A NEW CLASS OF DECISION SURFACES BASED ON THE MINIMIZATION OF WITHIN CLASS VARIANCE

Stefanos Zafeiriou[†]*, Anastasios Tefas*[†] *and Ioannis Pitas*[†]

[†]Aristotle University of Thessaloniki
Department of Informatics, Box 451, 54124 Thessaloniki, Greece

ABSTRACT

In this paper a modified class of Support Vector Machine (SVM) inspired from the optimization of Fisher's discriminant ratio is presented. The modified class of SVM is used in order to find decision surfaces by solving the corresponding optimization problem in arbitrary Hilbert spaces, defined by Mercer's kernels. The effectiveness of the proposed approach is demonstrated by comparing it with the maximum margin SVM in various experiments using artificial data. Moreover, we have applied the proposed approach in the recognition of neutral expression in facial images.

Index Terms— Support Vector Machines, linear discriminant analysis, facial expression recognition, kernel machines.

1. INTRODUCTION

Pattern recognition systems employing SVM [1] have drawn much attention due to their good performance in practical applications and their solid theoretical foundations. The property that distinguishes SVM from other nonparametric techniques, like nearest-neighbor classification or neural networks, is that it is based on structural risk minimization. Typical pattern recognition methods attempt to minimize the misclassification errors on the training set (empirical risk minimization). Instead, SVM minimize the structural risk, that is the probability of misclassifying a previously unseen data point drawn randomly from a fixed but unknown probability distribution. If the Vapnik-Chervonenkis (VC)-dimension [1] of the family of decision surfaces is known, the theory of SVM provides an upper bound for the probability of misclassification of the test set for any possible probability distributions of the data points [1]. The main reason that made SVM so popular is that they consist of quadratic optimization problems which can be solved very efficiently and it is guaranteed that they will find a global extremum.

Another aspect of SVM is that they can be used in order to construct non-linear decision surfaces. In order to find such surfaces, a non-linear function ϕ is firstly used in order to project the samples to a very high dimensional feature space, the so-called *Hilbert space*, where the vectors are linearly or near-linearly separable and a maximum margin hyperplane is found. The decision surface can be found without having to compute explicitly the mapping ϕ, but by only computing dot products in the Hilbert space by means of the so-called *kernel trick* [2], as long as the mapping ϕ satisfies the Mercer's conditions [3]. The interested reader may refer to [4] for details on the geometry of Hilbert spaces (also referred as feature spaces).

In [5] a unified framework in terms of a nonlinearized variant of the Rayleigh coefficients has been proposed and has been applied in order to formulate nonlinear generalizations of Fishers Discriminant Analysis and oriented PCA with kernel functions. In order to overcome the fact that both calculation and eigenanalysis of covariance matrices in arbitrary dimensional Hilbert spaces are generally ill-posed problems, regularization parameters have been incorporated in the optimization problem.

An effort to combine the Fisher's discriminant optimization problem [6] and SVM formulation has been done in [7], where a modified class of SVMs has been constructed. In detail, motivated by the fact that the Fisher's discriminant optimization problem for two classes is a constraint least-squares optimization problem [7, 8, 9], the problem of minimizing the within-class variance has been reformulated, so that it can be solved by constructing the optimal separating hyperplane for both separable and nonseparable cases. In the face verification problem, the modified class of SVM has been applied successfully in order to weight the local similarity value of the elastic graphs nodes according to their corresponding discriminant power for frontal face verification [7]. It has been shown that it outperforms the maximum margin SVM [7].

In [7], only the linear SVM case with the assumption that the number of training vectors is larger than the feature dimensionality has been considered (i.e., when the within scatter matrix of the samples is not singular). In this paper the modified class of SVM is extended in arbitrary dimensional dot product Hilbert spaces, in order to define decision surfaces using Mercer's kernels. We will show, using both artificial and real data, that the proposed SVM method can outperform

This work has been funded by the project 03ED849 co-funded by the European Union and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework for Mr. Zafeiriou.

maximum margin SVM.

2. DECISION HYPERPLANES AND SURFACES

Let a training set with finite number of elements $\mathcal{U} = \{(\mathbf{x}_i, y_i), i \in \{1, \ldots, N\}\}$, be separated into two different classes $\mathcal{C}_k$ and $\mathcal{C}_t$, with training samples $\mathbf{x}_i \in \Re^M$ and labels $y_i \in \{1, -1\}$. The cardinalities of $\mathcal{C}_k$ and $\mathcal{C}_t$ are $N(\mathcal{C}_k)$ and $N(\mathcal{C}_t)$, respectively. The simplest way to separate these classes is by finding a separating hyperplane:

$$\mathbf{w}^T\mathbf{x} - b = 0 \tag{1}$$

where $\mathbf{w} \in \Re^M$ is the normal vector of the hyperplane and $b \in \Re$ is the corresponding scalar term of the hyperplane, also known as bias term [7]. The decision whether a test sample $\mathbf{x}$ belongs to one of the different classes $\mathcal{C}_k$ and $\mathcal{C}_t$ is taken by using the linear decision function $g_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} - b)$, also known as canonical decision hyperplane [1].

In cases that the samples are not linearly separable, we would like to find more complex decision functions. To do so, we use a non-linear mapping $\phi : \Re^M \rightarrow \mathcal{H}$ that maps the training samples to the arbitrary dimensional feature space $\mathcal{H}$. In that space the training samples are linearly or near-linearly separable. Hence, a hyperplane can be found in $\mathcal{H}$ (a hyperplane in $\mathcal{H}$ corresponds to a surface in $\Re^M$) as:

$$\mathbf{w}^T\phi(\mathbf{x}) - b = 0 \tag{2}$$

and the corresponding canonical decision function is $g_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\phi(\mathbf{x}) - b)$. In this paper we will define novel decision surfaces, using mappings ϕ that satisfy the Mercer's condition [1]. It is worth noting here that the hyperplanes in $\Re^M$ are a special case of hyperplanes in $\mathcal{H}$, when using the mapping $\phi(\mathbf{x}) = \mathbf{x}$.

3. THE NOVEL CLASS OF DECISION SURFACES

In [7], inspired by the maximization of the Fisher's discriminant ratio, which requires the minimization of the within class variance, and the SVM separability constraints, a modified class of SVM has been introduced. The optimization problem is defined as:

$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{S}_w\mathbf{w}, \;\; \mathbf{w}^T\mathbf{S}_w\mathbf{w} > 0 \tag{3}$$

subject to the separability constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1, \;\; i = 1, \ldots, N \tag{4}$$

where the matrix $\mathbf{S}_w$ is the within class scatter matrix defined as:

$$\begin{aligned} \mathbf{S}_w = \; & \textstyle\sum_{\mathbf{x}\in\mathcal{C}_t}(\mathbf{x} - \mathbf{m}_{\mathcal{C}_t})(\mathbf{x} - \mathbf{m}_{\mathcal{C}_t})^T + \\ & + \textstyle\sum_{\mathbf{x}\in\mathcal{C}_k}(\mathbf{x} - \mathbf{m}_{\mathcal{C}_k})(\mathbf{x} - \mathbf{m}_{\mathcal{C}_k})^T, \end{aligned} \tag{5}$$

$\mathbf{m}_{\mathcal{C}_k}$ and $\mathbf{m}_{\mathcal{C}_t}$ are the mean sample vectors for the classes $\mathcal{C}_k$ and $\mathcal{C}_t$, respectively.

In [7] only the linear case has been considered (the decision surfaces proposed in [7] are not the the generalization of (3) subject to the constraints (4) in Hilbert spaces). In the following, the optimization problem that has as outcome the proposed decision surfaces will be defined and solved. These decision surfaces are derived from the minimization of the within class variance in a dot product Hilbert space $\mathcal{H}$ subject to separability constraints. The space $\mathcal{H}$ will be called feature space while the original $\Re^M$ space will be called input space. In the space $\mathcal{H}$ the within scatter is defined as:

$$\begin{aligned} \mathbf{S}_w^\Phi & = \textstyle\sum_{\mathbf{x}\in\mathcal{C}_t}(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_t}^\Phi)(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_t}^\Phi)^T \\ & + \textstyle\sum_{\mathbf{x}\in\mathcal{C}_k}(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_k}^\Phi)(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_k}^\Phi)^T. \end{aligned} \tag{6}$$

where $\mathbf{m}_{\mathcal{C}_k}^\Phi = \frac{1}{N(\mathcal{C}_k)}\sum_{\mathbf{x}\in\mathcal{C}_k}\phi(\mathbf{x})$ and $\mathbf{m}_{\mathcal{C}_t}^\Phi = \frac{1}{N(\mathcal{C}_t)}\sum_{\mathbf{x}\in\mathcal{C}_t}\phi(\mathbf{x})$ are the mean sample vectors in $\mathcal{H}$ for the classes $\mathcal{C}_k$ and $\mathcal{C}_t$, respectively. The optimization problem of the modified SVM approach (in *soft margin* formulation [10]) consists of finding a vector $\mathbf{w} \in \mathcal{H}$ such that:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathbf{w}^T\mathbf{S}_w^\Phi\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \mathbf{w}^T\mathbf{S}_w^\Phi\mathbf{w} > 0 \tag{7}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) - b) \geq 1 - \xi_i \; , \xi_i \geq 0, \;\; i = 1, \ldots, N. \tag{8}$$

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]$ is the vector of the non-negative slack variables and C is a given constant that defines the cost of the errors after the classification. The solution of the minimization of (7), subject to the constraints (8), is given by the saddle point of the Lagrangian:

$$\begin{aligned} L(\mathbf{w}, b, \mathbf{a}, \mathbf{r}, \boldsymbol{\xi}) & = \mathbf{w}^T\mathbf{S}_w^\Phi\mathbf{w} + C\textstyle\sum_{i=1}^N \xi_i - \\ & \textstyle\sum_{i=1}^N a_i[y_i(\mathbf{w}^T\phi(\mathbf{x}_i) - b) - 1 + \xi_i] \\ & - \textstyle\sum_{i=1}^N r_i\xi_i \end{aligned} \tag{9}$$

where $\mathbf{a} = [a_1, \ldots, a_N]^T$ and $\mathbf{r} = [r_1, \ldots, r_N]^T$ are the vectors of the Lagrangian multipliers for the constraints (8). The Karush-Kuhn-Tucker (KKT) conditions [11] imply that for the optimal choice of $\mathbf{w}, \mathbf{a}, \mathbf{r}, b, \boldsymbol{\xi}$ the following hold:

$$\begin{aligned} & \nabla_\mathbf{w} L|_{\mathbf{w}=\mathbf{w}_o} = \mathbf{0} \Leftrightarrow \mathbf{S}_w^\Phi\mathbf{w}_o = \tfrac{1}{2}\textstyle\sum_{i=1}^N a_{i,o}y_i\phi(\mathbf{x}_i) \\ & \tfrac{\partial L}{\partial b}|_{b=b_o} = 0 \Leftrightarrow \mathbf{a}_o^T\mathbf{y} = 0 \\ & \tfrac{\partial L}{\partial \xi_i}|_{\xi_i=\xi_{i,o}} = 0 \Leftrightarrow r_{i,o} = C - a_{i,o} \\ & r_{i,o} \geq 0, 0 \leq a_{i,o} \leq C, \xi_{i,o} \geq 0, r_{i,o}\xi_{i,o} = 0 \\ & y_i(\mathbf{w}_o^T\phi(\mathbf{x}_i) - b_o) - 1 + \xi_{i,o} \geq 0 \\ & a_{i,o}\{y_i(\mathbf{w}_o^T\phi(\mathbf{x}_i) - b_o) - 1 + \xi_{i,o}\} = 0 \end{aligned} \tag{10}$$

the subscript o denotes the optimal case and $\mathbf{y} = [y_1, \dots, y_N]$ is the vector denoting the class labels. Since the feature space is of arbitrary dimension the matrix $\mathbf{S}_w^\Phi$ is almost always singular. Thus, the optimal normal vector $\mathbf{w}_o$ cannot be directly found from the KKT conditions (10):

$$\mathbf{S}_w^\Phi \mathbf{w}_o = \frac{1}{2} \sum_{i=1}^{N} a_{i,o} y_i \phi(\mathbf{x}_i). \tag{11}$$

In this paper a solution of the optimization problem (7) subject to the separability constraints (8) will be described without having to assume that the within scatter matrix of the training data is invertible neither in the feature space nor in the input space. It will be proven that there is a solution to this optimization problem by proving that there is a mapping that makes the solution feasible. This mapping is the *Kernel Principal Component Analysis* (KPCA) transform [12].

Let us define the total scatter matrix $\mathbf{S}_t^\Phi$ in the feature space $\mathcal{H}$ as:

$$\mathbf{S}_t^\Phi = \sum_{i=1}^{N} (\phi(\mathbf{x}_i) - \mathbf{m}^\Phi)(\phi(\mathbf{x}_i) - \mathbf{m}^\Phi)^T, \tag{12}$$

where $\mathbf{m}^\Phi = \sum \phi \mathbf{x}$. The matrix $\mathbf{S}_t^\Phi$ is bounded, compact, positive and self-adjoint operator in the Hilbert space $\mathcal{H}$. Thus, according to the Hilbert-Schmidt Theorem [13, 14] its eigenvectors system is an orthonormal basis of $\mathcal{H}$. Let $\mathcal{B}^\Phi$ and $\mathcal{B}_\perp^\Phi$ be the complementary spaces spanned by the orthonormal eigenvectors of $\mathbf{S}_t^\Phi$ that correspond to non-zero eigenvalues and to zero eigenvalues, respectively. Thus, any arbitrary vector $\mathbf{w} \in \mathcal{H}$, can be uniquely represented as $\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta}$ with $\boldsymbol{\varphi} \in \mathcal{B}^\Phi$ and $\boldsymbol{\zeta} \in \mathcal{B}_\perp^\Phi$ [14].

Let us define the linear mapping $L^\Phi : \mathcal{H} \to \mathcal{B}^\Phi$ as:

$$\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta} \to \boldsymbol{\varphi}. \tag{13}$$

The following Theorem shows that the optimization of the (7) subject to the constraints (8) can be performed in the space $\mathcal{B}^\Phi$ instead of $\mathcal{H}$ without losing any information.

Theorem. Under the mapping L^Φ the optimization problems (7) subject to the constraints (8) is equivalent to:

$$\min_{\boldsymbol{\varphi}, b, \boldsymbol{\xi}} \boldsymbol{\varphi}^T \mathbf{S}_w^\Phi \boldsymbol{\varphi} + C \sum_{i=1}^{N} \xi_i, \ (\boldsymbol{\varphi}^T \mathbf{S}_w^\Phi \boldsymbol{\varphi} > 0), \ \boldsymbol{\varphi} \in \mathcal{B}^\Phi \tag{14}$$

subject to the constraints:

$$y_i(\boldsymbol{\varphi}^T \phi(\mathbf{x}_i) - b) \geq 1 - \xi_i \ , \xi_i \geq 0, \ i = 1, \dots, N, \quad \boldsymbol{\varphi} \in \mathcal{B}^\Phi \square \tag{15}$$

The proof of the above Theorem is omitted due to lack of space.

The optimal decision surface for the optimization problem (7) subject to the constraints (8) can be found in the reduced space $\mathcal{B}^\Phi$ spanned by the non-zero eigenvectors of $\mathbf{S}_t^\Phi$. The number of the non-zero eigenvectors of $\mathbf{S}_t^\Phi$ is $K \leq N - 1$ thus, the dimensionality of $\mathcal{B}^\Phi$ is $K \leq N - 1$ and according to the functional analysis theory [15] the space $\mathcal{B}^\Phi$ is isomorphic to the $(N-1)$-dimensional Euclidean space $\Re^{N-1}$. The isomorphic mapping is:

$$\boldsymbol{\varphi} = \mathbf{P} \boldsymbol{\eta}, \ \boldsymbol{\eta} \in \Re^{N-1}, \tag{16}$$

where $\mathbf{P}$ is the matrix with columns the eigenvectors of $\mathbf{S}_t^\Phi$ that correspond to non-null eigenvalues and is an one-to-one mapping from $\Re^{N-1}$ onto $\mathcal{B}$.

Under this mapping the optimization problem is reformulated as:

$$\min_{\boldsymbol{\eta}, b, \boldsymbol{\xi}} \boldsymbol{\eta}^T \tilde{\mathbf{S}}_w \boldsymbol{\eta} + C \sum_{i=1}^{N} \xi_i, \ \boldsymbol{\eta}^T \acute{\mathbf{S}}_w \boldsymbol{\eta} > 0, \ \boldsymbol{\eta} \in \Re^{N-1} \tag{17}$$

where $\tilde{\mathbf{S}}_w$ is the within scatter matrix of the projected vectors in $\Re^{N-1}$ given by $\tilde{\mathbf{S}}_w = \mathbf{P}^T \mathbf{S}_w^\Phi \mathbf{P}$ (KPCA transform). The equivalent separability constraints are:

$$\begin{gathered} y_i(\boldsymbol{\eta}^T \tilde{\mathbf{x}}_i - b) \geq 1 - \xi_i \\ , \xi_i \geq 0, \ i = 1, \dots, N, \ \boldsymbol{\eta} \in \Re^{N-1} \end{gathered} \tag{18}$$

where $\tilde{\mathbf{x}}_i = \mathbf{P}^T \phi(\mathbf{x}_i)$ are the projected vectors in $\Re^{N-1}$ using the KPCA transform. For details on calculation of the projections using the KPCA transform someone may refer to [12, 14]. Under the projection to KPCA mapping, the optimal decision surface for the optimization problem (14) subject to (15) in $\mathcal{H}$ can be found by solving the optimization problem (17) subject to (18) in $\Re^{N-1}$ and then moving to $\mathcal{H}$ using (16).

The solution of the optimization problem (17) subject to the constraint (18) is found by the saddle point of the following Lagrangian:

$$\begin{aligned} L(\boldsymbol{\eta}, b, \mathbf{a}, \mathbf{r}, \boldsymbol{\xi}) \quad & = \boldsymbol{\eta}^T \tilde{\mathbf{S}}_w \boldsymbol{\eta} + C \textstyle\sum_{i=1}^{N} \xi_i - \\ & - \textstyle\sum_{i=1}^{N} a_i [y_i(\boldsymbol{\eta}^T \tilde{\mathbf{x}}_i - b) - 1 + \xi_i] - \\ & - \textstyle\sum_{i=1}^{N} r_i \xi_i \end{aligned} \tag{19}$$

the KKT conditions are similar to the ones in (10). Thus, the optimal vector is given by:

$$\tilde{\mathbf{S}}_w \boldsymbol{\eta}_o = \frac{1}{2} \sum_{i=1}^{N} a_{i,o} y_i \tilde{\mathbf{x}}_i. \tag{20}$$

The problem here is that the matrix $\tilde{\mathbf{S}}_w$ may be still singular. But, if the matrix $\tilde{\mathbf{S}}_w$ is singular it contains only one eigenvector that corresponds to null eigenvalue.

3.1. The matrix $\tilde{\mathbf{S}}_w$ is singular

In order to find the decision surface for the case of $\tilde{\mathbf{S}}_w$ being singular, we proceed as follows. Let $\tilde{\boldsymbol{\Theta}}$ be the matrix with columns the $N-2$ non-null eigenvectors of $\tilde{\mathbf{S}}_w$. In order to find the decision surface in this case the training samples are projected to $\tilde{\mathbf{\Theta}}$. Let $\breve{\mathbf{S}}_w$ be the within class scatter matrix of the projected samples in $\Re^{N-2}$ given by $\breve{\mathbf{S}}_w = \tilde{\mathbf{\Theta}}^T \tilde{\mathbf{S}}_w \tilde{\mathbf{\Theta}}$. The optimal normal vector γ_o in $\Re^{N-2}$ of the hyperplane is given by:

$$\breve{\mathbf{S}}_w \gamma_o = \frac{1}{2}\sum_{i=1}^{N} a_{i,o} y_i \tilde{\breve{x}}_i \Leftrightarrow \gamma_o = \frac{1}{2}\breve{\mathbf{S}}_w^{-1}\sum_{i=1}^{N} a_{i,o} y_i \breve{\mathbf{x}}_i. \quad (21)$$

By replacing (21) to (19) and using the KKT conditions (10), the constraint optimization problem (17) subject to the constraints (18) is reformulated to the Wolfe dual problem:

$$\max_{\mathbf{a}} f(\mathbf{a}) = \mathbf{1}_N^T \mathbf{a} - \frac{1}{2}\mathbf{a}^T \mathbf{Q} \mathbf{a} \text{ subject to} \quad (22)$$
$$0 \le a_i \le C,\ i = 1, \dots, N,\ \ \mathbf{a}^T \mathbf{y} = 0$$

where $\mathbf{1}_N$ is a N dimensional vector of ones and $[\mathbf{Q}]_{i,j} = \frac{1}{2} y_i y_j \breve{\mathbf{x}}_i^T \breve{\mathbf{S}_w}^{-1} \breve{\mathbf{x}}_j$. The corresponding optimal normal vector that is derived from the optimization problem (17) subject to the constraints (18) is $\boldsymbol{\eta}_o = \tilde{\mathbf{\Theta}} \gamma_o$. The decision surface is given by:

$$\begin{aligned} g(\mathbf{x}) &= \text{sign}(\mathbf{w}_o^T \phi(\mathbf{x}) - b_o) = \text{sign}(\boldsymbol{\varphi}_o^T \phi(\mathbf{x}) - b_o) = \\ &= \text{sign}(\boldsymbol{\eta}_o^T \mathbf{P}^T \phi(\mathbf{x}) - b_o) \\ &= \text{sign}(\gamma_o^T \tilde{\mathbf{\Theta}}^T \mathbf{P}^T \phi(\mathbf{x}) - b_o) = \\ &= \text{sign}(\tfrac{1}{2}\textstyle\sum_{i=1}^{N} a_{i,o} y_i \phi(\mathbf{x}_i)^T \mathbf{P} \tilde{\mathbf{\Theta}} \breve{\mathbf{S}}_w^{-1} \tilde{\mathbf{\Theta}}^T \mathbf{P}^T \phi(\mathbf{x}) - \\ &\quad - b_o) \end{aligned} \quad (23)$$

The optimal threshold b_o can be found by exploiting the fact that for all support vectors $\tilde{\mathbf{x}}_i$ with $0 < a_{i,o} < C$, their corresponding slack variables are zero (KKT condition). Thus, for any support vector $\tilde{\mathbf{x}}_i$ with $i \in \mathcal{S} = \{i : 0 < a_i < C\}$ the following holds:

$$y_i\left(\frac{1}{2}\sum_{j=1}^{N} y_j a_{j,o} \breve{\mathbf{x}}_j^T \breve{\mathbf{S}}_w^{-1} \breve{\mathbf{x}}_i - b_o\right) = 1. \quad (24)$$

Averaging over these patterns yields a numerically stable solution:

$$b_o = \frac{1}{N}\sum_{i \in \mathcal{S}}\left(\frac{1}{2}\sum_{j=1}^{N} y_j a_{j,o} \breve{\mathbf{x}}_j^T \breve{\mathbf{S}}_w^{-1} \breve{\mathbf{x}}_i - y_i\right). \quad (25)$$

Summarizing, the training phase includes an initial projection to $\Re^{N-1}$ using the KPCA transform. The training samples are accordingly projected to $\Re^{N-2}$ using $\tilde{\mathbf{\Theta}}$. Then, the SVM optimization problem is solved in this space where $\breve{\mathbf{S}}_w$ is invertible. In the test phase, when a test vector arrives for classification, it should be first projected to $\Re^{N-2}$ by using the above procedure and finally classified using (23).

4. EXPERIMENTAL RESULTS

4.1. Experimental Results with Artificial Data

Several kernels have been used in the experiments and the parameter C in (7) has been set to infinity so that no training errors were allowed. The typical kernels that have been used in our experiments have been polynomial and Radial Basis Functions (RBF) kernels:

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d \quad (26)$$
$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = e^{-\gamma(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}$$

where d is the degree of the polynomial and γ is the spread of the Gaussian cluster.

Artificial data have been used in order to show that the proposed decision hypeplanes and surfaces are not so sensitive to outliers as the ones defined by the maximum margin SVM approach. A comparison of the linear maximum margin SVM against the linear modified SVM (case where $\phi(\mathbf{x}) = \mathbf{x}$) in the separable case is shown in Figure 1. The advantage of the modified SVM method is that it takes into account both the class distribution statistics and the vectors that are in the boundaries, in contrast to the maximum margin SVM that considers only the vectors that lie in the boundaries.

In case of a non-linear decision surface the suitability of the proposed approach against the maximum margin SVM can be seen in Figure 2. The SVM approach totally failed to capture the nonlinearity of the data (Figure 2a). On the other hand the proposed decision surface has successfully captured the underlying non-linearity of the data (Figure 2b).

4.2. Neutral Facial Expression Detection using Cohn- Kanade database

This experiment illustrates the application of the proposed technique to the neutral facial expression detection problem. The recognition of the neutral facial expression can be also used to assist face verification algorithms [16], that, in general, are sensitive to the change of facial expressions and ask the client to have a neutral facial expression when using the verification system.

The Cohn-Kanade database [17] was used for the facial expression recognition in 6 basic facial expressions (anger, disgust, fear, happiness, sadness and surprise) classes. This database, is anottated with Facial Action Units (FAUs) [18]. These combinations of FAUs were translated into facial expressions according to [18], in order to define the corresponding ground truth for the facial expressions. In order to form the dataset to be used for the experiments, every image sequence available was taken under consideration, for every subject (96 subjects in total). One image for the neutral state and one image for the fully intensed facial expression were chosen from each image sequence (first and last frame of the image sequence respectively). Not all six facial expressions

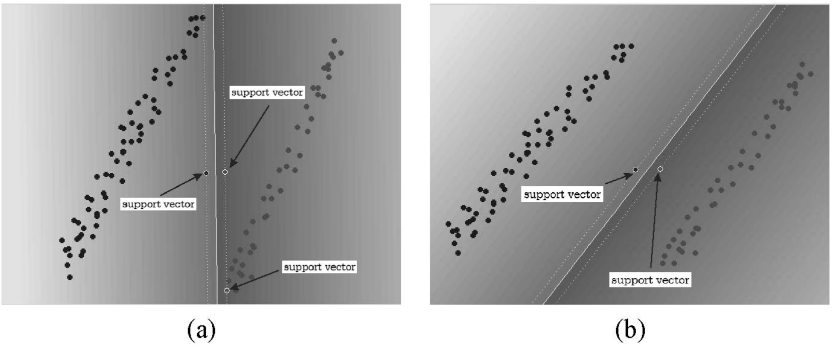

Fig. 1. a) The maximum margin SVM hyperplane; b) the proposed SVM hypeplane.

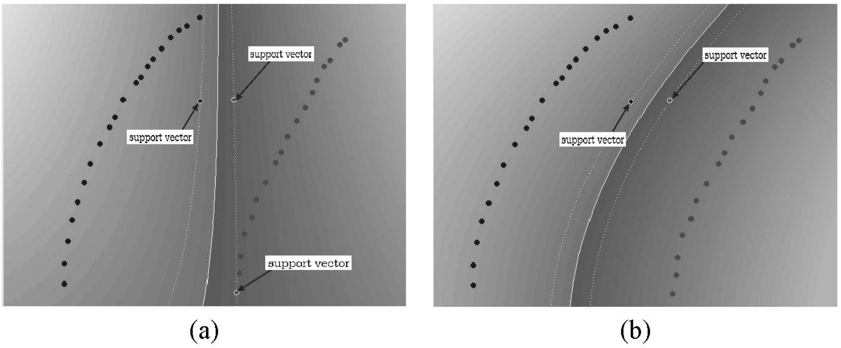

Fig. 2. The optimal decision surface using second order polynomial kernel and (a) maximum margin SVM, (b) the proposed SVM.

were present for every subject. For example a subject may have three video sequences posing happiness and none posing sadness, thus creating 3 samples for the happiness facial expression and 3 samples for the neutral facial expression, but none for the sadness facial expression. The chosen images were used to build the database, consisting of 704 images (equal number of samples for the neutral and fully expressive images). In Figure 3, a sample of image sequences of one poser from this database, is shown.

From the total of 704 ”face-prints” of the Cohn-Kanade database the 352 are neutral facial images while the remaining 352 are expressive images. The average size of the training set has been 564 facial images (282 expressive and 282 neutral images) and the average size of the test set has been 141 images (70.5 neutral and 70.5 expressive images).

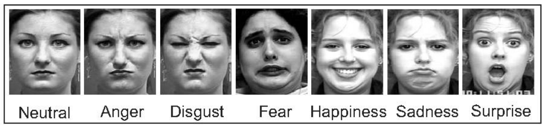

Fig. 3. Neutral Vs Expressive Images of a poser of Kanade database

Table 1. The best error rates of the tested classifiers for neutral state detection.

Algorithm	Overall Error Rate%
MVSVMs with 4-th degree polynomial kernel	**6**
SVMs with 4-th degree polynomial kernel	7.9
Regular CKFDA 4-th degree polynomial kernel	14

Figure 4 shows the results of the various tested approaches, i.e., a Kernel Fisher's Discriminant algorithm (the so-called Complete Kernel Fsher's Discriminant Analysis (CKFDA)[14]), SVMs, and the Modified Version of SVMs (MVSVMs), proposed in this paper, for the polynomial kernel and for various degrees. As can be seen MVSVMs approach is constantly better than SVMs and CKFDA for all the tested polynomial kernels. The lowest error rates are summarized in Table 1.

5. CONCLUSION

A novel class of decision surfaces inspired from the Fisher's discriminant ratio and SVM has been proposed. The advantage of the proposed technique is that they consider the class distribution statistics and not only the samples in the class

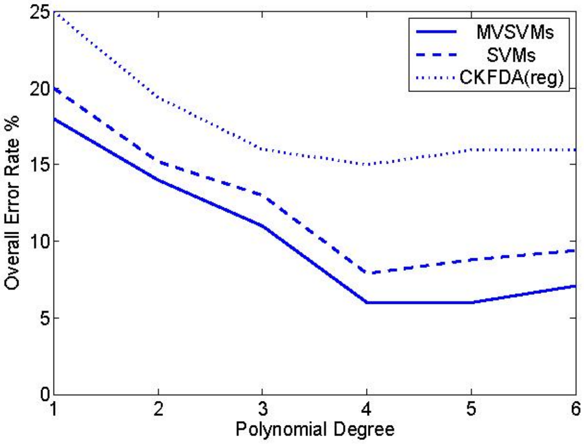

Fig. 4. Experimental results for neutral detection determination using polynomial kernel with various degrees.

boundaries. The demonstrated experiments have shown that the proposed class of decision surfaces can overcome some of the problems of the maximum margin SVM. Further research includes the theoretical investigation of the generalization capability of the proposed class of SVM.

6. REFERENCES

[1] Vapnik,V., *Statistical Learning Theory*, J.Wiley, New York, 1998.

[2] Scholkopf,B. and Smola,A., *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[3] Saitoh,S., *Theory of Reproducing Kernels and its Applications*, Harlow, UK: Longman Scientific & Technical, 1988.

[4] Scholkopf,B. and Mika,S. and Burges,C.J.C and Knirsch,P. and Muller,K.-R. and Ratsch,G and Smola,A.J., “Input space vs. feature space in kernel-based methods,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[5] Mika, S. and Ratsch G. and Weston, J. and Scholkopf, B. and Smola, A. and Muller, K.-R., “Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 623 – 628, 2003.

[6] Fukunaga,K., *Statistical Pattern Recognition*, CA: Academic, San Diego, 1990.

[7] Tefas,A. and Kotropoulos,C. and Pitas,I., “Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.

[8] Muller, K.-R. and Mika, S. and Ratsch, G. and Tsuda, K. and Scholkopf, B., “An introduction to kernel-based learning algorithms,” *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[9] Fisher, R.A., “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[10] Cortes,C. and Vapnik,V., “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] Hutson, V. and Pym, J.S., *Practical Methods of Optimization*, second ed. New York: John Wiley, 1987.

[12] Scholkopf,B. Smola,A. and Muller,K. R., “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[13] Hutson, V. and Pym, J.S., *Applications of Functional Analysis and Operator Theory*, London: Academic Press, 1980.

[14] Yang,J. and Frangi,A.F. and Yang,J. and Zhang, D. and Jin, Z., “KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[15] Kreyszig, E., *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1978.

[16] Tian,Y. and Bulle,R. M., “Automatic detecting neutral face for face authentication,” in *AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management*, California, USA, pp., 24–26.

[17] Kanade,T. and Cohn,J. and Tian,Y., “Comprehensive databases for facial expression analysis,” in *Proc. IEEE Inter. Conf. on Face and Gesture Recognition*, Grenoble, France, pp., 46–53.

[18] M. Pantic AND L. J. M. Rothkrantz, “Expert system for automatic analysis of facial expressions,” *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, August 2000.