ICONIP'95, Beijing, China

Unsupervized Feature Extraction for Nonlinear Supervized Classification with Application to Chromosome Analysis

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

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

Abstract- Three unsupervized feature extraction techniques were studied and compared using nonlinear supervized classifier. The first technique was the well-known Principal Component Analysis (PCA). The second was the incomplete eigenfeature approach in which only part of the features was chosen by an elimination criterion prior to the projection by the PCA. Along with the two linear projection methods, a feedforward neural network implementation of the Sammon's nonlinear projection algorithm was used. The three techniques were compared with a two layer perceptron classifier using chromosome data. The first two techniques enable a designer a trade off between pre-selection of measurements prior to the projection and extraction of minimal number of eigenfeatures. It emerges that the choice of the feature extraction approach is highly dependent on the data to be classified and on the ultimate purpose of the extraction, classification or exploratory projection.

1. Introduction

Feature extraction can be viewed as finding a set of reduced dimensionality vectors that represents an observation. In pattern recognition, as well as in neural network based classification, it is desirable to extract features that are focused on discrimination between classes. The selection of features that contain the most discriminatory information is important because the cost of decision making is directly related to the number of features used in the decision rule. Although a reduction in dimensionality is desirable, the error increment due to the reduction in dimensionality must be constrained to be adequately small [1].

Feature extraction (or projection) methods can be grouped into four categories based on two factors [2]: (i) supervized versus unsupervized, and (ii) linear versus nonlinear. Discriminant analysis is the commonly used linear supervized feature extraction technique. Mostly, the extracted features must be independent of class membership which implies that the feature extraction method should be unsupervized. Popular unsupervized methods are principal component analysis (PCA) (a linear mapping) and Sammon's algorithm [3] (a nonlinear mapping). The PCA attempts to preserve the variance of the projected data, whereas Sammon's projection tries to preserve the interpattern distances.

Nonlinear supervized classification is probably the most investigated paradigm for neural network based classification. In most cases, the feature extraction stage is integrated into the classification network which is then trained with supervized learning algorithms. However, for complex problems huge training periods along with large training sets may be necessary by this attitude. Thus, it is usually more useful to separate the two stages.

2. Projection methods for feature extracion

In projection methods for feature extraction the aim is to map a d-dimensional space to an m-dimensional space, m < d, such that the structure of the data is preserved. For the understanding of the multidimensional data structure, as well as for classification purposes, the projected space dimension, m, should be as low as possible without significant loss in the performance. Sometimes, features extracted for one of these objectives are not necessarily the preferable features for the other objective.

2.1 Linear methods

Among the unsupervized 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. This procedure finds the subspace in which the original sample vectors may be approximated with the least mean square error for a given dimensionality.

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

Eigenvectors of the covariance matrix of the mixture density are ordered according to the magnitudes of the respective eigenvalues. The fraction of variance accounted by the first m eigenvectors, F_m, is,

$$F_m = \left(\sum_{i=1}^{m} e_i\right) / \left(\sum_{i=1}^{d} e_i\right) \tag{1}$$

where e_i is the ith eigenvalue. The nth eigenfeature is,

$$E_n = \sum_{i=1}^{m} u_{ni}\, x_i \tag{2}$$

where u_{ni} is the ith component of the nth eigenvector and x_i is the ith component of the original feature.

The magnitudes of the weights, $|u_{ni}|$, of each feature i indicates its relative contribution to the nth eigenfeature. The average of $|u_{ni}|$ over all n, U_i, indicates the total contribution of the ith feature. The value of U_i is used as an elimination criterion for the rejection of some of the d original features. The lowest weighted features are eliminated first and then the PCA is implemented. If the retained features and their associated weights are denoted 1 to m', (2) becomes

$$E'_n = \sum_{i=1}^{m'} u_{ni}\, x_i . \tag{3}$$

This is the incomplete eigenfeature system.

2.2 A nonlinear method

Sammon [3] proposed a nonlinear projection technique that attempts to preserve all the interpattern distances. The parameter to minimize in Sammon's projection algorithm is the mapping error, also called Sammon's stress, defined as,

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d^*(i,j)} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{[d^*(i,j)-d(i,j)]^2}{d^*(i,j)} \tag{4}$$

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. 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. Sammon's algorithm has no generalization capability (it is not able to project new data after training).

Jain and Mao [2] suggested a feedforward neural network implementation of Sammon's algorithm. They derived weight updating rule for the multilayer feedforward network which minimizes Sammon's stress based on the gradient descent method. In their implementation the network is able to project new patterns after training.

3. The methodology

3.1 Data set

The data set contained amniotic fluid chromosomes which were acquired and segmented in a process described elsewhere [4]. Some of the experiments were held with 5 types of chromosomes (types "2", "4", "13", "19" and "x") and the other with only 3 types ("13", "19" and "x"). In the first case 84 samples were used in each class (80% of them for training) and in the second case 100 samples were used in each class (50% of them for training). The chromosomes in feature space were represented either by global features (length, area, perimeter and centromeric index (the ratio of the short arm length of the chromosome to the whole chromosome length)) or by the density profile (dp) features (integral intensities along and perpendicular to the medial axis of the chromosome) [4], [5].

3.2 The classifier

A two-layer feedforward neural network trained by the backpropagation learning algorithm was chosen for the classification. The number of input units was set by the projected space dimension and the number of output units was determined by the number of classes (5 or 3 classes). The number of hidden units of the network was set according to a mechanism described elsewhere [5].

4. Experiments and results

Figure 1 sketches the probability of correct classification of the test set and the probability of error of the 5 types of chromosomes versus the percentage of variance preserved by the PCA. The chromosomes were represented in the feature space by their dp features. In the same graph, the probability of correct test classification and the probability of error using the 64 dp features are plotted for comparison purposes. Only the first 4 large eigenfeatures are required to yield performance better than this of the dp features. Results are given for percentages of variance of 38% (1 eigenfeature) to 99% (24 eigenfeatures). Using more eigenfeatures will cause, eventually, a decline in the probability of correct classification toward the probability achieved by the original 64 dp features. This decline related to the decrease in the ratio of the number of vectors per class to the number of features (e.g., with nearly 100 vectors per class this ratio for the 4 largest eigenfeatures is 25 where for the 10 largest eigenfeatures this ratio is 10). For the 64 dp features the ratio is even lower, which explains the poor results using these features.

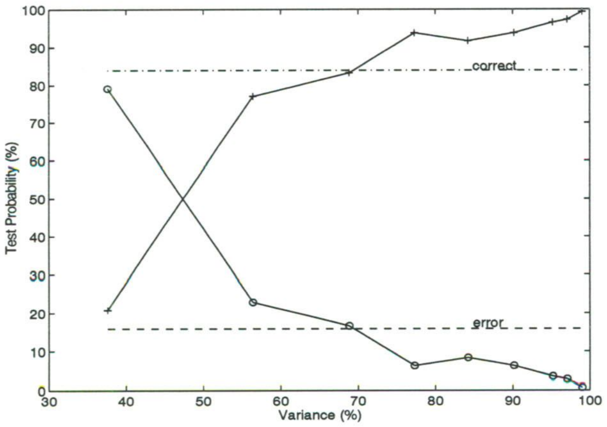

Fig. 1. The probability of correct classification of the test set (+) and the probability of error (o) vs. the percentage of variance preserved by the PCA. The same probabilities using the dp features marked with ·- and --, respectively.

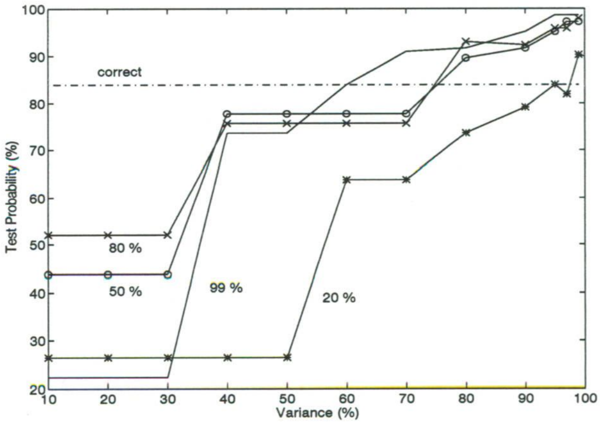

Fig. 2. The same as Fig. 1 for the incomplete eigenfeature approach for the "best" 20, 50, 80 and 99% of the original features.

Figure 2 outlines the results of an experiment based on the incomplete eigenfeature system. Four graphs are plotted, for 20, 50, 80 and 99% of the "best" original features selected according to the criterion described in section 2.1. These "best" features were projected by the PCA with preserved variances as plotted in the Figure (from 10 to 99%). This graph enables a designer a trade off between the number of pre-selected original features and the number of preserved projected features. As before, low ratio of the number of training vectors per class to the number of features dictates poor generalization performances when using all the 64 dp features compare to the case with only few eigenfeatures.

Figure 3 and 4 present results of a comparison of the linear (PCA) and the nonlinear (Sammon's algorithm) techniques, for the dp and the global features, respectively. The experiments were held for 3 classes of chromosomes and for different values of variance preserved by the PCA. The dimension of the projected space using Sammon's algorithm was 2 and the metric was Euclidean. The implementation was by a multilayer perceptron similarly to [2]. In Figure 3 the probability of correct test classification of all the 64 dp features, as well as of arbitrary 2 dp features (the first 2) is given for a comparison. It is interesting to compare the probability of correct classification using 2 arbitrary features (59%) to that of Sammon's algorithm (85%) and to that of the 2 largest PCA eigenfeatures (92%). Similar comparison of Sammon's algorithm to the PCA results for the global features (Figure 4) indicates an opposite conclusion, where the 2 features of the projected Sammon's space outperform the 2 largest eigenfeatures of the PCA (93% versus 59%). The probability of correct classification using all the 4 global features was 96%.

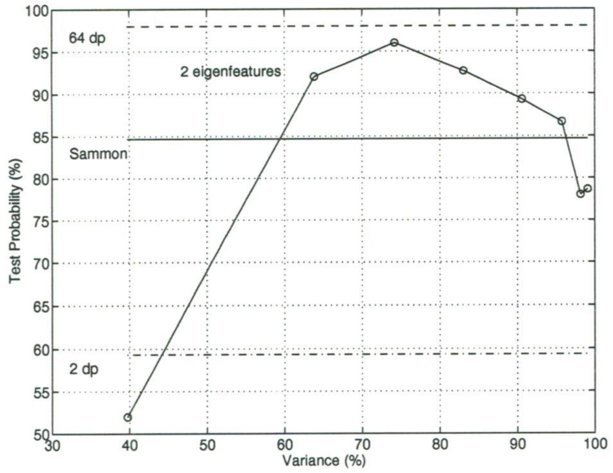

Fig. 3. A comparison between the PCA (o) and Sammon's algorithm (-) for the dp features and 3 classes of chromosomes. The probability of correct classification of the test set for the 64 and the 2 first dp features is also marked (-- and ·-, respectively).

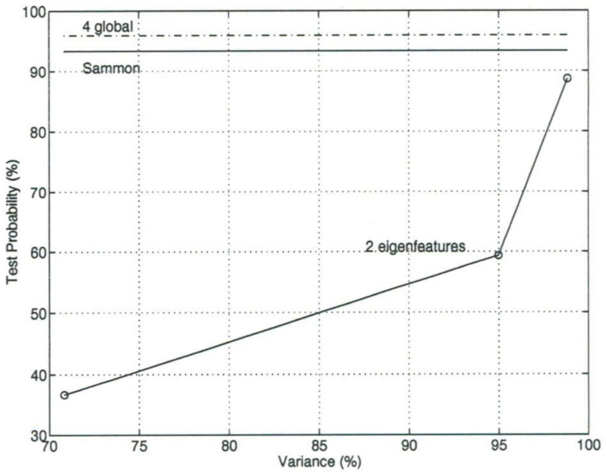

Fig. 4. A comparison between the PCA (o) and Sammon's algorithm (-) for the global features and 3 classes of chromosomes. The probability of correct classification of the test set for the global features is also marked (·-).

5. Discussion

The decision to either use a feature extraction method or not, as well as the constancy of the classification results are highly dependent on the ratio between training set size and the feature space dimension. High dimensional feature space and/or a small number of training vectors in each class could raise the problem of "curse of dimensionality" and to yield poor performance. This problem is especially encountered while trying to preserve large amount of information by a PCA when using small training sets.

The incomplete eigenfeature system enables feature reduction while a minor loss in performances with the benefit of using only small amount of measurements. The decision to preserve only few eigenfeatures by the PCA or to apply the incomplete eigenfeature system is a question of the classification problem and it is left to the designer. Likewise, the choice of which feature extraction method, linear or nonlinear, is preferable, is data dependent.

References

1. C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 388-400, 1993.
2. A. K. Jain and J. Mao, "Artificial neural network for nonlinear projection of multivariate data", *Proc. of the IJCNN'92*, Baltimore, June 1992, 335-340.
3. J. W. Sammon Jr., "A non-linear mapping for data structure analysis", *IEEE transactions on Computers*, 18, 401-409, 1969.
4. B. Lerner, H. Guterman, I. Dinstein and Y. Romem, "Medial axis transform based features and a neural network for human chromosome classification", Accepted for publication in *Pattern Recognition.*
5. B. Lerner, H. Guterman, I. Dinstein and Y. Romem, "Human chromosome classification using multilayer perceptron neural network", Accepted for publication in *International Journal of Neural Systems.*