01-07

Van Gestel T., Suykens J.A.K., De Brabanter J., De Moor B., Vandewalle J., ``Least squares support vector machine regression for discriminant analysis", in *Proc. of the International Joint Conference on Neural Networks (IJCNN'01)*, Washington DC, USA, Jul. 2001, pp. 2445-2450., Lirias number: 178685.

Least Squares Support Vector Machine Regression for Discriminant Analysis

Tony Van Gestel, Johan A.K. Suykens, Jos De Brabanter,
Bart De Moor & Joos Vandewalle

K.U.Leuven, Dept. of Electrical Engineering, ESAT-SISTA,
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
E-mail: {tony.vangestel,johan.suykens}@esat.kuleuven.ac.be.

Abstract

Support Vector Machine classifiers aim at constructing a large margin classifier in the feature space, while a nonlinear decision boundary is obtained in the input space by mapping the inputs in a nonlinear way to a possibly infinite dimensional feature space. Mercer's condition is applied to avoid an explicit expression for the nonlinear mapping and the solution follows from a finite dimensional quadratic programming problem. Recently, other classifier formulations related to a regularized form of Fisher Discriminant Analysis have been proposed in the feature space for which practical expressions are obtained in a second step by applying the Mercer condition. In this paper, we relate existing these techniques to least squares support vector machines, for which the solution follows from a linear Karush-Kuhn-Tucker system in the dual space. Based on the link with empirical linear discriminant analysis one can adjust the bias term in order to take prior information on the class distributions into account and to analyze unbalanced training sets.

Keywords. Support Vector Machines, Regression, Classification, Discriminant Analysis, Regularization

1 Introduction

In Support Vector Machines (SVMs) for classification [14], the inputs are mapped into a high dimensional feature space in which a large margin classifier is constructed. The solution follows from a finite dimensional Quadratic Programming problem in the dual space using the related positive definite kernel function by applying Mercer's theorem [3, 10, 12, 14]. While Vapnik's large margin formulation has been motivated from statistical learning theory, other kernel-based classification approaches have been proposed which are based on a regularized form of Fisher Discriminant Analysis in the feature space [2, 7].

Different problem formulations for linear discriminant analysis in the input space have already been proposed, like, e.g., solving a generalized eigenvalue problem, canonical correlation analysis and least squares regression [4, 5]. All these formulations result into equivalent linear discriminant functions representing the multidimensional classification problem into one-dimensional classification problem. The class label is assigned in the second step, where the bias term of the corresponding classifier is used or where the bias term has to be inferred from prior knowledge on the class distributions. In this paper, we will relate existing techniques in the feature space and show how to choose the bias term when the training set is unbalanced, i.e., the numbers of positively and negatively labeled data points are not equal. Practical expressions are then obtained using a regression interpretation of Least Squares Support Vector Machines (LS-SVMs) for classification [11].

This paper is organized as follows. In Section 2, linear discriminant analysis and the corresponding Bayesian decision rule are reviewed. The relations between the different classifier formulations in the feature space are discussed in Section 3. A practical algorithm based on the regression interpretation is presented in Section 4. The approach is validated on an example with an unbalanced training set in Section 5.

2 Linear Discriminant Analysis

We consider a binary classification problem between two classes $\mathcal{C}_+$, $\mathcal{C}_-$ with corresponding class labels $y = \pm 1$ and with n-dimensional input vector $x \in \mathbb{R}^n$. It is assumed that the inputs x corresponding to the class $\mathcal{C}_+$ are distributed according to a multivariate normal distribution with mean $m_+ = \mathcal{E}\{x|y = +1\}$ and co-

variance matrix $\Sigma_+ = \mathcal{E}\{(x-m_+)(x-m_+)^T|y=+1\}$, which is denoted as follows: $x \sim \mathrm{N}(m_+, \Sigma_+)$. Similarly, it is assumed that the input vector of class $\mathcal{C}_-$ are distributed as $x \sim \mathrm{N}(m_-, \Sigma_-)$, with corresponding mean $m_- = \mathcal{E}\{x|y=-1\}$ and covariance matrix $\Sigma_- = \mathcal{E}\{(\varphi(x_i)-m_-)(\varphi(x_i)-m_-)^T|y=-1\}$. Given the prior class probabilities $P(y=+1)$ and $P(y=-1)$, one assigns a new class label y given the feature vector x as follows: Decide $y=+1$ if $\log[P(y=+1|x,m_+,\Sigma_+)] > \log[P(y=+1|x,m_-,\Sigma_-)]$; else decide $y=-1$.

Applying Bayes' rule, we obtain

$$P(y=+1|x,m_+,\Sigma_+) = \frac{P(x|m_+,\Sigma_+)P(y=+1)}{P(x|m_+,\Sigma_+)P(y=+1)+P(x|m_-,\Sigma_-)P(y=-1)}, \quad (1)$$

where the denominator follows from normalization. Filling in the normal densities $P(x|m_+,\Sigma_+) = (2\pi)^{-n/2}\det(\Sigma_+)^{-1/2}\exp[-\frac{1}{2}(x-m_+)^T\Sigma_+^{-1}(x-m_+)]$ and $P(x|m_-,\Sigma_-) = (2\pi)^{-n/2}\det(\Sigma_-)^{-1/2}\exp[-\frac{1}{2}(x-m_-)^T\Sigma_-^{-1}(x-m_-)]$ a quadratic decision boundary is obtained [5]. When both covariance matrices are identical $\Sigma = \Sigma_+ = \Sigma_-$, the quadratic terms cancel and the following linear discriminant is obtained:

$$w^T(x-m) \geq \log P(y=-1) - \log P(y=-1), \quad (2)$$

with the linear discriminant

$$w = \Sigma^{-1}(m_+ - m_-) \quad (3)$$

and mean of both classes equal to $m = \frac{m_+ + m_-}{2}$. Defining the bias term

$$b = -w^T m - \log P(y=-1) + \log P(y=+1), \quad (4)$$

the Bayesian classification rule becomes: Decide $y=+1$ if

$$y = \mathrm{sign}[w^T x + b] \quad (5)$$

and decide $y=-1$ *otherwise.* Observe that the classification decision does not only depends upon the multivariate normal distribution defined by m_+, m_- and Σ, but also on the prior class probabilities $P(y=+1)$ and $P(y=-1)$; e.g., when $P(y=+1)/P(y=-1)$ increases, the bias term b from (4) will also increase since it will become more likely to decide $y=+1$.

3 Classification Approaches in the Feature Space

While the Bayesian classifier (5) was derived when the distribution of both classes $\mathcal{C}_+$ and $\mathcal{C}_-$ was known, these distributions have to be estimated from training data $D = \{(x_i, y_i)\}_{i=1}^N$. Different problem formulations have been proposed in order to construct a binary classifier from given training data D. First, we discuss Regularized Kernel Discriminant analysis (RKD) [5]. Then we show its equivalence with Regularized Kernel Fisher Discriminant (RKFD) [4]. While the RKFD involves the solution of a generalized eigenvalue problem, a computationally simpler ridge regression approach is discussed in the last Subsection.

Also observe that the Bayesian classifier (5) is linear in the input vector, while real-life problems may be nonlinear. Therefor, we will extend the discussion to a nonlinear class of algorithms related to Support Vector Machines (SVMs) and related kernel based algorithms [2, 6, 7, 10, 11, 12, 13, 14]. In these algorithms, one typically preprocesses the inputs $x \in \mathbb{R}^n$ in a nonlinear way by mapping them into a high-dimensional (possibly infinite dimensional) feature space using the nonlinear function $\varphi(\cdot)$: $\mathbb{R}^n \to \mathbb{R}^{n_h}$. On the other hand, this nonlinear function is not explicitly known and instead the corresponding positive definite kernel function $K(x_1,x_2) = \varphi(x_1)^T\varphi(x_2)$ is used by applying the Mercer condition. For the kernel function $K(\cdot,\cdot)$ one typically has the following choices: $K(x,x_i) = x_i^T x$ (linear kernel); $K(x,x_i) = (x_i^T x+1)^d$ (polynomial kernel of degree d) and $K(x,x_i) = \exp\{-\|x-x_i\|_2^2/\sigma^2\}$ (RBF kernel) where $d \in \mathbb{N}$ and $\sigma \in \mathbb{R}^+$ are constants.

We restrict the discussion to the feature space. A practical algorithm in the dual space based upon the derived equivalences will be presented in the next Section by applying the Mercer condition.

3.1 Regularized Kernel Discriminant Analysis in the Feature Space

In practice, the class centers m_+ and m_- and the corresponding covariance matrices Σ_{Φ_+} and Σ_{Φ_-} are not known and the classifier has to be estimated from the data $D = \{(x_i,y_i)\}_{i=1}^N$ consisting of N_+ and N_- data points of class $\mathcal{C}_+$ and $\mathcal{C}_-$, with sample means

$$\hat{m}_{\Phi_+} = \frac{1}{N_+}\sum_{i=1|y_i=+1}^N \varphi(x_i)$$
$$\hat{m}_{\Phi_-} = \frac{1}{N_-}\sum_{i=1|y_i=-1}^N \varphi(x_i)$$

and covariance matrices[1]

$$\hat{\Sigma}_{\Phi_+} = \frac{1}{N_+}\sum_{i=1|y_i=+1}^N (\varphi(x_i)-\hat{m}_{\Phi_+})(\varphi(x_i)-\hat{m}_{\Phi_+})^T$$
$$\hat{\Sigma}_{\Phi_-} = \frac{1}{N_-}\sum_{i=1|y_i=-1}^N (\varphi(x_i)-\hat{m}_{\Phi_-})(\varphi(x_i)-\hat{m}_{\Phi_-})^T,$$

[1] An alternative way to estimate the covariance matrices is by replacing N_+ and N_- by N_+-1 and N_--1, respectively, so as to obtain an unbiased estimate. A similar argument holds for (6).

respectively. Assuming equal covariance matrices, $\Sigma_\Phi = \Sigma_{\Phi_+} = \Sigma_{\Phi_-}$ can be estimated as

$$\hat{\Sigma}_\Phi = \frac{N_+\hat{\Sigma}_{\Phi_+} + N_-\hat{\Sigma}_{\Phi_-}}{N_+ + N_-}. \quad (6)$$

The prior class probabilities $P(y = +1)$ and $P(y = -1)$ are either known from the case at hand or can be estimated from the data D:

$$P(y = +1) = \frac{N_+}{N_+ + N_-} \quad (7)$$

$$P(y = -1) = \frac{N_-}{N_+ + N_-}. \quad (8)$$

However, the sampling of the covariance matrix will result in a poor estimate of the covariance matrices $\hat{\Sigma}_+$ and $\hat{\Sigma}_-$ with typically poor estimates of the smallest eigenvalues. Also observe that in the case $n_f > N$, the estimated covariance matrix will become rank deficient and the inverse may not exist. Motivated by an empirical Bayesian analysis, one adds a regularization term $\gamma^{-1} I_{n_f}$ to the covariance matrix $\hat{\Sigma}_\Phi$ [5]:

$$\hat{\Sigma}_{\Phi,r} = \hat{\Sigma}_\Phi + \gamma^{-1} I_{n_f}, \quad (9)$$

where the tuning parameter $\gamma \in \mathbb{R}^+$ is determined, e.g., by cross-validation methods.

According to RKD, the class label y is assigned to a new input x as $y = \text{sign}[w_{\text{RKD}}^T \varphi(x) + b_{\text{RKD}}]$, where

$$w_{\text{RKD}} = \hat{\Sigma}_\Phi^{-1}(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}) \quad (10)$$

$$b_{\text{RKD}} = -w_{\text{RKD}}^T \frac{\hat{m}_{\Phi_+} + \hat{m}_{\Phi_-}}{2} + \log(\frac{P(y=-1)}{P(y=+1)}). \quad (11)$$

Observe that w_{RKD} and b_{RKD} are never calculated explicitly in this way since the mapping $\varphi(x)$ is not explicitly known. Using equivalences with other classifier formulations in the next Subsections, the discriminant function will be expressed in terms of the corresponding positive definite kernel function $K(x_1, x_2) = \varphi(x_1)^T \varphi(x_2)$ by applying Mercer's theorem.

3.2 Regularized Kernel Fisher Discriminant

A useful criterion [4] to select the optimal linear decision rule $w_{\text{RKFD}}^T \varphi(x) + b_{\text{RKFD}}$ would be the distance $|w_{\text{RKFD}}^T(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-})|$ between the projected class centers, which is somewhat related to the large margin idea in the SVM formulation [14]. Notice that an arbitrary increase of this distance is obtained by a simple scaling of the discriminant w_{RKFD}. A good discriminant minimizes the overlap between the distributions, which is characterized, up to second order characteristics, by both the distance $|w_{\text{RKFD}}^T(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-})|$ and the variance or scatter $s = s_+ + s_-$ around the projected class centers, with

$$s_+ = \sum_{i=1|y_i=+1}^N (w_{\text{RKFD}}^T(\varphi(x_i) - \hat{m}_{\Phi_+}))^2 \quad (12)$$

$$s_- = \sum_{i=1|y_i=-1}^N (w_{\text{RKFD}}^T(\varphi(x_i) - \hat{m}_{\Phi_-}))^2. \quad (13)$$

In order to avoid overfitting, we will punish for large weights of w_{RKFD} by adding a regularization term $\gamma^{-1} w_{\text{RKFD}}^T w_{\text{RKFD}}$ to the variance s. The linear discriminant w_{RKFD} is then found by minimizing

$$\min_{w_{\text{RKFD}}} \mathcal{J}(w_{\text{RKFD}}) = \frac{(w_{\text{RKFD}}^T(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}))^2}{s + \gamma^{-1} w_{\text{RKFD}}^T w_{\text{RKFD}}}. \quad (14)$$

This optimization problem can be written as a generalized Rayleigh quotient and generalized eigenvalue problem by defining the within class scatter matrix $S_{\mathcal{W},\Phi} = S_{\mathcal{W},\Phi_+} + S_{\mathcal{W},\Phi_-} + N\gamma^{-1} I_{n_f}$ and the between class scatter matrix $S_{\mathcal{B},\Phi}$:

$$S_{\mathcal{W},\Phi_+} = \sum_{i=1|y_i=+1}^N (\varphi(x_i) - \hat{m}_{\Phi_+})(\varphi(x_i) - \hat{m}_{\Phi_+})^T = \frac{1}{N_+}\hat{\Sigma}_{\Phi_+} \quad (15)$$

$$S_{\mathcal{W},\Phi_-} = \sum_{i=1|y_i=-1}^N (\varphi(x_i) - \hat{m}_{\Phi_-})(\varphi(x_i) - \hat{m}_{\Phi_-})^T = \frac{1}{N_-}\hat{\Sigma}_{\Phi_-} \quad (16)$$

$$S_{\mathcal{B},\Phi} = (\hat{m}_+ - \hat{m}_-)(\hat{m}_+ - \hat{m}_-)^T. \quad (17)$$

Using $s + N\gamma^{-1} w_{\text{RKFD}}^T w_{\text{RKFD}} = w_{\text{RKFD}}^T S_{\mathcal{W},\Phi} w_{\text{RKFD}}$ and $(w_{\text{RKFD}}^T(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}))^2 = w_{\text{RKFD}}^T S_{\mathcal{B},\Phi} w_{\text{RKFD}}$, (14) can be written as the generalized Rayleigh quotient $\mathcal{J}(w_{\text{RKFD}}) = \frac{w_{\text{RKFD}}^T S_{\mathcal{B},\Phi} w_{\text{RKFD}}}{w_{\text{RKFD}}^T S_{\mathcal{W},\Phi} w_{\text{RKFD}}}$ for which the linear discriminant w_{RKFD} is found as the solution to the generalized eigenvalue problem

$$S_{\mathcal{B},\Phi} w_{\text{RKFD}} = \lambda_w S_{\mathcal{W},\Phi} w_{\text{RKFD}}. \quad (18)$$

Since $S_{\mathcal{B},\Phi} w_\Phi \propto (\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-})$, the solution w_{RKFD} is given by

$$w_{\text{RKFD}} = S_{\mathcal{W},RKFD}^{-1}(m_{\Phi_+} - m_{\Phi_-}) = N w_{\text{RKD}}, \quad (19)$$

since $S_{\mathcal{W},RKFD} = N(\hat{\Sigma}_\Phi + \gamma^{-1} I_{n_f})$. In [2, 7] the corresponding classifier is obtained by solving a generalized eigenvalue problem in the dual space, the the bias term is determined, e.g., by cross-validation. The linear discriminant (19) results into an optimal discrimination between the two classes $\mathcal{C}_+$ and $\mathcal{C}_-$, while the optimal bias term b_{RKFD} is equal to

$$b_{\text{RKFD}} = -w_{\text{RKFD}}^T(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}) - N \log(\frac{P(y=+1)}{P(y=-1)}). \quad (20)$$

3.3 Regression Approaches to Classification

Assume now a regression approach, where one assigns targets $t_+ = +1$ and $t_- = -1$ to classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively. Then one can formulate the regression approach:

$$\min_{w_{\text{R}}, b_{\text{R}}} \mathcal{J}(w_{\text{R}}, b_{\text{R}}) = \frac{\gamma^{-1}}{2} w_{\text{R}}^T w_{\text{R}} + \frac{1}{2}\sum_{i=1}^N (t_i - (w_{\text{R}}^T \varphi(x_i) + b_{\text{R}}))^2, \quad (21)$$

where γ is a tuning parameter that controls the tradeoff between regularization and error minimization. Putting the derivatives of (21) with respect to w_{R} and b_{R} equal to zero, the normal equations are obtained

$$\begin{bmatrix} \sum_{i=1}^{N} \varphi(x_i)\varphi(x_i)^T + \gamma^{-1} I_{n_f} & (N_+\hat{m}_{\Phi_+} + N_-\hat{m}_{\Phi_-}) \\ (N_+\hat{m}_{\Phi_+} + N_-\hat{m}_{\Phi_-})^T & N \end{bmatrix} \cdot \begin{bmatrix} w_{\mathrm{R}} \\ b_{\mathrm{R}} \end{bmatrix} = \begin{bmatrix} N_+t_+\hat{m}_{\Phi_+} - N_-t_-\hat{m}_{\Phi_-} \\ \sum_{i=1}^{N} t_i \end{bmatrix}. \quad (22)$$

From the second equation, we obtain $b_{\mathrm{R}} = -\frac{1}{N_+ + N_-}(N_+\hat{m}_{\Phi_+} + N_-\hat{m}_{\Phi_-})^T w_{\mathrm{R}} + \frac{1}{N_+ + N_-}(N_+t_+ - N_-t_-)$. Substituting this relation into the first equation, we obtain

$$\left(S_{W,\Phi} + \frac{N_+N_-}{N^2}(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-})(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-})^T\right) w_{\mathrm{R}} = \frac{N_+N_-}{N^2}(t_+ - t_-)(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}). \quad (23)$$

Observe that the between class scatter matrix only results in a scaling of the solution. Hence, w is proportional to w_{RKFD}, i.e., $w = \alpha_w w_{\mathrm{RKFD}} = N\alpha_w w_{\mathrm{RKD}}$ with

$$w_{\mathrm{RKFD}} = S_{W,\Phi}^{-1}(\hat{m}_{\Phi_+} - \hat{m}_{\Phi_-}). \quad (24)$$

Substituting this expression into (23), we obtain the scale factor α_w:

$$\alpha_w = \frac{N_+N_-}{N^2}(w_{\mathrm{R}}^T(\hat{m}_{\Phi_-} - \hat{m}_{\Phi_+}) + t_+ - t_-). \quad (25)$$

Solving the least squares problem (21) for w_{R} and calculating the scale factor α_w from (25), we obtain w_{RKFD} and w_{RKFD} as

$$w_{\mathrm{RKD}} = N w_{\mathrm{RKFD}} = \alpha_w N w_{\mathrm{R}}. \quad (26)$$

This allows us to calculate b_{RKD} from (11) and assign the class label y in an optimal way to the new input x.

For single output problems there exist explicit links between Kernel Canonical Correlation Analysis (CCA) [6, 13] and least squares regression, which allows to derive the classifier formulation starting from CCA.

4 LS-SVM Classifier and Discriminant Analysis

In the previous Section, it was shown that all classifier formulations yielded equivalent discriminants. Not only the linear discriminant function, but also the classifier could be obtained from the regression solution. While this analysis was performed in the feature space, the (nonlinear and possibly infinite dimensional) mapping $\varphi(\cdot)$ from the input space to the feature space is not explicitly known for Support Vector Machines. Instead, the Mercer condition is applied and the related Kernel function $K(x_1, x_2) = \varphi(x_1)^T\varphi(x_2)$ is to obtain practical expressions. In [11], Vapnik's SVM formulation for classification was modified by introducing equality constraints and a 2-norm. This resulted in the Least Squares Support Vector Machine classifier formulation

$$\min_{w_C, b_C} \mathcal{J}(w_{\mathrm{C}}, b_{\mathrm{C}}) = \tfrac{1}{2} w_{\mathrm{C}}^T w_{\mathrm{C}} + \tfrac{\gamma}{2} \sum_{i=1}^{N} e_{\mathrm{C},i}^2, \quad (27)$$

subject to the constraints

$$y_i(w_{\mathrm{C}}^T\varphi(x) + b_{\mathrm{C}}) = 1 - e_{\mathrm{C},i}, \qquad i = 1, \ldots, N. \quad (28)$$

This formulation corresponds to a regression approach with binary targets $t_i = y_i \in \{-1, +1\}$. Indeed, by multiplying the error $e_{\mathrm{C},i}$ with $y_i \in \{-1, +1\}$, the error term can be written as: $\sum_{i=1}^{N} e_{\mathrm{C},i}^2 = \sum_{i=1}^{N}(y_i e_i)^2 = \sum_{i=1}^{N} e_i^2$, with

$$e_i = y_i - (w_{\mathrm{R}}^T\varphi(x) + b_{\mathrm{R}}), \qquad i = 1, \ldots, N. \quad (29)$$

In Least Squares Support Vector Machines for Regression [11], one minimizes the following cost function

$$\min_{w_{\mathrm{R}}, b_{\mathrm{R}}} \mathcal{J}(w_{\mathrm{R}}, b_{\mathrm{R}}) = \tfrac{1}{2} w_{\mathrm{R}}^T w_{\mathrm{R}} + \tfrac{\gamma}{2} \sum_{i=1}^{N} e_i^2, \quad (30)$$

subject to the constraints (29).

This formulation is equivalent with (21) up to the scale factor γ. In order to avoid the explicit expression for $\varphi(\cdot)$, one solves the the minimization problem (30)-(29), by constructing the Lagrangian $\mathcal{L}(w_{\mathrm{R}}, b_{\mathrm{R}}, e; \alpha) = \mathcal{J}(w_{\mathrm{R}}, e) - \sum_{i=1}^{N} \alpha_i(w_{\mathrm{R}}^T\varphi(x_i) + b_{\mathrm{R}} + e_i - y_i)$, where $\alpha_i \in \mathbb{R}$ $(i = 1, \ldots, N)$ are the Lagrange multipliers (also called support values). By taking the conditions for optimality and eliminating w and e, the following linear Karush-Kuhn-Tucker system is obtained [11, 12]:

$$\left[\begin{array}{c|c} 0 & 1_v^T \\ \hline 1_v & \Omega + \gamma^{-1} I_N \end{array}\right] \left[\begin{array}{c} b_{\mathrm{R}} \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \quad (31)$$

with $I_N \in \mathbb{R}^{N \times N}$ the identity matrix and[2] $y = [y_1; \ldots; y_N]$, $1_v = [1; \ldots; 1]$, $\alpha = [\alpha_1; \ldots; \alpha_N]$. Mercer's condition [3, 9, 10, 14] is applied within the Ω matrix

$$\Omega_{ij} = \varphi(x_i)^T\varphi(x_j) = K(x_i, x_j). \quad (32)$$

The LS-SVM regressor is then obtained by applying the Mercer condition:

$$w_{\mathrm{R}}^T\varphi(x) + b_{\mathrm{R}} = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b_{\mathrm{R}}. \quad (33)$$

[2] The matlab notation $[X_1; X_2]$ is used, where $[X_1; X_2] = [X_1^T\ X_2^T]^T$.

From this discriminant, the scale factor α_w (25) is expressed in the dual space

$$\alpha_w = \frac{N_+N_-}{N^2}(t_+ + \frac{1}{N_-}\sum_{j=1}^{N}\alpha_j\sum_{i=1|y_i=-1}^{N}K(x_i,x_j) - \frac{1}{N_+}\sum_{j=1}^{N}\alpha_j\sum_{i=1|y_i=+1}^{N}K(x_i,x_j) - t_-). \quad (34)$$

The support values $\alpha_{\text{RKD},i} = \alpha_w\alpha$ of the RKD $w_{\text{RKD}} = \sum_{i=1}^{N}\alpha_{\text{RKD},i}\varphi(x_i)$ are now obtained. The bias term b_{RKD} is equal to

$$b_{\text{RKD}} = -\frac{1}{2N_+}\sum_{j=1}^{N}\sum_{i=1|y_i=1}^{N}\alpha_{\text{RKD},j}K(x_i,x_j) - \frac{1}{2N_+}\sum_{j=1}^{N}\sum_{i=1|y_i=-1}^{N}\alpha_{\text{RKD},j}K(x_i,x_j) + \log(\frac{P(y=-1)}{P(y=+1)}). \quad (35)$$

Efficient algorithms exist in numerical linear algebra for solving large scale systems. By reformulating the linear system (31) into two linear systems with positive definite data matrices as in [12], iterative methods can be applied such as, e.g., the Hestenes-Stiefel conjugate gradient algorithm, which are computationally more interesting than solving a generalized eigenvalue problem in the dual space [2, 7]. The scale factor α_w and the bias term b_{RKD} are easily obtained from (34) and (35) in order to obtain the regularized kernel discriminant

$$y = \text{sign}[\sum_{i=1}^{N}\alpha_{\text{RKD},i}K(x_i,x) + b_{\text{RKD}}]. \quad (36)$$

5 Example

We illustrate the approaches for the use of an RBF-kernel on the synthetic binary classification dataset from [8]: first we will train the classifier on a balanced data set with equal prior class probabilities, then we will illustrate the bias correction on an intentionally unbiased training set. The data set [8] consists of a training set and test set of $N = 250$ and $N_{test} = 1000$ data points, respectively. Both data sets are balanced, i.e., both classes $\mathcal{C}_-$ and $\mathcal{C}_+$ have equal prior probabilities. Each class is an equal mixture of two normal distributions. Due to the overlap of the distributions, the optimal theoretical performance that can be achieved is 92.0%. Both inputs were normalized to zero mean and unit variance [1]. The corresponding training data and test set data are depicted in Figure 1 and 2, respectively. Using Bayesian inference [1], we obtained the hyperparameters $\sigma = 1.113$ and $\gamma = 1.6445$.

First we trained the LS-SVM on the full training set, which yields a test set performance of 90.6%. The corresponding decision line is depicted by the dotted line in Figures 1, 2 and 3. Then we intentionally unbalanced the training data set by removing the even data

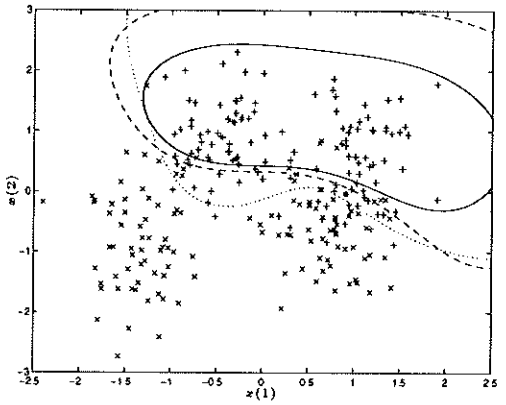


Figure 1: Original training set for the synthetic binary data set [8]. The training data of classes $\mathcal{C}_-$ an $\mathcal{C}_+$ are denoted by × and +, respectively. The corresponding decision boundaries for the classifiers obtained on the balanced and unbalanced data set are depicted by the dotted and full line, respectively. After adjusting the bias term, the improved dashed decision boundary is obtained.

points from class $\mathcal{C}_-$. The training data set now consists of $N_- = 63$ data points in class $\mathcal{C}_-$ and $N_+ = 125$ data points in class $\mathcal{C}_+$. The corresponding unbalanced training data set is depicted in Figure 3. The test set was not changed. We now trained an LS-SVM classifier on this unbalanced data set with the same (now suboptimal set) of hyperparameters and obtained an out-of-sample performance of 83.1% on the balanced

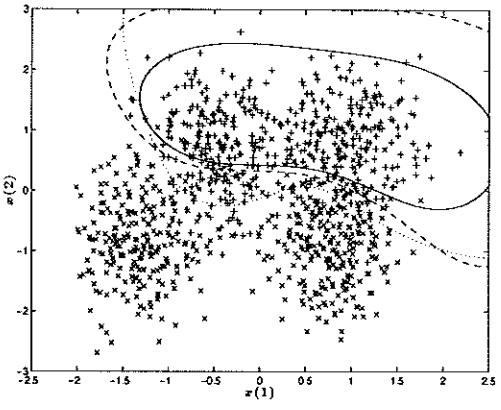


Figure 2: Test set for the synthetic binary data set [8]. The test data of classes $\mathcal{C}_-$ an $\mathcal{C}_+$ are denoted by × and +, respectively. The decision boundaries corresponding to the different classifiers are depicted using the convention of Figure 1.

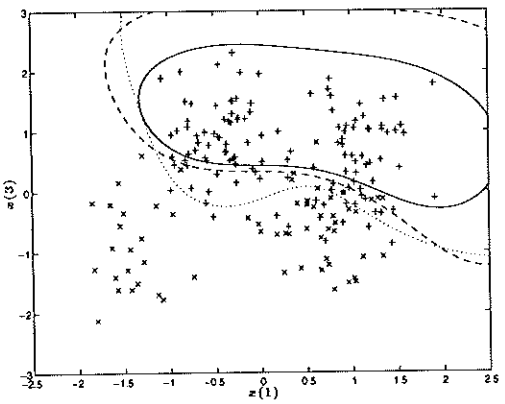


Figure 3: Unbalanced training set obtained by removing even data points of class $\mathcal{C}_-$. The training data of classes $\mathcal{C}_-$ an $\mathcal{C}_+$ are denoted by $\times$ and $+$, respectively. The decision boundaries corresponding to the different classifiers are depicted using the convention of Figure 1.

test set. The corresponding decision boundary is depicted by the full line in Figures 1, 2 and 3. Taking into account the unbalanced training set, we adjusted the bias term according to (35) which resulted in an increased test set performance of 87.1%. Notice that this increase is because the classification performance in the overlap between the two classes is much better. The corresponding decision line is depicted by the dashed line in Figures 1, 2 and 3. The decision line is now much closer to the original decision boundary (dotted line) obtained on the balanced training set. Its accuracy is lower because of the reduced number of training data and because the hyperparameters σ and γ were not adjusted.

6 Conclusions

In this paper, several kernel based classification techniques have been related with regularized discriminant analysis in the feature space, while a practical expression is obtained in a second step by applying the Mercer condition for the regression interpretation. The corresponding framework also allows to use prior knowledge on the class distributions and to adjust the bias term when the training or test set are unbalanced.

Acknowledgments

T. Van Gestel and J. Suykens are a Research Assistant and a Postdoctoral Researcher with the Fund for Scientific Research-Flanders (FWO-Vlaanderen), respectively. This work was partially supported by grants and projects from the Flemish Gov.: (Research council KULeuven: GOA-Mefisto 666; FWO-Vlaanderen: Res. Proj. G.0240.99, G.0256.97, and Comm. (ICCoS and ANMMM); AWI: Bil. Int. Coll.; IWT: STWW, IMPACT); from the Belgian Fed. Gov. (IUAP-IV/02, IV/24; Program Dur. Dev.); from the Eur. Comm.: (TMR Netw. Niconet; Science ERNSI). The scientific responsibility is assumed by its authors.

References

[1] Bishop, C.M. *Neural Networks for Pattern Recognition*, Oxford University Press.

[2] Baudat, G., & Anouar, F. Generalized Discriminant Analysis Using a Kernel Approach. *Neural Computation*, 12, 2385-2404, 2000.

[3] Cristianini, N., & Shawe-Taylor, J. *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[4] Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7(2), 179-188, 1936.

[5] Friedman, J. "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165-175, 1989.

[6] Lai, P.L., & Fyfe, C. "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, 2001, accepted for publication.

[7] Mika, S., Rätsch, G., & Müller, K.-R. "A Mathematical Programming Approach to the Kernel Fisher Algorithm," in *Proc. Advances in Neural Information Processing Systems NIPS 13*, 2001. (To appear).

[8] Ripley, B.D. "Neural Networks and Related Methods for Classification," *Journal Royal Statistical Society B, 56(3)*, 409-456, 1994.

[9] Saunders, C., Gammerman, A., & Vovk, V. "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. on Machine Learning ICML-98*, pp. 515-521, Madison-Wisconsin, 1998.

[10] Smola, A., & Schölkopf, B., "On a kernel-based method for pattern recognition, regression, approximation and operator inversion," *Algorithmica*, vol. 22, pp. 211-231, 1998.

[11] Suykens, J.A.K., & Vandewalle, J., "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293-300, 1999.

[12] Suykens, J.A.K., "Least Squares Support Vector Machines for Classification and Nonlinear Modelling," *Neural Network World, Special Issue on PASE'2000*, vol. 10, pp. 29-48, 2000.

[13] Van Gestel, T., Suykens J.A.K., De Brabanter, J., De Moor, B., & Vandewalle, J. "Kernel Canonical Correlation Analysis and Least Squares Support Vector Machines", In *Proc. International Conference on Artificial Neural Networks (ICANN 2001)*, Vienna, Austria, 2001, to appear.

[14] Vapnik, V., *Statistical Learning Theory*, Wiley, New-York, 1998.