ORIENTED SOFT LOCALIZED SUBSPACE CLASSIFICATION

Thiagarajan Balachander, and Ravi Kothari

Artificial Neural Systems Laboratory
Department of Electrical & Computer Engineering and Computer Science
University of Cincinnati, Cincinnati, OH 45221-0030, USA
{tbalacha,rkothari}@ececs.uc.edu

ABSTRACT

Subspace methods of pattern recognition form an interesting and popular classification paradigm. The earliest subspace method of classification was the CLass Featuring Information Compression (CLAFIC) which associated with each class a linear subspace. Local subspace classification methodologies which have enhanced classification power by associating multiple linear subspaces with each class have also been investigated. In this paper, we introduce the Oriented Soft Regional Subspace Classifier (OS-RSC). The highlights of this classifier are (i) Class specific subspaces are formed to specifically maximize the average projection of one class while minimizing that of the rival class (ii) Multiple manifolds are formed for each class increasing classification power (iii) soft sharing of the training patterns again allows for consistent classification performance. It turns out that the cost function for forming class specific subspaces is maximized for a subspace of unit dimensionality. The performance of the proposed classifier is tested on real-world classification problems.

1. INTRODUCTION

Subspace methods of pattern recognition classify a pattern based on its distance from different vector subspaces, with the presumption that each class is linearly spanned by a unique set of basis vectors. Thus the subspace classifier design reduces to the determination of projection subspaces for each class. Different subspace classifiers vary in the method of determining the subspaces and the number of subspaces associated with each class. For example, in CLAFIC [1] the subspaces are formed from the Principal Components of training patterns in that class.

Local subspace methods of pattern recognition are based on a more general data model and allow patterns of the same class to be associated with more than one sub manifold. Each pattern is thus associated with a distinct sub manifold depending on its location in the feature space. A (global) subspace classifier draws at most quadratic decision boundaries and thus the local subspace classifier forms piecewise quadratic decision boundaries and has enhanced classification ability. The Regional Subspace Classifier (RSC) [2] (see also [3]) and its soft version the S-RSC [4] were proposed by us earlier based on local sub manifolding of subspace classifiers. The S-RSC also included a mechanism for soft sharing of the training patterns between multiple sub manifolds. An approach to introduce locality by combining Nearest Neighbor technique with subspace methodology has also been proposed [5]. In an effort to improve CLAFIC the mean square (representational) error criterion can be replaced by error criterion which directly aid in classification. Learning subspace methods that directly reduce the number of misclassifications on the training set have also been proposed ([1]). The Adaptive Subspace SOM [6], performs unsupervised subspace classification, by allowing multiple units (clusters) to tune to input features and partition the input space.

In this paper, we propose a new subspace classification paradigm called the OS-RSC. In this, each class subspace is formed by an Oriented Principal Component Analysis (OPCA), such that the ratio of the average projection of patterns from own to rival class is maximized. The proposed OS-RSC thus generates subspaces that directly aid classification (as opposed to PCA based methods which use faithful representation as an indirect approach for classification). Also training patterns are allowed to be shared softly between multiple sub manifolds. Further, training patterns that show greater memberships to a sub manifold are allowed to influence the projection matrices for that sub manifold to a larger extent.

The rest of the paper has been organized as follows. In Section 2 we review some background material on CLAFIC and OPCA. We show that the OPCA based cost function is maximized (in general) for a subspace of dimensionality one. Section 3 outlines how we can modify the CLAFIC to bring about the improvements we suggest. An algorithm for the OS-RSC design is developed in Section 4. In Section 5 we present simulation results and we conclude the paper with a discussion in Section 6.

2. BACKGROUND

2.1. CLAFIC

Let the feature vectors be represented by $x \in R^n$ or $x = [x_1\ x_2\ \ldots\ x_n]^T$ and come from K classes $\omega^{(1)}, \ldots, \omega^{(K)}$. Each class $\omega^{(i)}$ is represented by a $p^{(i)}$ dimensional subspace $\mathcal{L}^{(i)}$ and the goal in CLAFIC is to maximize the average projection of the vectors of a given class $\omega^{(i)}$ on its

Part of this research was supported by a grant from The Whitaker Foundation.

own subspace $\mathcal{L}^{(i)}$. Thus we seek to[1]:

$$\mathrm{J}\uparrow = \sum_{i=1}^{K} E[x^T P^{(i)} x \mid x \in \omega^{(i)}] \quad (1)$$

by finding a set of orthonormal basis vectors $\{u_i^{(1)}, \ldots, u_i^{(p^{(i)})}\}$ and thus a unique projection matrix for the subspace $\mathcal{L}^{(i)}$ is computed as $P^{(i)} = \sum_{j=1}^{p^{(i)}} u_i^{(j)} u_i^{(j)\,T}$. Then for any given input vector the classification rule is

$$\text{Classify } x \in \omega^{(i)} \quad \textbf{if} \quad x^T P^{(i)} x > x^T P^{(j)} x \quad \forall\, j \neq i,$$

The basis vectors for the i^{th} class subspace of CLAFIC can be shown to be (see for *e.g.* [1]) the eigenvectors corresponding to the $p^{(i)}$ largest eigenvalues of the *class correlation* matrix given by

$$Q^{(i)} = E[xx^T | x \in \omega^{(i)}] \quad (2)$$

2.2. ORIENTED PCA

Let $\{x\}$ and $\{v\}$ both be stationary stochastic vector processes in R^n (x and v could represent signal/noise or in a classification setting signals from two classes). In oriented principal component analysis [7] the goal is to find the subspace $\mathcal{L}$ with an associated matrix $P = UU^T$ ($U = [u_1 \ldots u_m]$) such that

$$\mathrm{J_{OPC}}\uparrow = \frac{E[x^T P x]}{E[v^T P v]} = \frac{\mathrm{tr}\{U^T R_x U\}}{\mathrm{tr}\{U^T R_v U\}} = \frac{\sum_{i=1}^m u_i^T R_x u_i}{\sum_{i=1}^m u_i^T R_v u_i} \quad (3)$$

where $R_x = E[xx^T]$ and $R_v = E[vv^T]$. The solutions to the above equation are called the oriented principal components. The directions u_i that achieve the above maxima are obtained by differentiating (3) wrt u_i and solving by setting to zero. Along u_i's, x has maximum variance subject to the fact that v has minimum variance (i.e. maximum signal-to-signal ratio). The u_i's are the principal generalized eigenvectors (g.e.vector) of the symmetric generalized eigenvalue (g.e.value) problem

$$R_x u_i = \lambda_i R_v u_i \quad (4)$$

If R_v is invertible then the above can be rewritten as a normal eigenvalue problem $R_v^{-1} R_x u_i = \lambda_i u_i$. Thus the direction u_i is steered by the distribution of v. The directions u_i corresponding to g.e.values λ_i and are the principal g.e.vectors and each oriented eigenvector is an extremal direction subject to the constraint that it is R_x and R_v orthogonal to all the previous g.e.vectors i.e. $R_x -$ orthogonality : $u_i^T R_x u_j = 0 \;\forall\; j < i$ and $R_v -$orthogonality : $u_i^T R_v u_j = 0 \quad \forall \quad j < i$. To show (see also [8]) that the maximum value of J_{OPC} is λ_1 (largest g.e.value), substituting (4) into (3) we have, $\mathrm{J_{OPC}} = \sum_{i=1}^m \lambda_i \beta_i$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ and $\beta_i = u_i^T R_v u_i / \sum_{i=1}^m u_i^T R_v u_i$, which restricts $\sum_{i=1}^m \beta_i = 1$. Thus obviously the maximum $\mathrm{J_{OPC}}$ of λ_1 is obtained (for unique g.e.values) setting $\beta_1 = 1$ and the other β_i's to zeros. In other words, the maximizing subspace for J_{OPC} is of unit dimensionality (m = 1). If g.e.values are degenerate (i.e. repeated eigenvalues) then $m > 1$ can be used but maximal J_{OPC} is still λ_1 and no additional advantage is derived.

[1] $J\uparrow$ ($J\downarrow$) would imply we seek to maximize (minimize) the cost function J

3. ORIENTED SOFT REGIONAL SUBSPACE CLASSIFIERS

In CLAFIC the projection directions for each class are formed from PC's of the training patterns from a specific class. However, common sense dictates that that it might be more beneficial to take into account patterns from other classes also while finding the projection matrix for a given class. Making use of the OPCA discussed in the previous section, the Oriented Principal Component based Soft Regional Subspace Classifier (OS-RSC) is developed.

The overall cost function for the oriented subspace classification can be given for the two output class case as,

$$\mathrm{J}\uparrow = \sum_{c=1}^{L} \sum_{i=1, j\neq i}^{2} \frac{E[\alpha_c(x)(x-r_c)^T P_c^{(i)} (x-r_c) \mid x \in \omega^{(i)}]}{E[\alpha_c(x)(x-r_c)^T P_c^{(i)} (x-r_c) \mid x \in \omega^{(j)}]} \quad (5)$$

Thus in the above cost function the objective is to maximize the ratio of the average projection of a class of vectors on its own class projection matrix to the expected projection of that of the other class. Allowing patterns to be shared in a soft-fashion between the L clusters through $\alpha_c(x)$, the above equation can now be rewritten as,

$$\mathrm{J}\uparrow = \sum_{c=1}^{L} \sum_{i=1, i\neq j}^{2} \frac{\mathrm{tr}\{U_c^{(i)T} Q_{x-r_c}^{(i)} U_c^{(i)}\}}{\mathrm{tr}\{U_c^{(i)T} Q_{x-r_c}^{(j)} U_c^{(i)}\}} \quad (6)$$

where $Q_{x-r_c}^{(i)} = E[\alpha_c(x)(x-r_c)^T (x-r_c) \mid x \in \omega^{(i)}]$ and the basis vectors each class-cluster are given by an OPCA,

$$Q_{x-r_c}^{(i)} u_c{}^{(i)} = \lambda Q_{x-r_c}^{(j)} u_c{}^{(i)} \quad (7)$$

$$\text{Classification rule is } x \in \omega^{(i)} \quad \textbf{if } (x-r_c)^T P_c^{(i)} (x-r_c) > (x-r_{c'})^T P_{c'}^{(j)} (x-r_{c'}) \quad \forall\, j \neq i, \quad c, c' \in \{1, 2, \ldots L\}$$

4. ALGORITHM FOR OS-RSC DESIGN

Let $X = \{(x^1, y^1), (x^2, y^2), \ldots, (x^N, y^N)\}$ represent the N training patterns and their associated class labels, where $x^j \in R^n$ and $y^j \in \{1, 2\}$. Then it is required to come up with the following (i) the cluster associations of each of the training vectors with the L clusters, (ii) the cluster centers r_c and (iii) the own space projection matrix $P_c^{(i)}$ corresponding to all classes and clusters such that J is maximized. Adapting the generalized Lloyd's Algorithm [9] for VQ to OS-RSC (*i.e.* to iteratively refine α, $P_c^{(i)}$ and r_c to maximize J) the following algorithm is obtained,

1. Initialize the L cluster centers r_c to L randomly chosen patterns from the training set and initialize the cluster associations by

$$\alpha_i(x) = \frac{e^{\gamma d(x, r_i)}}{\sum_{c=1}^{L} e^{\gamma d(x, r_c)}} \quad (8)$$

where $d(x, r_c) = (x-r_c)^T P_c^{(i)} (x-r_c)$ for $x \in \omega^{(i)}$, and γ is a constant (which controls the membership degree). However since the $P_c^{(i)}$ matrices are not initialized just the Euclidean distance of each training

pattern from the cluster centers could be used in (8) for initialization. Use the soft memberships to initialize the projection matrices.

2. The current cluster memberships are available and each of the cluster centers r_c are to be updated. The total error corresponding to each cluster $C^{(c)}$ can be expressed as only the inner summation of (5) and that is to be maximized by proper update of r_c's. Finding the partial wrt r_c and solving,

$$r_c = \{(D_1\alpha_1 - N_1\alpha_2)D_2^2{P_c}^{(1)} + (D_2\alpha_2 - N_2\alpha_1) D_1^2{P_c}^{(2)}\}^{-1}\{(D_1D_2^2{P_c}^{(1)} - N_2D_1^2{P_c}^{(2)})(\alpha x)_1 + (D_1^2D_2{P_c}^{(2)} - N_1D_2^2{P_c}^{(1)})(\alpha x)_2\}$$

where

$$N_i = E[\alpha_c(x)(x-r_c)^T{P_c}^{(i)}(x-r_c) \mid x \in \omega^{(i)}]$$
$$D_i = E[\alpha_c(x)(x-r_c)^T{P_c}^{(i)}(x-r_c) \mid x \notin \omega^{(i)}]$$

and

$$(\alpha x)_i = E[\alpha_c(x)x \mid x \in \omega^{(i)}], \alpha_i = E[\alpha_c(x) \mid x \in \omega^{(i)}]$$

and $[\cdot]^-$ is the generalized inverse and is equal to the normal inverse if the latter exists.

3. The *class cluster covariance matrix* $Q_c^{(i)}$ is calculated and projection matrices $P_c^{(i)}$ are updated
4. The cluster associations α_c of individual training patterns are again updated using (8)
5. Steps 2-4 are iterated until the cluster centers stabilize or until a maximum number of iterations.

5. SIMULATIONS

The efficacy of the OS-RSC was evaluated on following real world data sets. Each dataset was divided into three parts - training, testing and validation. For each classification problem, 10 independent runs were simulated based on different initializations of the cluster centers, for varying number of clusters (L). The L giving the best performance on the validation dataset was retained and was then used with the testing data (over 10 runs) to provide the average accuracies.

Sonar Data: This is a two class identification problem of undersea targets (rock or cylinder) [10]. The inputs are in R^{60}. There are 111 cylinder patterns and 97 rock patterns. 104 training points, and 52 patterns each for validation and testing. The Sonar data was projected into R^{10} keeping the 10 most significant directions after PCA. This was necessary as the performance was unsatisfactory in R^{60}.

Lymphoma and Lymphoma1 Data: Lymphoma (Lymphoma1) is a two class - malignant or benign - identification problem. The inputs are in R^9 (R^{10}) and represent textural, tonal and boundary features extracted from segmented cytological preparations of lymph node cells [11]. There are a total of 439 patterns, 145 benign cases and 234 malignant cases. One-fifth of each class was divided equally for validation and testing.

Diabetes Data: This is a two class (Diabetes present or absent) identification problem of diabetes in Pima Indians [12]. The inputs are in R^8 and are personal data and result of medical examinations. 500 patterns are no diabetes class and 268 patterns are for diabetes class, making it a total of 768 patterns. 384 training samples, and 192 each for validation and testing.

Echocardiogram Data: This is a two class dataset (Patient survives for more than one year or not). There are a total of 132 patterns and all 11 inputs are numeric-valued. Of the eleven, two inputs are binary valued [13]. 108 patterns are 'dead' class 24 are 'alive' class. 66 training samples, and 33 each for validation and testing

6. DISCUSSION AND CONCLUSION

Table 1 summarizes the performance of OS-RSC on the datasets considered. Table 2 puts the performance of OS-RSC vis-a-vis best case performance reported elsewhere in literature in perspective. Figure 1 gives the plot of average accuracies and standard deviations over varying number of clusters.

Since an iterative algorithm was employed for the minimization of the cost function a convergence criterion could be specified. In all of our simulations we found it convenient to run the iteration for a fixed number of times (30). The changes in the cluster centers usually stabilized by then. The standard deviations in the accuracies obtained are high (see Table 1). This was probably due to the fact that the GLA that was applied for the minimization procedure did not converge to the global minima. Future effort will be directed towards applying global optimization (GA, SA) strategies.

The softness in the OS-RSC is distinct from soft competitive Vector Quantization algorithms in which memberships start out to be soft and then are annealed to become crisp. In OS-RSC the memberships determine how much each of the training patterns influence the projection subspaces and even the final memberships could remain soft. Also, although RSC, S-RSC and OS-RSC may be looked upon as creating a set of invariant feature banks they differ from ASSOM in that no specific attempt is made to introduce spatial ordering in the evolution of the clusters.

By virtue of forming subspaces that explicitly aid classification OS-RSC outperforms RSC, S-RSC and EALSM (which are similar in spirit to the OS-RSC as they involve variations of local subspace classification).

It is in general difficult if not impossible to build a classifier that is superior to all possible classification paradigms on all possible datasets. From Table 2 it can be observed that although the proposed methodology performs well across a wide range of datasets there do exist other classifiers that perform better. However,

(i) Across 5 different datasets the OS-RSC gives reasonably competitive performance
(ii) Being a subspace classifier it is simple and fast in operation
(iii) It has less tunable parameters and the dimensionality of the best subspace is 1

In fact other than determining the number of clusters, the only other 'tunable' parameter is γ (the softness factor).

It was found that the results were not overly sensitive to a range of γ (chosen so that softmax function does not always saturate). Having said that, it is felt that there is a need and scope for improving the performance of the proposed classifier. Inclusion of higher order combinations of the inputs is to be pursued as a means to that end. A preliminary investigation along these lines suggests strongly that even for a simple linear classification strategy inclusion of higher order combinations improve classification performance. Ongoing work is also focused on developing a bias-variance framework for the general case of subspace classifiers.

Data Set	% Accuracy	σ	# Clusters
SONAR	84.2	2.8	2
LYMPHOMA	74.0	4.8	8
LYMPHOMA1	77.3	2.2	5
DIABETES1	73.7	1.0	4
ECHO	87.3	5.4	2

Table 1: Results obtained with the OS-RSC.

Data Set	Classifier	% Accuracy	Ref
SONAR	EALSM	92.0‡	[3]
	MLP	90.4	[3]
	RSC	93.1	[2]
	S-RSC	92.3	[4]
	OS-RSC	84.2	†
LYMPHOMA	MLP	83.6	[11]
	RSC	66.5	[2]
	S-RSC	76.3	[4]
	OS-RSC	74.0	†
LYMPHOMA1	MLP	81.7	[11]
	RSC	69.8	†
	S-RSC	66.2	†
	OS-RSC	77.3	†
DIABETES1	CW	77.3	[14]
	k-NN	74.2	[14]
	MLP	76.4	[12]
	RSC	69.0	†
	S-RSC	72.8	†
	OS-RSC	73.7	†
ECHO	RSC	70.3	†
	S-RSC	82.4	†
	OS-RSC	87.3	†

Table 2: Comparative performance of different classifiers (‡BestCase; other numbers are average accuracies) for a *qualitative* assessment of the performance of different classifiers. In some cases the conditions under which the performance of other classifiers are obtained are not fully known, a direct comparison of the numbers is less meaningful. († These simulations were run by us for this paper and are not reported elsewhere)

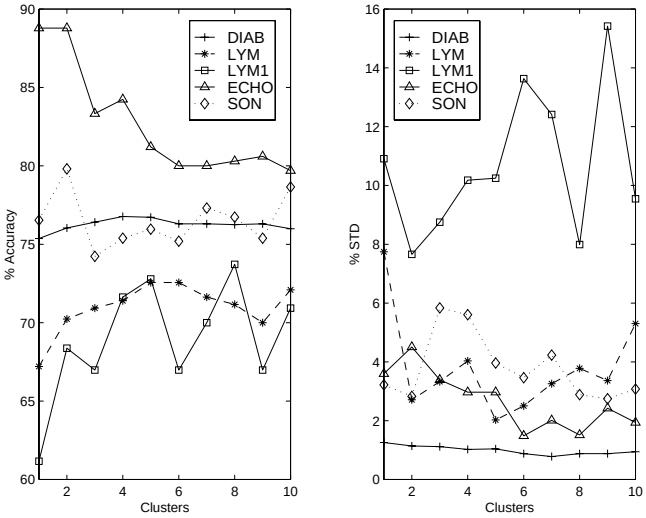

Figure 1: OS-RSC. Left: Generalization Accuracy in % , Right: Standard Deviation of Generalization Accuracy

7. REFERENCES

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

[2] T. Balachander and R. Kothari, "Localized Subspace Pattern Classification," *In Proc. IEEE IJCNN*, pp 1804-1809, 1998.

[3] M. Prakash and M. N. Murty, "Extended Subspace Methods of Pattern Recognition," *Pattern Recognition Letters*, Vol. 17, No. 11, pp. 1131-1139, 1996.

[4] T. Balachander, and R. Kothari, "Localized Soft Subspace Pattern Classification," *ICAPR*, 1998.

[5] J. Laaksonen, "Local Subspace Classifier and Local Subspace SOM," *Workshop On SOM*, Finland, 1997.

[6] T. Kohonen, S. Kaski and H. Lappalainen, "Self-Organized Formation of various Invariant-Feature Filters in the ASSOM," *Neural Computation*,Vol. 9, pp. 1321-1344, 1997.

[7] K. I. Diamantaras, and S. Y. Kung, *Principal Component Neural Networks*, John Wiley, 1997.

[8] Y. Yamashita and H. Ogawa, "Relative KL Transform," *IEEE Tran. Signal Processing*, Vol. 44, No. 2, pp. 371-378, 1996.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic, 1992.

[10] Data available from CMU-AI Repository. *ftp://ftp.cs.cmu.edu/afs/cs/project/connect/bench/*

[11] T. Balachander, R. Kothari, and H. Cualing, "An Empirical Comparison of Dimensionality Reduction Techniques for Pattern Classification," In LNCS, Vol. 1327, Springer-Verlag, pp. 589-594, 1997,

[12] L. Prechelt, Proben1 dataset available from *ftp://ftp.ira.uka.de/pub/neuron/proben1.tar.gz*

[13] Data at *http://pages.prodigy.com/upso/datasets.htm*

[14] C. Ji and S. Ma, "Combinations of Weak Classifiers," *IEEE Tran. Neural Networks*, Vol. 8, No. 1, pp. 32-42, 1997.