Nonlinear Common Vectors for Pattern Classification

Hakan Cevikalp[1] and Marian Neamtu[2]
[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
[2] Department of Mathematics, Vanderbilt University, Nashville, Tennessee, USA
hakan.cevikalp@gmail.com, neamtu@math.vanderbilt.edu

Abstract

The Common Vector (CV) method is a linear method, which allows to discriminate between classes of data sets, such as those arising in image and word recognition. In this paper a variation of this method is introduced for finding the projection vectors of each class as elements of the intersection of the null space of that class' covariance matrix and the range space of the covariance matrix of the pooled data. Then, a novel approach is proposed to apply the method in a nonlinearly mapped higher-dimensional feature space. In this approach, all samples are mapped to a higher-dimensional feature space using a kernel mapping, and then the modified CV method is applied in the transformed space. As a result, each class gives rise to a unique common vector. This approach guarantees a 100% recognition rate for the samples of the training set. Moreover, experiments with several test cases also show that the generalization ability of the proposed method is superior to the kernel-based nonlinear subspace method.

1. Introduction

The subspace classifier is a pattern recognition method, which uses a linear subspace for each class [1]. The motivation behind the subspace classifiers is the optimal reconstruction of multi-dimensional data with linear principal components. In this approach, it is assumed that the vector distribution in each class lies in a lower-dimensional subspace of the original feature space. The subspaces representing classes are defined in terms of basis vectors that are linear combinations of the sample vectors of each class. Therefore, basis vectors spanning those subspaces must first be computed. Then, a test sample vector is classified based on the lengths of the projections of that sample onto each of the subspaces or, alternatively, on the distances of the test vector from these subspaces.

Watanabe et al. proposed the first subspace method, the Class-Featuring Information Compression (CLAFIC), for pattern classification [2]. This method employs the Principal Component Analysis (PCA) to compute the basis vectors spanning the subspaces of each class. However, the subspaces found by the CLAFIC method may sometimes have a large subspace in common, which causes poor classification of pattern samples. Therefore, the Method of Orthogonal Subspaces (MOSS) was proposed, which removes the common subspace of the classes and makes the subspaces mutually orthogonal [3]. Fukunaga and Koontz proposed a new method, which enabled to select the basis vectors in such a way that the projections onto the so-called rival subspaces are minimized [4]. Lastly, learning subspace methods, capable of learning in a decision-directed fashion, have been proposed in [5], [6].

Subspace classifiers are linear methods in nature and therefore may not extract nonlinear features of classes. The kernel-based nonlinear subspace method was developed to as to overcome this limitation [7], [8]. In this approach all data samples are mapped to a higher-dimensional feature space. The kernel trick is then used to compute the kernel PCA components in this transformed space for each class separately. It has been reported that the performance of nonlinear subspace methods is superior to linear ones [8].

All linear subspace methods discussed here are optimal from a reconstruction point of view and employ eigenvectors corresponding to nonzero eigenvalues of the correlation matrix of each class to compute the required basis vectors. The basis vectors are then used to construct the subspaces that represent classes in the database. However, these subspaces might not be optimal for discrimination of class samples. Therefore, Gulmezoglu et al. [9] proposed a new method for the case of a small sample size problem. For feature extraction, the method employs eigenvectors corresponding to zero eigenvalues of the covariance matrices of classes. It was proved that such basis vectors are optimal from the classification point of view and that all training set samples can be classified correctly when using these projection vectors for feature extraction.

In this paper we introduce a variation of the CV method and extend it to the nonlinear case. The new method, which will be referred to as the Kernel CV method, consists in applying the modified CV method in the setting of a nonlinearly mapped higher-dimensional feature space. The remainder of the paper is organized as follows: In Section 2, we introduce the modified CV method, Section 3 describes the Kernel CV method, and in Section 4, we discuss our experimental results.

2. A Variation of Common Vector Method

The CV method was originally proposed for isolated word recognition problems, where the number of samples in each class is smaller than or equal to the dimensionality of the sample space (i.e., the covariance matrix of each class is singular). This method extracts the features that are common

to all samples in each class. In order to accomplish its goal, the method eliminates all features that are in the direction of eigenvectors corresponding to nonzero eigenvalues of the covariance matrices of the classes. Therefore, each class is represented by the null space of its own class covariance matrix.

We showed recently that the null space of the covariance matrix (or scatter matrix) of the pooled data does not contain any discriminative information for classification of data samples [10], [11]. Therefore, this subspace can be discarded from our consideration. Then, the new subspace representing each class will be defined as the intersection of the null space of that class' covariance matrix and the range space of the covariance matrix of the pooled data.

In particular, let the training set be composed of C classes, where the i-th class contains N_i samples, and let x_m^i be a d-dimensional column vector, which denotes the m-th sample from the i-th class. There will be a total of $M = \sum_{i=1}^{C} N_i$ samples in the training set. Suppose that $d > N_i$, for $i = 1,...,C$. In this case, the covariance matrix of each class is defined as

$$\Sigma_i = \frac{1}{N_i} \sum_{m=1}^{N_i} (x_m^i - \mu_i)(x_m^i - \mu_i)^T, \quad i = 1,...,C, \tag{1}$$

where μ_i is the mean of the samples in the i-th class. The covariance matrix of the pooled data is defined as

$$\begin{aligned} \Sigma &= \frac{1}{M} \sum_{i=1}^{C} \sum_{m=1}^{N_i} (x_m^i - \mu)(x_m^i - \mu)^T \\ &= AA^T, \quad i = 1,...,C, m = 1,...,N_i, \end{aligned} \tag{2}$$

where μ is the mean of all samples and $A \in R^{dxM}$ is given by

$$A = [(x_1^i - \mu)/\sqrt{M} \ ... \ (x_{N_C}^C - \mu_i)/\sqrt{M}\ ]. \tag{3}$$

The projection matrix (also called the orthogonal projection operator) of the null space $N(\Sigma_i)$ of the covariance matrix of the i-th class and the projection matrix of the range space $R(\Sigma)$ of the covariance matrix of the pooled data, commute in the sense that

$$P^{(i)} P = P P^{(i)}, \tag{4}$$

where $P^{(i)}$ is the projection matrix of $N(\Sigma_i)$ and P is the projection matrix of $R(\Sigma)$. Therefore, the projection matrix P_{int}^i of the intersection $N(\Sigma_i) \cap R(\Sigma)$ for each class can be found as

$$P_{\text{int}}^i = P^{(i)} P = P P^{(i)}, \quad i = 1,...,C. \tag{5}$$

The basis vectors spanning each mentioned intersection space can be found by using an eigen-decomposition. More precisely, the eigenvectors corresponding to the eigenvalues 1 of P_{int}^i span the intersection subspaces representing the classes of interest. However, this approach is not always practical since the size of the projection matrices can be very large (e.g., images of size 256 by 256 yield projection matrices of size 65,536 by 65,536). On the other hand, since the projection matrices commute, we can first project the samples onto $R(\Sigma)$ and then find the null spaces of the classes in the transformed space, so as to compute basis vectors of the intersection subspaces. The algorithm carrying out this idea can be summarized as follows:

Step 1: Project the training set samples onto $R(\Sigma)$:

i) Compute the nonzero eigenvalues and corresponding eigenvectors α_k of Σ using the matrix $A^T A \in R^{MxM}$, where $\Sigma = AA^T \in R^{dxd}$ and A is given by (3) [10]. Set $U = [\alpha_1 \quad ... \quad \alpha_r]$, where r is the rank of Σ.

ii) Project the training set samples onto $R(\Sigma)$ by

$$y_m^i = U^T x_m^i, \quad i = 1,...,C, \quad m = 1,...,N_i. \tag{6}$$

Step 2: Find the null spaces of classes in the transformed space: In the transformed space, the new covariance matrices of the classes will be

$$\widetilde{\Sigma}_i = U^T \Sigma_i U, \quad i = 1,...,C. \tag{7}$$

Apply eigen-decomposition to each covariance matrix, $\widetilde{\Sigma}_i \in R^{rxr}$. Let $q_k^{(i)}$ be the eigenvectors corresponding to the nonzero eigenvalues of $\widetilde{\Sigma}_i$. Set $Q^{(i)} = [q_1^{(i)} \quad ... \quad q_{n_i}^{(i)}]$, where n_i is the dimensionality of $N(\widetilde{\Sigma}_i)$.

Step 3: Compute the final basis vectors of the intersection space $N(\Sigma_i) \cap R(\Sigma)$: The final basis vectors spanning the intersection subspaces will be

$$W^{(i)} = U Q^{(i)}, \quad i = 1,...,C. \tag{8}$$

Note that the basis vectors span the intersection subspace $N(\Sigma_i) \cap R(\Sigma)$ and therefore the following holds:

$$P_{\text{int}}^{(i)} = W^{(i)} W^{(i)^T}, \quad i = 1,...,C. \tag{9}$$

When the samples of each class are projected onto their corresponding intersection subspace, the feature vector $\Omega_{com}^i = [< x_m^i, w_1^i > \quad .... \quad < x_m^i, w_{n_i}^{(i)} >]^T$ of each sample is the same for all samples in that class. These feature vectors are called the common vectors. To recognize a test sample, we compute the Euclidean distances between the test sample feature vector and the common vectors of each class, using the Euclidean distance

$$\kappa_i = \| \Omega_{test}^i - \Omega_{com}^i \|, \ i = 1,...,C. \tag{10}$$

Then we assign the test sample to the class that minimizes this distance.

3. The Kernel Common Vector Method

This method consists in mapping the given training set samples to an implicit higher-dimensional space $\Im$ using a nonlinear kernel mapping and applying the above-described version of the linear CV method in the transformed space.

Let $\Phi(x_1^1),\Phi(x_2^1),...,\Phi(x_{N_1}^1),\Phi(x_1^2),...,\Phi(x_{N_C}^C)$ represent the transformed samples in $\Im$. The covariance matrix Σ^Φ of the pooled data in $\Im$ is given by

$$\begin{aligned}\Sigma^\Phi &= \frac{1}{M}\sum_{i=1}^{C}\sum_{m=1}^{N_i}(\Phi(x_m^i)-\mu^\Phi)(\Phi(x_m^i)-\mu^\Phi)^T \\ &= \frac{1}{M}(\Phi-\Phi 1_M)(\Phi-\Phi 1_M)^T,\end{aligned} \tag{11}$$

where μ^Φ is the mean of all samples, and Φ is the matrix whose columns are the mapped training set samples in $\Im$. Here $1_M \in R^{MxM}$ is a matrix with entries $1/M$.

Our aim is to find a basis vectors for the intersection subspaces $N(\Sigma_i^\Phi)\cap R(\Sigma^\Phi)$, for each class. Here, Σ_i^Φ represents the covariance matrix of the *i*-th class in $\Im$. To find these basis vectors, we follow the steps given in the previous section; we first project all training samples onto $R(\Sigma^\Phi)$ and then find the null spaces of the classes in the transformed space. The projection of training set samples onto $R(\Sigma^\Phi)$ can be done easily by employing the Kernel PCA method. The algorithm can be summarized as follows:

Step 1: Project the training set samples onto $R(\Sigma^\Phi)$ using the Kernel PCA. Let

$$\widetilde{K} = K - 1_M K - K 1_M + 1_M K 1_M \in R^{MxM} = P\Lambda P^T, \tag{12}$$

where the diagonal elements of Λ are nonzero and $K \in R^{MxM}$ is given by $K = \Phi^T\Phi = (K^{ij})_{\substack{i=1,...,C\\ j=1,...,C}}$,where the matrices $K^{ij} \in R^{N_i xN_j}$ are defined as

$$K^{ij} = (k_{mn}^{ij})_{\substack{m=1,...,N_i\\ n=1,...,N_j}} =< \Phi(x_m^i),\Phi(x_n^j) >= k(x_m^i,x_n^j)_{\substack{m=1,...,N_i\\ n=1,...,N_j}}. \tag{13}$$

The matrix that transforms the training set samples onto $R(\Sigma^\Phi)$ is $(\Phi-\Phi 1_M)P\Lambda^{-1/2}$. The new covariance matrix $\widetilde{\Sigma}_i^\Phi \in R^{rxr}$ (r is the rank of $R(\Sigma^\Phi)$ and cannot be larger than M-1) of each class in the reduced space becomes

$$\begin{aligned}\widetilde{\Sigma}_i^\Phi &= ((\Phi-\Phi 1_M)P\Lambda^{-1/2})^T\Sigma_i^\Phi(\Phi-\Phi 1_M)P\Lambda^{-1/2} \\ &= \frac{1}{N_i}\Lambda^{-1/2}P^T\widetilde{K}^{(i)}\widetilde{K}^{(i)T}P\Lambda^{-1/2}, \quad i=1,...,C.\end{aligned} \tag{14}$$

Here, the matrix $\widetilde{K}^{(i)} \in R^{MxN_i}$ is given by

$$\begin{aligned}\widetilde{K}^{(i)} &= K^{(i)} - K^{(i)}G^{(i)} - 1_M K^{(i)} + 1_M K^{(i)}G^{(i)} \\ &= (K^{(i)} - 1_M K^{(i)})(I - G^{(i)})\end{aligned}, \tag{15}$$

where $G^{(i)} \in R^{N_i xN_i}$ is a matrix whose elements are all $1/N_i$ and the matrix $K^{(i)} \in R^{MxN_i}$ is given by $K^{(i)} = \Phi^T\Phi^{(i)} = (K^{(i)j})_{j=1,...,C} \in R^{MxN_i}$, where $\Phi^{(i)}$ is the matrix whose columns are the mapped samples of the *i*-th class in $\Im$, and where each matrix $K^{(i)j} \in R^{N_j xN_i}$ is defined as

$$K^{(i)j} = (k_{mn}^{(i)j})_{\substack{m=1,...,N_j\\ n=1,...,N_i}} =< \Phi(x_m^j),\Phi(x_n^i) >= k(x_m^j,x_n^i)_{\substack{m=1,...,N_j\\ n=1,...,N_i}}. \tag{16}$$

Step 2: For each class, find a basis of the null space of $\widetilde{\Sigma}_i^\Phi$. This can be done by an eigen-decomposition. The normalized eigenvectors corresponding to the zero eigenvalues of $\widetilde{\Sigma}_i^\Phi$ form an orthonormal basis for the null space of $\widetilde{\Sigma}_i^\Phi$. Let $Q^{(i)}$ be a matrix whose columns are the computed eigenvectors corresponding to the zero eigenvalues, such that

$$Q^{(i)T}\widetilde{\Sigma}_i^\Phi Q^{(i)} = 0, \quad i=1,...,C. \tag{17}$$

Step 3: The basis vector matrix $W^{(i)}$ whose columns span the intersection subspace of the *i*-th class, is

$$W^{(i)} = (\Phi-\Phi 1_M)P\Lambda^{-1/2}Q^{(i)}, \quad i=1,...,C. \tag{18}$$

The number of basis vectors spanning the intersection subspaces is determined for each class by the dimensionality of $N(\widetilde{\Sigma}_i^\Phi)$. After performing feature extraction, all training set samples in each class generate the common vector of that class. Therefore, similarly to the linear CV case, a 100% recognition accuracy is also guaranteed for this method. Moreover, to recognize a given test sample, we compare the Euclidean distances between the common vectors and the feature vector of the test sample for each class, using (10), and we assign the test sample to the class that minimizes the distance.

4. Experimental Results

In our experiments we used the ORL (Olivetti-Oracle Research Lab) face database [12] to test the proposed method. The ORL face database contains *C*=40 individuals, with 10 images per person. The images are taken at different time instances with different lighting conditions (slightly), facial expressions, and facial details. The size of each image is 92x112. Some individuals from the ORL face database are shown in Fig. 1.

We have experimented with the polynomial kernel $k(x,y) = (< x,y >)^2$ of degree 2 and the Gaussian kernel $k(x,y) = \exp(-\| x-y \|^2/\gamma)$, for all data sets. The parameter γ was chosen as 1.06e8, based on empirical observations. Beside the methods proposed here, we also tested the linear CLAFIC method and the kernel-based nonlinear subspace method (Kernel CLAFIC). Class correlation matrices were used for finding the basis vectors spanning the subspaces of classes for the CLAFIC and the kernel-based nonlinear

subspace method. For both the CLAFIC and the kernel-based subspace methods the dimension of each subspace was determined by the rank of the corresponding correlation matrix since there are only a few training samples in each class. In particular, the dimension of each subspace was 5.

We selected randomly five samples from each class for training and the remaining samples were used for testing. We did not apply any preprocessing to the images. Then, recognition rates were computed and this process was repeated five times. The recognition rates were found by averaging the recognition rates in each run. The computed recognition rates are shown in Table I.

Fig.1. Three individuals from the ORL face database.

TABLE I
Recognition Rates of the ORL Face Database

Linear Methods	**Recognition Rates(%) & Standard Deviations**	
CLAFIC	95.3, $\sigma = 1.68$	
Variation of CV	96, $\sigma = 1.58$	
Nonlinear Methods	Polynomial Kernel	Gaussian Kernel
Kernel CLAFIC	95.3, $\sigma = 1.85$	95.9, $\sigma = 1.67$
Kernel CV	96, $\sigma = 1.83$	95.8, $\sigma = 1.68$

As can be seen from the results, although there is not a significant difference between the results, the modified CV method outperforms the CLAFIC method and similarly, the Kernel CV method outperforms the Kernel CLAFIC method. These results show that the basis vectors, which span the intersection of the null space of a class' covariance matrix and the range space of the covariance matrix of pooled data give an optimal set of projection directions for feature extraction.

5. Conclusion

In this paper we proposed a new method, which uses kernel functions for recognition. The method employs the intersection subspace of the null space of a class' covariance matrix and the range space of the covariance matrix of pooled data, to represent each class. When the training set samples are projected onto these intersection subspaces, all training set samples in each class give rise to a unique vector, called a common vector. Thus, a 100% recognition rate is guaranteed for the training set samples. Our test results show that the generalization ability of the proposed method compares favorably with the kernel-based nonlinear subspace method.

6. References

[1] E. Oja, *Subspace Methods of Pattern Recognition.* Research Studies Press, 1983.

[2] S. Watanabe, P.F. Lambert, C.A. Kulikowski, J.L. Buxton, and R. Walker, "Evaluation and selection of variables in pattern recognition," in Computer and Information Sciences II, pp. 91, 1967.

[3] S. Watanabe and N. Pakvasa, "Subspace method in pattern recognition," in *Proceedings of the 1st International Conference on Pattern Recognition*, Washington, D.C. 1973.

[4] K. Fukunaga and W. L. Koontz, "Application of the Karhunen-Loeve expansion to feature selection and ordering," *IEEE Transactions on Computers C-19(4)*, pp. 311-318, 1970.

[5] T. Kohonen, G. Nemeth, K.J. Bry, M. Jalanko, and H. Riittinen, "Spectral classification of phonemes by learning subspaces," in *Proceedings of the 5th International Conference on Acoustics, Speech and Signal Processing*, Washington D.C., pp. 97-100, 1979.

[6] M. Kuusela and E. Oja, "The averaged learning subspace method for spectral pattern recognition," in *Proceedings of the 6th International Conference on Pattern Recognition,* Munchen, pp. 134-137, 1982.

[7] K. Tsuda, "Subspace classifier in the Hilbert space," *Pattern Recognition Letters*, vol. 20, pp. 513-519, February 1999.

[8] S.-W. Kim and B. J. Oommen, 'On utilizing search methods to select subspace dimensions for kernel-based nonlinear subspace classifiers," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 136-141, January 2005.

[9] M. B. Gulmezoglu, V. Dzhafarov, and A. Barkana, "The common vector approach and its relation to principal component analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, September 2001.

[10] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 4-13, January 2005.

[11] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Transaction on Neural Networks*, in review.

[12] The ORL Database of Faces, AT&T Laboratories Cambridge.Available: http://www.uk.research.att.com/facedata-base.html.