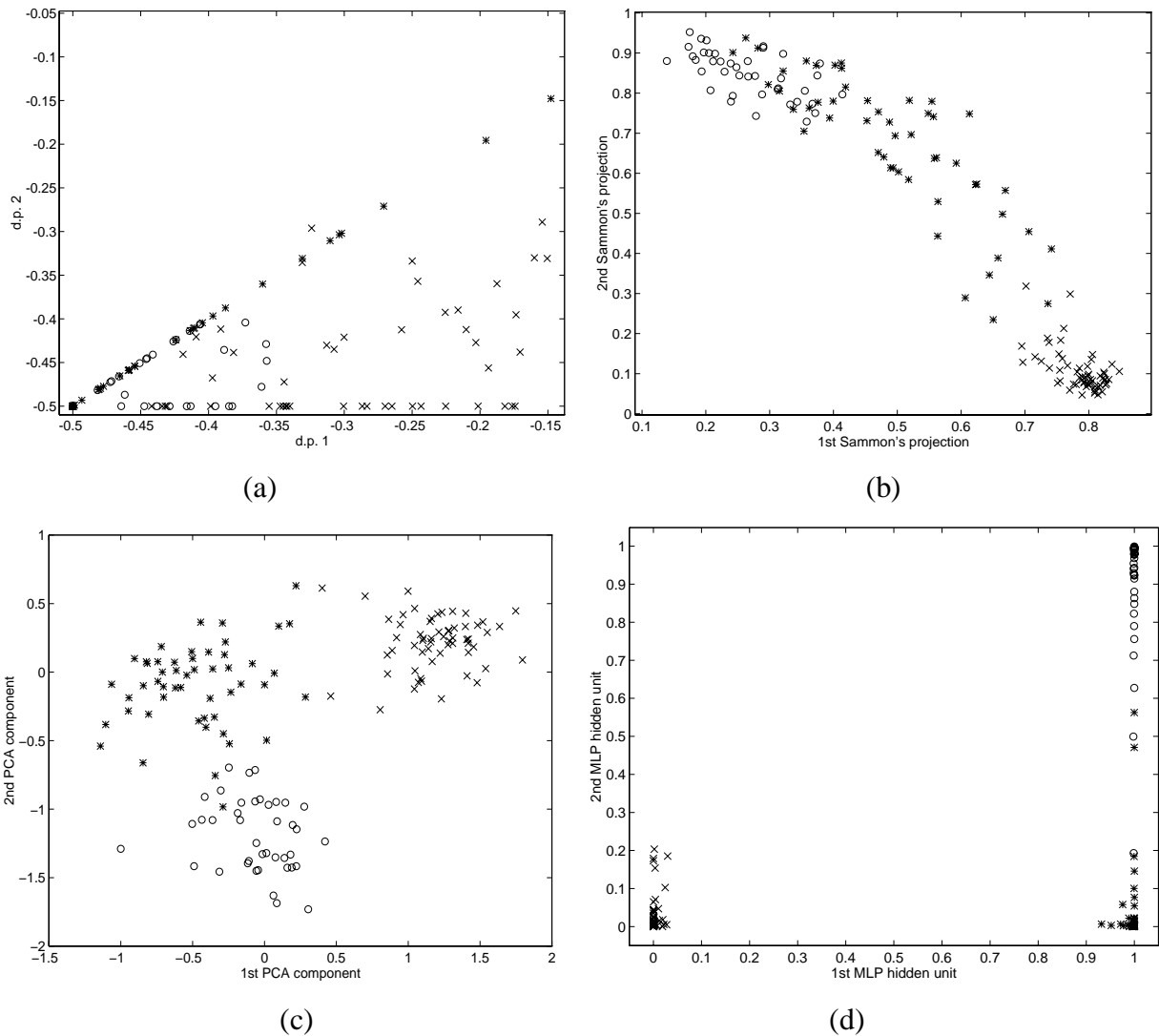


Fig. 3. The two-dimensional projection maps of: (a) two d.p. features (the 1st and the 2nd), (b) Sammon's mapping, (c) the PCA based feature extractor and (d) the MLP feature extractor (o, * and x for chromosome types 13, 19 and x, respectively).

features are amongst the most discriminative d.p. features [11]. The second projection map (Fig. 3b) is formed by Sammon's mapping onto two-dimensional space. Random initialization is preferred to be used in Sammon's mapping experiments mainly because the PCA based initialization is frequently yields very similar maps to the PCA based feature extractor maps [4]. The third projection map (Fig. 3c) is produced by data projection along the two principal components. The fourth projection map (Fig. 3d) is produced by the two hidden units of the two layer perceptron feature extractor. All the maps are based on the test set and on the parameters of

the networks which were previously specified (Sections 3C2-3C4). The maps were obtained in an experiment in which 50% of the data set were used for training. Producing the same maps for the case which was experimented in the classification experiment (90% of the data set used for training) is of less interest because only ten test patterns per class were available for the experiment. The figure reveals the difference in the way the three feature extraction paradigms project data. Visually analyzed, the maps of the PCA based and the MLP NN feature extractors are more perceptive than the map of Sammon's mapping and the pattern spread is more evident. Moreover, the ratio of the cluster between scatter to the cluster within scatter of these two maps is larger. Not to be forgotten however, that projecting along the axes with the data largest and second largest variances, as the PCA does, is the easiest way to interpret projection maps. Considering discriminative power, the map of the MLP feature extractor is superior. It is important to mention, however, that the MLP is a supervised feature extraction paradigm where the other two are unsupervised. However, the MLP severely distorts the structure of the data and the interpattern distances while the PCA based feature extractor and Sammon's mapping preserve them very well. Another interesting point to observe is the way the MLP shrinks each class pattern to almost one point (or line), a quality which eases the classification process. These shrunk clusters are (almost) concentrated in three of the four map corners corresponding to the ultimate values of the hidden unit activation function (sigmoid). All the projection maps, especially these of the PCA based and the MLP paradigms reveal that the projected features are less correlated between themselves than the original features (Fig. 3a).

D2. Classification

We have used the MLP NN probability of correct classification of the test set as the criterion to evaluate the classification performances of the three feature extraction paradigms. The comparison of the probabilities of correct classification using the three feature extraction paradigms is given in Fig. 4 for 1 to 7 extracted features. Each point in the graph is an average over 210 experiments (see Sec. 3C1). For a comparison, the probabilities of correct classification using the original first 10 and all the 64 d.p. features are 86.6% and 83.7%, respectively.

As is shown in Fig. 4, the MLP feature extractor is responsible for achieving the best probability of correct classification. Sammon's mapping and the PCA based feature extractor

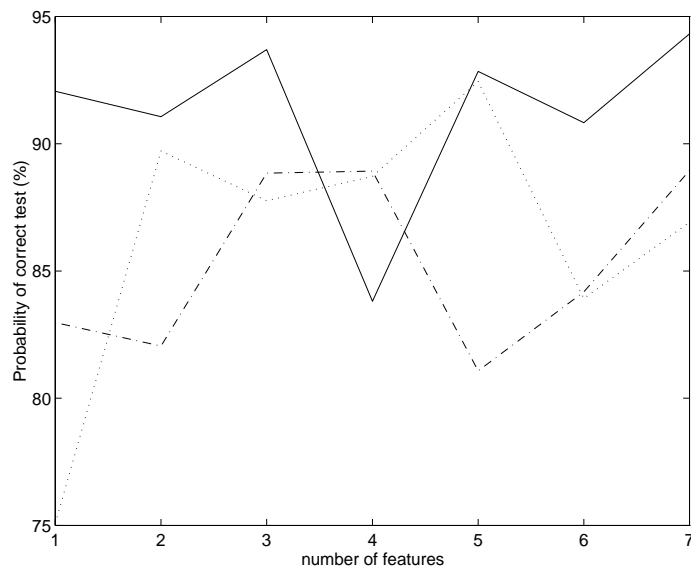


Fig. 4. The probability of correct classification using the three paradigms for increasing number of extracted features (\cdot for the PCA, $- \cdot$ for Sammon's mapping (random initialization) and solid line for the two layer perceptron).

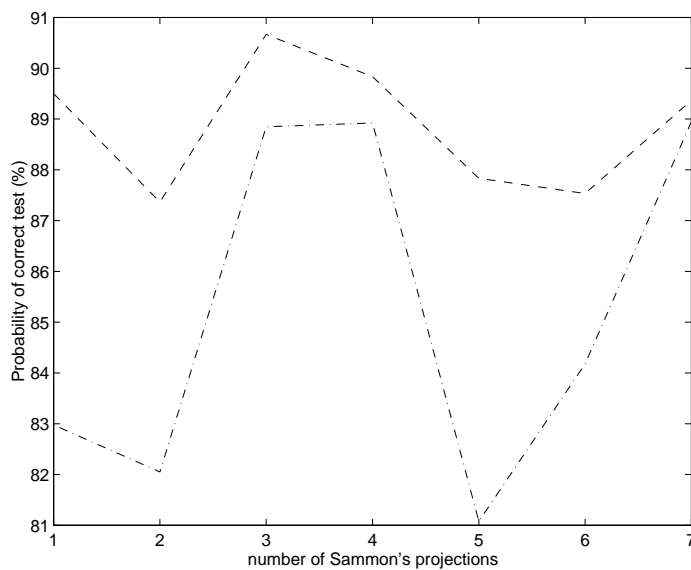


Fig. 5. The probability of correct classification based on two initializations of Sammon's mapping for increasing number of projections ($- \cdot$ for the random initialization and $--$ for the PCA based initialization).

lead to similar results which are inferior to the MLP feature extractor. Only three extracted features are needed using each of the paradigms to achieve superior classification performances compare to these achieved by the first 10 or all the 64 d.p. features. In Fig. 5 the random and the PCA based initializations of Sammon's mapping are compared. The experiments were held for the same previous ranges of projections as in Fig. 4. The superiority of the PCA based initialization over the random initialization is apparent. Moreover, decreasing randomness aids the PCA based initialization to achieve more stable classification results than the random initialization.

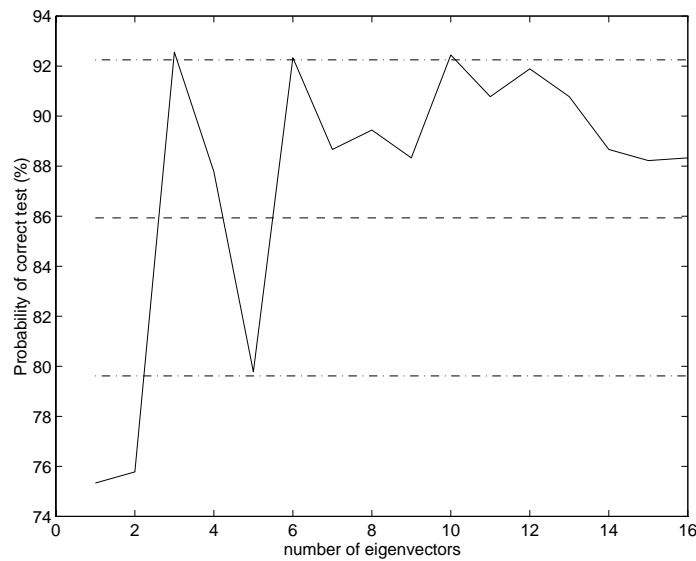


Fig. 6. The probability of correct classification using 2 Sammon's mapping projections. The average (dashed line) and the standard deviation (dashdot line) of 10 random initializations are compared to the PCA based initialization (solid line) for increasing number of eigenvectors.

Fig. 6 presents the results of another experiment to compare both Sammon's mapping initializations for only two extracted features. Ten Sammon's mapping random initializations were examined and the average and the standard deviation of the probability of correct classification of the 10 experiments are plotted. The average and the standard deviation are compared to the probability of correct classification using the PCA based initialization when additional eigenvectors are appended to the input-hidden initial weight matrix. Fig. 6 shows that only very few (six or more) eigenvectors are sufficient to initialize the PCA based feature extractor to

outperform the average performance of the random based initialization. Furthermore, the substantial advantage of the PCA based initialization over the random initialization is that only one experiment of Sammon's mapping is required. The random initialization requires several experiments with different random generator seeds before selecting the best (or the averaged) initialization. Concerning the fact that the computational complexity of Sammon's mapping is $O(n^2)$ for n patterns, this advantage is crucial.

4. Discussion

We study the classification capabilities of the well-known Sammon's mapping, which is originally applied for exploratory data projection. A comparison of the classification performance of a NN implementation of Sammon's mapping [4] with the PCA based and the MLP feature extractors, is made. The three paradigms are evaluated using a chromosomal feature set.

Although originally aimed and used for exploratory data projection, Sammon's mapping has an admirable classification capability. Only one experiment of Sammon's mapping is required when the PCA eigenvectors of the sample covariance matrix corresponding to the largest eigenvalues are used to initialize the algorithm. This fact has an enormous computational impact on the feature extraction process. In addition, the improved initial directions, the PCA provides, enable a classification performance improvement based on only very few eigenvectors. A combination of a nonlinear feature extraction paradigm and class information improves the discriminative capability. The MLP feature extractor which is a supervised nonlinear paradigm, is found in this study as the best feature extraction paradigm for both exploratory data projection and classification.

References

1. H. Bourland, and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybern.*, vol. 59, pp. 291-294, 1988.
2. Y. Chien, *Interactive Pattern Recognition*. NY: Marcel Dekker, Inc., 1978.
3. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (2nd ed.). New York: Academic Press, 1990.

4. J. Mao, and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Trans. Neural Networks*, vol. 6, pp. 296-317, 1995.
5. E. Oja, "Principal components, minor components and linear neural networks," *Neural Networks*, vol. 5, pp. 927-935, 1992.
6. J. W. Sammon Jr., "A nonlinear mapping for data structure analysis," *IEEE Trans. Comput.*, vol. 18, pp. 401-409, 1969.
7. T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, pp. 459-473, 1989.
8. T. Kohonen, "The self organizing map," *Proc. IEEE*, vol. 78, pp. 1464-1480, 1990.
9. P. A. Devijver, and J. Kittler, *Pattern Recognition- A Statistical Approach*. NJ :Prentice Hall, 1982.
10. R. P. Gorman, and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, pp. 75-89, 1988.
11. B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Medial axis transform based features and a neural network for human chromosome classification," *Pattern Recognition*, vol. 28, pp. 1673-1683, 1995.
12. B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Human chromosome classification using multilayer perceptron neural network," *Int. J. Neural Systems*, vol. 6, pp. 359-370, 1995.