A method for designing nonlinear kernel-based discriminant functions from the class of second-order criteria

Fahed Abdallah, Cédric Richard, Régis Lengellé

Laboratoire de Modélisation et Sûreté des Systèmes (LM2S),

Université de Technologie de Troyes (UTT),

12 rue Marie Curie, B.P. 2060,

10010 Troyes cedex, France.

tel: +33.3.25.71.56.92 fax: +33.3.25.71.56.99

E-mail: fahed.abdallah@utt.fr

Extended abstract (500-1000 words)

Data-driven design of linear receivers $d(X) = X^T W \gtrless \nu$ consists in finding optimum W and ν in the sense of a preselected criterion, e.g., a second-order criterion such as Fisher criterion or generalized signal-to-noise ratio [Richard et al, 2002], from a data set $\{(X_i, Y_i)\}_{i=1,\ldots,n}$. Here the X_i's are training samples and the Y_i's indicate either class $\mathcal{C}_0$ or $\mathcal{C}_1$. Let m_i and σ_i^2 be the first and second-order conditional moments of $d(X)$, respectively. In [Fukunaga, 1990] and [Richard, 1998], it has been shown that linear discriminant functions optimizing second-order criteria $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ can be derived by solving the following linear system

$$[\rho\Sigma_0 + (1-\rho)\Sigma_1]\, W = [M_1 - M_0], \tag{1}$$

where M_i and Σ_i denote respectively the conditional expected vectors and covariance matrices of the observation X. The parameter ρ depends on the second-order criterion $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ as follows:

$$\rho = \frac{\frac{\partial \Psi}{\partial \sigma_0^2}}{\frac{\partial \Psi}{\partial \sigma_0^2} + \frac{\partial \Psi}{\partial \sigma_1^2}}, \tag{2}$$

Since W in (1) only depends on $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ via ρ, this parameter can be adjusted to pick the receiver $d(X)$ that has the best performance. Obviously, the resulting structure performs better than or equal to the receiver maximizing the Fisher criterion, which corresponds to $\rho = \mathbf{P}(X \in \mathcal{C}_0)$. This approach thus leads to the optimum receiver in the sense of the best second-order criterion $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$. Note that this criterion is never set up.

A simple method to obtain a nonlinear discriminant is to map the samples into a high dimensional feature space $\mathcal{F}$ using a nonlinear function ϕ, and then to perform a linear discriminant analysis in $\mathcal{F}$. Clearly, if $\mathcal{F}$ is a very high, or even infinitely, dimensional space, deriving $d(x)$ may be a computationally intractable problem. However, by using the theory of reproducing kernels [Vapnik, 1995], such a problem can be solved without explicitly mapping into the feature space $\mathcal{F}$. Recently, a powerful method of obtaining nonlinear kernel Fisher discriminant, called KFD method, has been proposed [Mika et al., 1999] and widely studied [Mika et al., 2001-1], [Mika et al., 2001-2]. A closed form solution to this problem has also been obtained in [Billings and Lee, 2002]. In this paper, we present an extension of the KFD method. It is also based on Mercer kernels, and it consists in determining the optimum nonlinear receiver in the sense of the best second-order criterion.

As for the KFD approach, input samples are first mapped into a high-dimensional feature space $\mathcal{F}$ by using a nonlinear function $X \mapsto \phi(X)$. Let Ψ be any second-order criterion. From (1), it follows that the discriminant function $d^\phi(X) = \phi(X)^T W^\phi$ operating in $\mathcal{F}$ is optimum in the sense of any second-order criterion Ψ if it satisfies:

$$\left[\rho\Sigma_0^\phi + (1-\rho)\Sigma_1^\phi\right] W^\phi = [M_1^\phi - M_0^\phi]. \tag{3}$$

where M_i^ϕ and Σ_i^ϕ denote the conditional expected vectors and covariance matrices of $\phi(X)$, respectively. The parameter ρ is defined in (2). Equation (3) may be complicated to solve when $\mathcal{F}$ is a very high-dimensional space. However, one can get around this by using Mercer functions ϕ, which satisfy [Vapnik, 1995]:

$$k(X_i, X_j) = \phi^T(X_i)\,\phi(X_j). \tag{4}$$

The polynomial kernel $k(X_i, X_j) = (1 + X_i^T X_j)^d$, whose separating surface in the input space is a polynomial surface of degree d, and the exponential radial basis kernel $k(X_i, X_j) = \exp\left(-\|X_i - X_j\|/\sigma^2\right)$, which produces a piecewise linear separating surface, are typical kernels that satisfies the Mercer condition. Other examples can be found in [Vapnik, 1995]. As shown below, this property can be used to solve the problem without ever mapping explicitly in the feature space $\mathcal{F}$. Following [Mika et al., 1999], W^ϕ must lie in the span of all training samples in $\mathcal{F}$ according to the theory of reproducing Kernels.

This means that W^ϕ can be expressed as:

$$W^\phi = \sum_{i=1}^{n} \alpha_i \phi(X_i) = \mathbf{Q}\alpha, \tag{5}$$

where n is the number of training samples, and $\mathbf{Q}$ denotes the matrix $[\phi(X_1) \cdots \phi(X_n)]$. Multiplying (3) by $\mathbf{Q}^T$ and using (5) yields

$$\left[\rho\, \mathbf{Q}^T \Sigma_0^\phi \mathbf{Q} + (1-\rho)\, \mathbf{Q}^T \Sigma_1^\phi \mathbf{Q}\right] \alpha = \mathbf{Q}^T [M_1^\phi - M_0^\phi]. \tag{6}$$

Using Mercer condition (4), this expression can be reformulated as follows:

$$\left[\frac{\rho}{n_0}\, \mathbf{K}_0(\mathbf{I} - \mathbf{1}_{n_0})\mathbf{K}_0^T + \frac{1-\rho}{n_1}\, \mathbf{K}_1(\mathbf{I} - \mathbf{1}_{n_1})\mathbf{K}_1^T\right] \alpha = M, \tag{7}$$

where α has to be determined. Here n_i is the number of samples from class $\mathcal{C}_i$ in the training set, $\mathbf{K}_i$ is a n by n_i matrix such that $\mathbf{K}_i(j,l) = k(X_j, X_l)$ with $X_j \in (\mathcal{C}_0 \cup \mathcal{C}_1)$ and $X_l \in \mathcal{C}_i$, $\mathbf{I}$ is the identity matrix and $\mathbf{1}_{n_i}$ is the matrix with all elements set to $\frac{1}{n_i}$. The elements of the vector M are defined as:

$$M(j) = \frac{1}{n_1} \sum_{X \in \mathcal{C}_1} k(X, X_j) - \frac{1}{n_0} \sum_{X \in \mathcal{C}_0} k(X, X_j). \tag{8}$$

The discriminant function under consideration is given by:

$$d^\phi(X) = \phi(X)^T W^\phi = \sum_{i=1}^{n} \alpha_i\, k(X_i, X). \tag{9}$$

Our approach, called nonlinear kernel second-order discriminant (KSOD), is very much in the spirit of KFD. Both are based on the kernel trick that allows the efficient computation of linear discriminants in feature space. They correspond to powerful nonlinear decision functions in input space. However, classifiers obtained with KSOD obviously performs better than or equal to those resulting from KFD, since they are optimum in the sense of the best second-order criterion. Figures 1 and 2 show an illustrative comparison of the performance of the receivers found by KFD and KSOD on a toy data set. Each class consists of two noise parabolic shapes mirrored at the x and y axis, respectively. Figure 1 represents the average training and test errors over 50 runs. Figure 2 gives the separating surfaces in the input space obtained with KFD and KSOD. A polynomial kernel of degree 2 was used and, as in [Mika et al., 1999], the threshold ν was estimated with a linear Support Vector Machine.

References

[Billings and Lee, 2002] S. A. Billings and K. L. Lee. Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. *Neural Networks*, vol. 15, 2002.

[Fukunaga, 1990] K. Fukunaga. *Statistical pattern recognition.* Academic Press, 1990.

[Mika et al., 2001-1] S. Mika, G. Rätsch and K. R. Müller. A mathematical programming approach to the kernel Fisher algorithm. *Advances Neural Information Processing Systems XIII*, T. K. Leen, T. G. Dietterich, V. Tresp, editors, pp. 591-597, MIT Press, 2001.

[Mika et al., 2001-2] S. Mika, A. J. Smola and B. Schölkopf. An improved training algorithm for kernel Fisher discriminants. *Proceedings AISTATS 2001*, T. Jaakkola, T. Richardson, editors, pp. 98-104, Morgan Kaufmann, 2001.

[Mika et al., 1999] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K. R. Müller. Fisher discrimiant analysis with kernels. *Neural networks for signal processing IX*, Y. H. Hu, J. Larsen, E. Wilson, S. Douglas, editors, pp. 41-48, 1999.

[Richard et al, 2002] C. Richard, R. Lengellé and F. Abdallah. Bayes-optimal detectors design using relevant second-order criteria. *IEEE Signal Processing Letters*, vol. 9, no. 1, 2002.

[Richard, 1998] C. Richard. *Une méthodologie pour la détection à structure imposée.* PhD Thesis, Compiègne University of Technology, France, 1998.

[Vapnik, 1995] V. Vapnik. *The nature of statistical learning theory.* Springer Verlag, 1995.

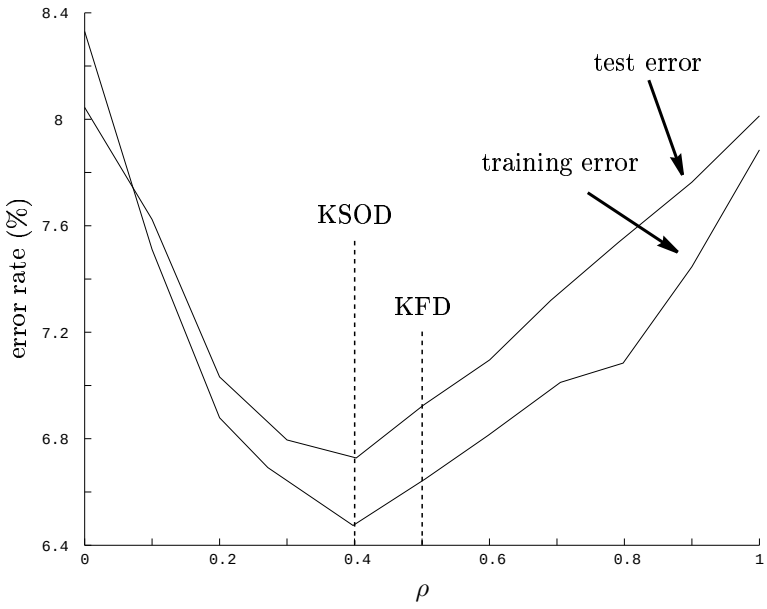

Figure 1: Comparison of the error rates of the receivers found with KFD and KSOD. This experiment shows that the solution obtained with KSOD method performs better ($\rho = 0.4$) than that resulting from KFD approach ($\rho \equiv \mathbf{P}(X \in \mathcal{C}_0) = 1/2$).

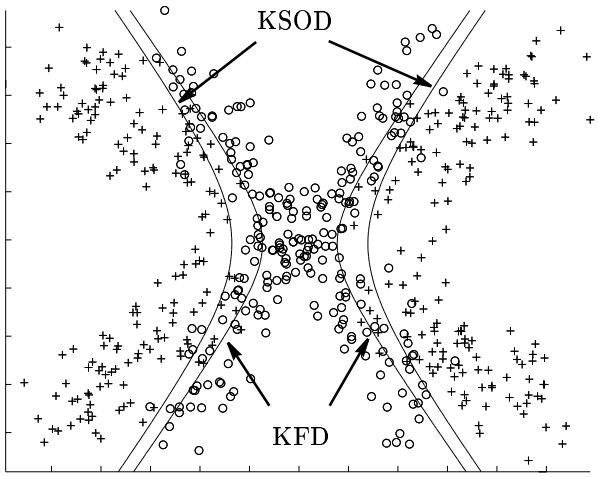

Figure 2: Separating surfaces in the input space obtained with KFD and KSOD. The samples from the two classes are represented by crosses and circles.