KERNEL SECOND-ORDER DISCRIMINANTS VERSUS SUPPORT VECTOR MACHINES

Fahed ABDALLAH, Cédric RICHARD, Régis LENGELLE

Laboratoire LM2S
Université de Technologie de Troyes
B.P. 2060, F-10010 Troyes Cedex, FRANCE
tel: +33.3.25.71.56.92 fax: +33.3.25.71.56.99
fahed.abdallah@utt.fr, cedric.richard@utt.fr, regis.lengelle@utt.fr

ABSTRACT

Support vector machines (SVMs) are the most well known nonlinear classifiers based on the Mercer kernel trick. They generally leads to very sparse solutions that ensure good generalization performance. Recently Mika *et al.* have proposed a new nonlinear technique based on the kernel trick and the Fisher criterion: the nonlinear kernel Fisher discriminant (KFD). Experiments show that KFD is competitive to the SVM classifiers. Nevertheless, it can be shown that there exists distributions such that even though the two classes are linearly separable, the Fisher linear discriminant has an error probability close to 1. In this paper, we propose an alternative strategy based on Mercer kernels that consists in picking the optimum nonlinear receiver in the sense of the best second-order criterion. We also present a strategy for controlling the complexity of the resulting classifier. Finally we compare this new method with SVM and KFD.

1. INTRODUCTION

In the last few years there have been very significant developments in classification methods based on kernels. Support Vector Machines (SVMs) were introduced and first applied as alternatives to multi-layer neural networks [17]. The high generalization ability provided by these learning machines has inspired recent works in discriminant analysis as well as the fundamental theory of model complexity and generalization. SVMs consist in mapping the data into a high dimensional space $\mathcal{F}$ where the two classes of data are more readily separable, and maximizing the margin [16, 17]. Recently, a powerful method of obtaining nonlinear kernel Fisher discriminants (KFD) has been proposed, and very promising results were reported when compared with the other state of the art classification techniques [10]. Nevertheless, it can be shown that there exists distributions such that even though the two classes are linearly separable, the Fisher linear discriminant has an error probability close to 1 [4]. In this paper, we present an extension of the KFD method that is also based on Mercer kernels. Our approach, called *nonlinear kernel second-order discriminant* (KSOD), consists in determining the optimum nonlinear receiver in the sense of the best second-order criterion [5, 6, 13]. In order to obtain a sparse solution, we also propose a strategy to control the complexity of the resulting classifier.

The present paper starts with a brief description of SVMs for binary classification. Next we present our nonlinear approach, which is based on the kernel trick for obtaining a simple, computationally inexpensive algorithm. We also propose a procedure for controlling the complexity of receivers and then improving their generalization performance. Finally, experiments using artificial and real world data are performed in order to make comparisons between our approach, KFD and SVM.

2. SUPPORT VECTOR MACHINES

We consider the binary classification problem with d-dimensional patterns $\{\boldsymbol{x}_i\}_{i=1,\ldots,n}$ having labels $y_i = \pm 1$ that indicate either class $\mathcal{C}_0$ or $\mathcal{C}_1$. The discriminant function approach uses a real-valued function $f(\boldsymbol{x})$, the sign of which determines the class label prediction. SVMs implement complex discriminant functions of the form

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b, \tag{1}$$

by using a nonlinear function $\boldsymbol{\phi}$ to map data into a high dimensional (and possibly infinite dimensional) feature space $\mathcal{F}$. The separating hyperplane $(\boldsymbol{w}, b)$ is found by maximizing the distance between itself and the closest points in the training set, called *support vectors*. Using a linearly separable training set, $\boldsymbol{w}$ and b are solutions of the quadratic programming problem [3, 17]

$$\min_{\boldsymbol{w},b} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 \right), \tag{2}$$

subject to the constraints $y_i(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1$, $\forall i$. When the training data cannot be separated linearly in $\mathcal{F}$, a more general setting can be used to allow misclassified points. This gives rise to a slightly different optimization problem [3, 17]

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + c \sum_{i=1}^{n} \xi_i \right), \tag{3}$$

subject to $y_i(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, $\forall i$. The sum of slack-variables ξ_i is related to the number of misclassification errors, and the positive real constant c is a tuning parameter in the algorithm. Let κ be any kernel that satisfies the Mercer condition. Then we have

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j), \tag{4}$$

which means that $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ corresponds to the inner product of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in $\mathcal{F}$. It can be shown that the optimum discriminant

function (1) in the sense of the quadratic programming problem (3) can be rewritten as

$$f^*(\boldsymbol{x}) = \sum_i^n \upsilon^*(i)\, y_i\, \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b^*. \tag{5}$$

Here the $\upsilon^*(i)$'s are components of a dual vector $\boldsymbol{\upsilon}^*$ which maximizes the following expression

$$\delta(\boldsymbol{\upsilon}) = \sum_{i=1}^n \upsilon(i) - \frac{1}{2} \sum_{i,j=1}^n \upsilon(i)\, \upsilon(j)\, y_i\, y_j\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{6}$$

subject to the constraints

$$\sum_{i=1}^n \upsilon(i)\, y_i = 0, \qquad 0 \le \upsilon(i) \le c. \tag{7}$$

Note that this dual problem is identical to the one obtained in the separable case (2) except for the upper bound c of the Lagrange multipliers $\upsilon^*(i)$. Every pattern $\boldsymbol{x}_i$ such that $0 < \upsilon^*(i) < c$ is called *support vector*. One of the most important property of SVM is that the solution is sparse in $\boldsymbol{\upsilon}^*$, i.e., the $\upsilon^*(i)$'s associated with patterns that are outside the margin area are zero. SVMs would hardly be practical for large data sets without this characteristic [17].

3. SECOND-ORDER DISCRIMINANTS

Kernel Fisher discriminant (KFD) is a nonlinear generalization of Fisher discriminant that is based on Mercer kernels [10], in analogy to SVMs. On a large number of problems, KFD has shown classification accuracies on a par with SVMs. Nevertheless, it is stated in [4] that there are distributions such that even though the two classes are linearly separable, the Fisher linear discriminant has an error probability close to 1. In this section, we show how designing optimum linear receivers in the sense of the best second-order criterion, given any binary classification problem. Next, this approach is generalized to nonlinear kernel discriminant functions using the Mercer trick. Finally, performance of such receivers are compared with KFDs and SVMs on a large number of benchmarks.

3.1. Linear second-order discriminant

Let $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ be an arbitrary linear discriminant function. If the d-dimensional random vector $\boldsymbol{X}$ is normally distributed, $f(\boldsymbol{X})$ is also normally distributed. Therefore, the error in the projected 1-dimensional space is determined by the first and second-order moments of $f(\boldsymbol{X})$

$$\eta_i = \mathrm{E}\{f(\boldsymbol{X}) \,|\, \boldsymbol{X} \in \mathcal{C}_i\} = \boldsymbol{w}^T \boldsymbol{m}_i + b \tag{8}$$

$$\sigma_i^2 = \mathrm{Var}\{f(\boldsymbol{X}) \,|\, \boldsymbol{X} \in \mathcal{C}_i\} = \boldsymbol{w}^T\, \Sigma_i\, \boldsymbol{w}, \tag{9}$$

where $\boldsymbol{m}_i$ and Σ_i are the conditional expected vectors and covariance matrices of $\boldsymbol{X}$. Even if $\boldsymbol{X}$ is not normal, $f(\boldsymbol{X})$ can be close to normal for large d since it is the summation of d terms and the central limit theorem may come into effect [5]. Then second-order criteria $\Psi(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2)$ appear as reasonable measures of separability in the projected space. In addition, it has been shown in [13] that there exists a broad class of second-order criteria that guaranty the best receiver in the Bayes sense for general nonlinear detector design.

Let Ψ be any second-order criterion. The optimal statistic $f(\boldsymbol{X})$ is obtained by equating to zero the partial derivatives of Ψ with respect to $\boldsymbol{w}$ and b:

$$\begin{cases} \frac{\partial \Psi}{\partial \boldsymbol{w}} = & \frac{\partial \Psi}{\partial \sigma_0^2} \frac{\partial \sigma_0^2}{\partial \boldsymbol{w}} + \frac{\partial \Psi}{\partial \sigma_1^2} \frac{\partial \sigma_1^2}{\partial \boldsymbol{w}} + \frac{\partial \Psi}{\partial \eta_0} \frac{\partial \eta_0}{\partial \boldsymbol{w}} + \frac{\partial \Psi}{\partial \eta_1} \frac{\partial \eta_1}{\partial \boldsymbol{w}} = 0 \\ \frac{\partial \Psi}{\partial b} = & \frac{\partial \Psi}{\partial \sigma_0^2} \frac{\partial \sigma_0^2}{\partial b} + \frac{\partial \Psi}{\partial \sigma_1^2} \frac{\partial \sigma_1^2}{\partial b} + \frac{\partial \Psi}{\partial \eta_0} \frac{\partial \eta_0}{\partial b} + \frac{\partial \Psi}{\partial \eta_1} \frac{\partial \eta_1}{\partial b} = 0. \end{cases}$$

Solving this system with

$$\frac{\partial \sigma_i^2}{\partial \boldsymbol{w}} = 2\, \Sigma_i\, \boldsymbol{w} \quad \frac{\partial \eta_i}{\partial \boldsymbol{w}} = \boldsymbol{m}_i \quad \frac{\partial \sigma_i^2}{\partial b} = 0 \quad \frac{\partial \eta_i}{\partial b} = 1 \tag{10}$$

leads directly to the following result [5, 12].

Proposition 1. *Let $f(\boldsymbol{x}) \triangleq \boldsymbol{w}^T \boldsymbol{x} + b$ be any linear decision statistic. The optimum projection vector $\boldsymbol{w}$ under which the maximum value of any given second-order criterion Ψ is reached satisfies*

$$[\rho \Sigma_0 + (1-\rho)\Sigma_1]\, \boldsymbol{w}_\rho = [\boldsymbol{m}_1 - \boldsymbol{m}_0], \tag{11}$$

where $\boldsymbol{m}_i$ and Σ_i are the conditional expected vectors and covariance matrices of $\boldsymbol{X}$. The parameter ρ depends on the criterion Ψ according to

$$\rho = \frac{\frac{\partial \Psi}{\partial \sigma_0^2}}{\frac{\partial \Psi}{\partial \sigma_0^2} + \frac{\partial \Psi}{\partial \sigma_1^2}}. \tag{12}$$

Hence, the optimum projection direction $\boldsymbol{w}_\rho$ given by (11) depends on Ψ via a single parameter $\rho \in]-\infty, +\infty[$, which can be adjusted to pick the receiver $f(\boldsymbol{x})$ that has the best performance. This approach thus leads to the optimum receiver in the sense of the best second-order criterion Ψ, without setting it up. Note that $\rho \in [0, 1]$ if, and only if, $\partial\Psi/\partial\sigma_0^2$ and $\partial\Psi/\partial\sigma_1^2$ are of the same sign. This condition means that Ψ varies in the same way with σ_0^2 and σ_1^2, which is a desirable but non-mandatory requirement for design criteria [1]. This interval contains several well known second-order criteria such as Fisher criterion, generalized signal-to-noise ratio and mean square error.

3.2. Nonlinear second-order discriminant

Nonlinear kernel second-order discriminant (KSOD) can be obtained by using (11) in the feature space $\mathcal{F}$. According to Proposition 1, the function $f(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) + b$ operating in $\mathcal{F}$ is optimum in the sense of any given second-order criterion Ψ if it satisfies

$$\left[\rho \Sigma_0^\Phi + (1-\rho)\Sigma_1^\Phi\right] \boldsymbol{w} = [\boldsymbol{m}_1^\Phi - \boldsymbol{m}_0^\Phi], \tag{13}$$

where $\boldsymbol{m}_i^\Phi$ and Σ_i^Φ denote the conditional expected vectors and covariance matrices of $\Phi(\boldsymbol{X})$, respectively. Using training data, these moments can be estimated as follows:

$$\boldsymbol{m}_i^\Phi = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{C}_i} \Phi(\boldsymbol{x}) \tag{14}$$

$$\Sigma_i^\Phi = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{C}_i} \Phi(\boldsymbol{x})\, \Phi^T(\boldsymbol{x}) - (\boldsymbol{m}_i^\Phi)\, (\boldsymbol{m}_i^\Phi)^T, \tag{15}$$

where n_i is the number of samples from class $\mathcal{C}_i$ in the training set. When $\mathcal{F}$ is a very high-dimensional space, (13) may be difficult

to solve except if the Mercer trick is used. From the theory of reproducing kernels [15], we know that any solution $\boldsymbol{w} \in \mathcal{F}$ must lie in the span of all training samples in $\mathcal{F}$. Therefore $\boldsymbol{w}$ can be written as follows:

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha(i)\, \Phi(\boldsymbol{x}_i) = \mathbf{Q}\, \boldsymbol{\alpha} \tag{16}$$

where $\mathbf{Q}$ denotes the matrix $[\Phi(\boldsymbol{x}_1) \cdots \Phi(\boldsymbol{x}_n)]$, and the $\alpha(i)$'s are the dual parameters. Multiplying (13) by $\mathbf{Q}^T$ and using (16) yields

$$\left[\rho\, \mathbf{Q}^T \Sigma_0^{\Phi} \mathbf{Q} + (1-\rho)\, \mathbf{Q}^T \Sigma_1^{\Phi} \mathbf{Q}\right] \boldsymbol{\alpha} = \mathbf{Q}^T [\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_0^{\Phi}]. \tag{17}$$

By using a kernel which verifies (4), this expression can be reformulated as

$$\mathbf{N}_\rho\, \boldsymbol{\alpha} = \boldsymbol{m}, \tag{18}$$

where $\boldsymbol{\alpha}$ has to be determined and $\mathbf{N}_\rho$ is a n by n matrix which is given by

$$\mathbf{N}_\rho = \left[\frac{\rho}{n_0}\, \mathbf{K}_0(\mathbf{I} - \mathbf{1}_{n_0})\mathbf{K}_0^T + \frac{1-\rho}{n_1}\, \mathbf{K}_1(\mathbf{I} - \mathbf{1}_{n_1})\mathbf{K}_1^T\right]. \tag{19}$$

In the above expression, $\mathbf{K}_i$ is a n by n_i matrix with elements

$$\mathbf{K}_i(p,q) = \kappa(\boldsymbol{x}_p, \boldsymbol{x}_q), \tag{20}$$

for all $\boldsymbol{x}_p \in (\mathcal{C}_0 \cup \mathcal{C}_1)$ and $\boldsymbol{x}_q \in \mathcal{C}_i$. $\mathbf{I}$ is the identity matrix and $\mathbf{1}_{n_i}$ is the matrix with all elements set to $\frac{1}{n_i}$. The components of $\boldsymbol{m}$ in (18) are defined as

$$m(j) = \frac{1}{n_1} \sum_{\boldsymbol{x} \in \mathcal{C}_1} k(\boldsymbol{x}, \boldsymbol{x}_j) - \frac{1}{n_0} \sum_{\boldsymbol{x} \in \mathcal{C}_0} k(\boldsymbol{x}, \boldsymbol{x}_j). \tag{21}$$

To determine the projection of any new sample $\boldsymbol{x}$ onto $\boldsymbol{w}$, we have to calculate the n-dimensional vector $\boldsymbol{\alpha}$ from (18). Eq. (16) yields:

$$\Phi(\boldsymbol{x})^T\, \boldsymbol{w} = \sum_{i=1}^{n} \alpha(i)\, k(\boldsymbol{x}_i, \boldsymbol{x}). \tag{22}$$

Finally one can use different strategies to determine the bias b, as shown in [10]. The method presented above is called *kernel second-order discriminant* since it leads to the optimum nonlinear receiver in the sense of the best second-order criterion without setting it up. Obviously, classifiers obtained with KSOD perform better than or equal to those resulting from the KFD method developed by Mika *et al.* in [10].

4. COMPLEXITY CONTROL OF KSOD

As the number of parameters in (22) increases, overfitting problems may arise with devastating effects on the generalization performance [14]. A technique proposed by LeCun *et al.*, called *optimal brain damage (OBD)*, has been widely used to reduce the size of neural networks by selectively pruning weights. We present here a similar strategy to be applied to the dual vector $\boldsymbol{\alpha}$ provided by (18).

The goal is to find the best components of the dual vector $\boldsymbol{\alpha}$ for pruning. It is judicious to set to zero those involving the smallest variations of the squared error $\mathcal{E}_\rho$ defined from (18) by

$$\mathcal{E}_\rho = \|\mathbf{N}_\rho\, \boldsymbol{\alpha} - \boldsymbol{m}\|^2. \tag{23}$$

To facilitate our decision about the parameters to be pruned, the process is performed on a basis of normalized eigenvectors of $\mathbf{N}_\rho$, where $\mathcal{E}_\rho$ can be rewritten as

$$\mathcal{E}_\rho = \sum_{i=1}^{n} [\lambda_\rho(i)\, \tilde{\alpha}(i) - \tilde{m}(i)]^2, \tag{24}$$

with $\tilde{\boldsymbol{\alpha}} = \mathbf{P}_\rho^T\, \boldsymbol{\alpha}$ and $\tilde{\boldsymbol{m}} = \mathbf{P}_\rho^T\, \boldsymbol{m}$. The i^{th} column of the matrix $\mathbf{P}_\rho$ is the eigenvector corresponding to the i^{th} eigenvalue $\lambda_\rho(i)$ of $\mathbf{N}_\rho$. A perturbation $\delta\tilde{\boldsymbol{\alpha}}$ modifies the objective function $\mathcal{E}_\rho$ by the quantity

$$\delta\mathcal{E}_\rho = \sum_i \frac{\partial \mathcal{E}_\rho}{\partial \tilde{\alpha}(i)}\, \delta\tilde{\alpha}(i) + \frac{1}{2} \sum_i \frac{\partial^2 \mathcal{E}_\rho}{\partial \tilde{\alpha}(i)^2}\, \delta\tilde{\alpha}(i)^2 + \frac{1}{2} \sum_{i \neq j} \frac{\partial^2 \mathcal{E}_\rho}{\partial \tilde{\alpha}(i)\, \partial \tilde{\alpha}(j)}\, \delta\tilde{\alpha}(i)\, \delta\tilde{\alpha}(j) + O(\|\delta\tilde{\boldsymbol{\alpha}}\|^2), \tag{25}$$

with $\tilde{\alpha}(i)$ the i^{th} component of $\tilde{\boldsymbol{\alpha}}$. If $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}_\rho \triangleq \mathbf{P}_\rho^T\, \boldsymbol{\alpha}_\rho$, where $\boldsymbol{\alpha}_\rho$ satisfies (18) given ρ, replacing $\mathcal{E}_\rho$ in (25) yields

$$\delta\mathcal{E}_\rho = \sum_{i=1}^{n} [\lambda_\rho(i)\, \delta\tilde{\alpha}(i)]^2. \tag{26}$$

Pruning the i^{th} component of $\tilde{\boldsymbol{\alpha}}_\rho$ then increases $\mathcal{E}_\rho$ by

$$\delta\mathcal{E}_\rho(i) = [\lambda_\rho(i)\, \tilde{\alpha}_\rho(i)]^2 \tag{27}$$

since $\delta\tilde{\alpha}(i) = \tilde{\alpha}_\rho(i)$ when $\tilde{\alpha}_\rho(i)$ is set to zero. Therefore the components of $\tilde{\boldsymbol{\alpha}}_\rho$ associated with the smallest variations of $\mathcal{E}_\rho$ given by (27) are good candidates for pruning.

We shall now show that $\delta\mathcal{E}_\rho(i)$ does not depend on ρ. This avoids time consuming OBD-based procedures for each ρ. Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric non-negative matrices. Under weak conditions on $\mathbf{A}$ and $\mathbf{B}$, there exists a nonsingular matrix $\mathbf{P}$ such that both $\mathbf{P}^T\, \mathbf{A}\, \mathbf{P}$ and $\mathbf{P}^T\, \mathbf{B}\, \mathbf{P}$ are diagonal [7]. We can apply this result to $\mathbf{Q}^T\, \Sigma_0^{\Phi}\, \mathbf{Q}$ and $\mathbf{Q}^T\, \Sigma_1^{\Phi}\, \mathbf{Q}$ and then conclude that the matrix $\mathbf{P}_\rho$, whose columns are eigenvectors of $\mathbf{N}_\rho$, is independent of ρ. It directly follows that $\tilde{\boldsymbol{m}} = \mathbf{P}^T\, \boldsymbol{m}$ does not depend on ρ. Since $\boldsymbol{\alpha}_\rho$ satisfies (18), we have $\lambda_\rho(i)\, \tilde{\alpha}(i) = \tilde{m}(i)$ and

$$\delta\mathcal{E}_\rho(i) = \tilde{m}(i)^2 \tag{28}$$

if $\lambda_\rho(i) \neq 0$. This shows that the variations of $\mathcal{E}_\rho$ do not depend on ρ. In addition, note from relation (27) that $\delta\mathcal{E}_\rho(i) = 0$ if $\lambda_\rho(i) = 0$, which means that the associated component $\tilde{\alpha}_\rho(i)$ can be directly set to zero. These results then lead to an efficient algorithm for controlling the complexity of KSOD receivers.

5. EXPERIMENTATIONS

The two spirals problem [8], which is a difficult benchmark classification problem, has been chosen as the first experiment to test the efficiency of the OBD-based method applied to KSOD. Figure 1 shows the decision function of a SVM classifier (dotted line) based on an exponential radial basis function (ERBF) kernel having a width equal to 1. Note that 100% of the 194 training data were selected as support vectors by the algorithm. A similar decision function was obtained (solid line) with the KSOD method, using only the 4 most significant components $\tilde{\alpha}(i)$ selected by the

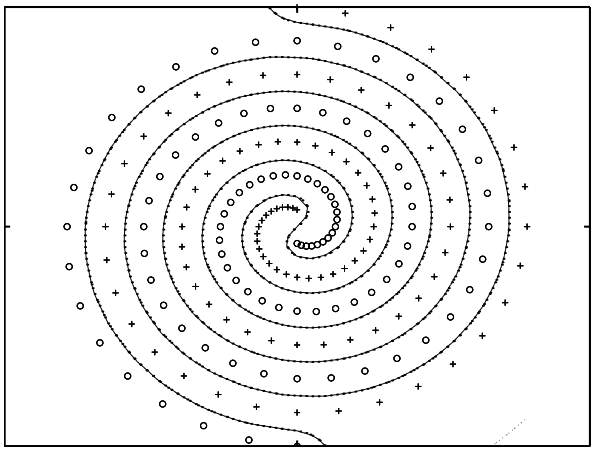

Fig. 1. Comparison of KSOD (solid) with SVM (dotted). The KSOD decision function was obtained with only 4 components of $\tilde{\boldsymbol{\alpha}}$ out of a total of 194 whereas SVM needed 194 support vectors.

OBD procedure out of a total of 194. This experiment then illustrates the ability of our approach to provide sparse solutions.

To compare our method to KFD and SVM, 10 experiments were conducted on artificial and real world data downloaded from `http://www.first.gmd.de/~raetsch`. For each of the 10 problems, Table 1 shows the average test error over 40 runs on 400 training samples and 8000 test samples chosen arbitrarily from a mixture of the available data sets. The kernel function was selected as the RBF having a width equal to 1. The results presented in Table 1 clearly show that the KSOD and the KSOD-OBD methods are more efficient in most cases than SVM and KFD. In addition the number of support vectors needed by the SVM method is very large in comparison with the number of components $\tilde{\alpha}(i)$ selected by the OBD method. Such very economical solutions generally have good generalization performance and produce short testing times.

Data set	KFD	SVM	KSOD	KSOD-OBD
Banana	10.60	10.43 (132)	10.59	**10.37** (35)
Thyroid	0.39	0.33 (156)	0.25	**0.23** (75)
B. cancer	8.14	7.10 (364)	6.70	**6.60** (75)
Diabetes	17.79	17.68 (308)	17.39	**17.11** (150)
German	21.36	21.06 (400)	20.96	**20.90** (200)
Heart	4.44	4.52 (388)	4.41	**4.24** (100)
Solar	32.42	32.73 (364)	31.61	**31.04** (40)
Waveform	11.14	**11.07** (400)	11.14	11.14 (400)
Ringnorm	1.53	1.52 (396)	1.53	**1.50** (20)
Titanic	28.88	28.88 (400)	28.55	**27.72** (15)

Table 1. Comparison of the mean error rates obtained with KFD, SVM, KSOD and KSOD-OBD. This table also gives the number of components $\tilde{\alpha}(i)$ that were selected by the OBD algorithm and the number of support vectors.

6. CONCLUSION

In this paper, KSOD and KSOD-OBD methods were presented. These methods have been applied to an extensive number of artificial and real world data sets. The results obtained suggest that the KSOD-OBD method performs often better than KFD and SVM. Furthermore, very sparse descriptions of the classification functions were obtained for most of the 10 experiments, which imply a very short testing time for new data and best generalization performance.

7. REFERENCES

[1] F. Abdallah, C. Richard and R. Lengellé. "On virtues and vices of second-order measures of quality for binary classification," in *Proc. ANNIE Conference*, November 10-13, 2002, Saint Louis, USA.

[2] C. Burges. "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[3] C. Cortes and V. Vapnik. "Support vector networks," *Machine Learning*, vol. 20, pp. 1–25, 2002.

[4] L. Devroye, L. Gyorfi and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag, 1996.

[5] K. Fukunaga. *Statistical Pattern Recognition*, San Diego: Academic Press, 1990.

[6] W. A. Gardner. "A unifying view of second-order measures of quality for signal classification," *IEEE Transactions on Communications*, vol. 28, no. 6, pp. 807–816, 1980.

[7] G. H. Golub and C. F. Van Loan. *Matrix Computations*. London: The Johns Hopkins University Press, 1993.

[8] K. J. Lang and M. J. Witbrock. "Learning to tell two spirals apart," in *Proc. Connectionnist Summer Schools*, Morgan Kaufmann, 1988.

[9] Y. LeCun, J. S. Denker and S. A. Solla. "Optimal brain damage," in *Proc. Advances in Neural Information Processing Systems*, vol. 2, pp. 598–605, 1990.

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

[11] K. R. Müller, S. Mika, Rätsch, K. Tsuda and B. Schölkopf. "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, no. 2, pp. 181-201, May 2001.

[12] C. Richard and R. Lengellé. "Data-driven design and complexity control of time-frequency detectors," *Signal Processing*, vol. 77, no. 1, pp. 37–48, 1999.

[13] 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.

[14] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

[15] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, 1988.

[16] B. Schölkopf, C. Burges and A. Smola. *Advances in Kernel Methods-Support Vector Learning*, MIT Press, 1999.

[17] V. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.