TOWARDS A UNIFICATION OF INFORMATION THEORETIC LEARNING AND KERNEL METHODS

Robert Jenssen[1]*, Deniz Erdogmus[2], Jose C. Principe[2] and Torbjørn Eltoft[1]

[1]Department of Physics, University of Tromsø, Norway
[2]Computational NeuroEngineering Laboratory, University of Florida, USA

Abstract. In this paper, we discuss an intriguing relationship between information theoretic learning (ITL), based on Parzen window density estimation, and kernel-based learning algorithms. We show that some of the widely used ITL cost functions, when estimated by the Parzen method, can be expressed in terms of inner products in a kernel feature space defined by a Mercer kernel, where the Mercer kernel, in fact, is the Parzen window. This link gives a theoretical criterion for the selection of the Mercer kernel, based on density estimation. Also, we show that the support vector machine (SVM), as an example of a well-known kernel-based learning algorithm, can be examined in an information theoretic framework, using weighted Parzen windows for density estimation.

INTRODUCTION

During the last decade, research on Mercer kernel-based learning algorithms, predominantly the support vector machine (SVM) theory [1, 2], but also methods like kernel Fisher discriminant (KFD) analysis [3] and kernel principal component analysis (KPCA) [4], have flourished. These methods have proven to achieve excellent results on a number of applications, ranging from e.g. pattern and object recognition [5], time series prediction [6] to DNA and protein analysis [7]. One problem with the kernel methods though, is that it is not clear how to choose the actual kernel function. Often the Gaussian radial-basis-function (RBF) is used, in which case it still remains an open question exactly how to choose the width of the RBF kernel.

Independently of the activity on kernel methods, another line of research has recently emerged that is coined *information theoretic learning* [8]. ITL addresses the issue of extracting information directly from data in a nonparametric manner. The learning-from-examples scenario starts with a data

*Corresponding author. Phone: (+47) 776 46493. Email: robertj@phys.uit.no

0-7803-8608-6/04/$20.00 ©2004 IEEE

set that globally conveys information about a real-world event, and the goal is to capture the information in the parameters of a learning machine. The backbone of ITL has been the utilization of Renyi's measure of entropy as a cost function for learning, in addition to approximations to the Kullback-Leibler probability density divergence. These quantities lends themself nicely to non-parametric estimation via Parzen windowing for density estimation. The ITL framework has been successfully applied in a variety of learning scenarios, such as object recognition [8], time series prediction [9], blind deconvolution [10], blind source separation [11] and clustering [12].

The purpose of this paper is to provide a first step towards unifying the two aforementioned frameworks for learning, by demonstrating an intriguing duality between the Parzen and the Mercer kernels. We show that the most widely used ITL cost functions, when estimated by the Parzen method, can be expressed in terms of inner products in a kernel feature space defined by a Mercer kernel, which in fact is the kernel in the Parzen window method. Having illustrated this fact by several examples, we turn to the most famous kernel learning machine, the SVM, and show that it can be expressed in terms of one of the ITL density divergence measures, when the pdfs are estimated by a weighted Parzen window estimator. Based on the discussion we give in this paper, we conjecture that the kernel-based learning algorithms that are expressed in terms of inner products in the kernel feature space, are in fact learning by implicitly utilizing non-parametric estimates of probability densities in the input space. This view gives a theoretical criterion for selecting the Mercer kernel to be used in the kernel-based methods, namely the kernel that would lead to a relatively accurate estimate, if used as the Parzen window in density estimation.

The organization of this paper is as follows. In section 2 we review the basic theory of nonlinear kernel feature spaces. In section 3 we discuss some of the cost functions utilized in information theoretic learning, and show how they can be translated into quantities defined in the Hilbert feature space via Parzen windowing. Thereafter, in section 4, we show how the SVM classifier cost function can be expressed in terms of an information theoretic density divergence via weighted Parzen window estimation. Finally, in section 5 we make our concluding remarks.

KERNEL FEATURE SPACES

Kernel-based learning algorithms make use of the following idea: via a nonlinear mapping

$$\begin{aligned} \mathbf{\Phi} : R^d &\rightarrow \mathcal{F} \\ \mathbf{x} &\rightarrow \mathbf{\Phi}(\mathbf{x}) \end{aligned} \tag{1}$$

the data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in R^d$ is mapped into a potentially much higher dimensional feature space $\mathcal{F}$. For a given learning problem one now considers the same algorithm in $\mathcal{F}$ instead of in R^d, that is, one works with $\mathbf{\Phi}(\mathbf{x}_1), \ldots, \mathbf{\Phi}(\mathbf{x}_N) \in \mathcal{F}$.

This mapping is of particular interest in cases where the learning algorithm is expressed only in terms of inner products. The reason is that one can use a highly effective trick for computing inner products in the feature space using *kernel functions*, without even knowing the exact mapping $\mathbf{\Phi}$. This can be advantageous since we do not have to execute the learning algorithm in a very high dimensional space, which can cause intractable problems.

Consider a symmetric kernel function $k(\mathbf{x}, \mathbf{y})$. If $k : \mathcal{C} \times \mathcal{C} \rightarrow R$ is a continuous kernel of a positive integral operator in a Hilbert space $L_2(\mathcal{C})$ on a compact set $\mathcal{C} \in R^d$, i.e.

$$\forall f \in L_2(\mathcal{C}) : \int_{\mathcal{C}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0, \tag{2}$$

then there exists a space $\mathcal{F}$ and a mapping $\mathbf{\Phi} : R^d \rightarrow \mathcal{F}$, such that by Mercer's theorem [13]

$$k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{y}) \rangle = \sum_{i=1}^{N_{\mathcal{F}}} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y}), \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product, the ψ_i's are the eigenfunctions of the kernel and $N_{\mathcal{F}} \leq \infty$ [6, 1]. In this case

$$\mathbf{\Phi}(\mathbf{x}) = [\sqrt{\lambda_1} \psi_1(\mathbf{x}), \sqrt{\lambda_2} \psi_2(\mathbf{x}), \ldots]^T, \tag{4}$$

can potentially be realized.

The most widely used Mercer kernel is the radial-basis-function (RBF)

$$k(\mathbf{x}, \mathbf{y}) = \exp \left\{ - \frac{||\mathbf{x} - \mathbf{y}||^2}{2\sigma^2} \right\}. \tag{5}$$

A RBF kernel function corresponds to an infinite-dimensional Hilbert feature space, since the RBF has an infinite number of eigenfunctions.

ITL COST FUNCTIONS IN THE KERNEL SPACE

In this section, we examine some of the most widely used cost functions in information theoretic learning, and show how they can be estimated directly from data via the Parzen window method, utilizing the convolution theorem for Gaussians. Most importantly, we also show that these cost functions can in fact be expressed in terms of inner products in a Hilbert feature space defined by a Mercer kernel, where the Mercer kernel is identical to the window function used in Parzen density estimation.

In this section, probability density functions (pdfs) are estimated by the well known Parzen window method [14]. Let $\hat{f}(\mathbf{x})$ be an estimate of $f(\mathbf{x})$.

Then a *non-parametric* asymptotically unbiased and consistent estimate of f can be defined as [14]

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} W(\mathbf{x}, \mathbf{x}_i), \tag{6}$$

where W is the Parzen window, or kernel. The Parzen window must integrate to one. It is often chosen to be the Gaussian kernel

$$W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{||\mathbf{x} - \mathbf{x}_i||^2}{2\sigma^2} \right\}. \tag{7}$$

In [8], a principled approach of designing practical information theoretic criteria using Renyi's entropy of order two (quadratic entropy) was proposed. Renyi's quadratic entropy can be easily integrated with the Parzen window estimator, hence providing a means to estimate the entropy directly from the data set. Renyi's entropy is given by

$$H_2(\mathbf{x}) = -\log \int f^2(\mathbf{x}) d\mathbf{x}. \tag{8}$$

Since the logarithm is a monotonic function, the quantity of interest is $V(\mathbf{x}) = \int f^2(\mathbf{x}) d\mathbf{x}$, which was called the *information potential* [8], because of an analogy to a potential energy field. We have that

$$\begin{aligned} V(\mathbf{x}) &= \int \frac{1}{N} \sum_{i=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i) \frac{1}{N} \sum_{j=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j) d\mathbf{x} \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i) W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j) d\mathbf{x} \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \tag{9}$$

where in the last step the convolution theorem for Gaussians has been employed [8].

The key point of this paper, is to note is that $W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_j)$ is a Gaussian RBF kernel function, and hence it is also a *kernel function that satisfies Mercer's theorem.* Hence

$$W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle, \tag{10}$$

where $\mathbf{\Phi}(\mathbf{x}_i) = [\sqrt{\lambda_1}\psi_1(\mathbf{x}_i), \sqrt{\lambda_2}\psi_2(\mathbf{x}_i), \ldots]^T$, $i = 1, \ldots, N$. Now we rewrite

(9) as follows

$$\begin{aligned} V(\mathbf{x}) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j)\rangle \\ &= \left\langle \frac{1}{N}\sum_{i=1}^{N}\mathbf{\Phi}(\mathbf{x}_i), \frac{1}{N}\sum_{j=1}^{N}\mathbf{\Phi}(\mathbf{x}_j)\right\rangle \\ &= \langle \mathbf{m}^{\Phi}, \mathbf{m}^{\Phi}\rangle \\ &= ||\mathbf{m}^{\Phi}||^2, \end{aligned} \tag{11}$$

where $\mathbf{m}^{\Phi}$ is the mean vector of the $\mathbf{\Phi}$-transformed data. That is, the information potential turns out to be equal to the sum of inner products between all pairs of data points in the Hilbert kernel space, which can be expressed as the squared norm of the mean vector of the data in that space. This connection was previously pointed out by Girolami [15] in a study relating orthogonal series density estimates to KPCA. It is also interesting to note that (11) can be interpreted as the 2-norm of the probability mass function $P = (p_1, p_2, \ldots, p_N)$, when P is considered a point in a N-dimensional space [8].

In [8], two approximations to the Kullback-Leibler density divergence were proposed, that, like Renyi's entropy, can be easily integrated with a Parzen window density estimator. If the pdfs under consideration are the joint density and the product of marginals, these divergence measures approximate the Kullback-Leibler mutual information, as a measure of independence between random variables.

First we consider the Cauchy-Schwarz (CS) pdf divergence. It was named so because it was obtained by replacing inner products between vectors in the Cauchy-Schwarz inequality, by inner products between pdfs. It is defined as [8]

$$D_{CS}(p,q) = -\log \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p^2(\mathbf{x})d\mathbf{x}\int q^2(\mathbf{x})d\mathbf{x}}} \geq 0. \tag{12}$$

It can be seen that D_{CS} is zero iff the two densities are equal, and goes to infinity as the overlap between the two pdfs goes to zero. Again, since the logarithm is a monotonic function, the quantify of interest is the quantity in the argument of the log in (12). This quantity was called the Information Cut (IC) in [16], because it was shown that it is closely related to the graph theoretic notion of a cut, which is a measure of the cost of partitioning a graph into two pieces. Again, we estimate the two pdfs by the Parzen window method

$$\hat{p}(\mathbf{x}) = \frac{1}{N_1}\sum_{i=1}^{N_1} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i), \quad \hat{q}(\mathbf{x}) = \frac{1}{N_2}\sum_{j=1}^{N_2} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j). \tag{13}$$

By a similar calculation as above, the Information Cut can be expressed as;

$$IC = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} W_{2\sigma^2}(\mathbf{x}_i,\mathbf{x}_j)}{\sqrt{\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1} W_{2\sigma^2}(\mathbf{x}_i,\mathbf{x}_{i'}) \; \sum_{j=1}^{N_2}\sum_{j'=1}^{N_2} W_{2\sigma^2}(\mathbf{x}_j,\mathbf{x}_{j'})}}. \tag{14}$$

Equation (14) provides a means for estimating the divergence between two continuous densities directly from a set of data samples. In analogy to (11), this expression can also be expressed in terms of inner products in the Hilbert feature space, since $W_{2\sigma^2}(\mathbf{x}_i,\mathbf{x}_j) = k(\mathbf{x}_i,\mathbf{x}_j)$ is a Mercer kernel. When we carry out an exact same type of calculation as in (11), we obtain

$$IC = \frac{\langle \mathbf{m}_p^{\Phi}, \mathbf{m}_q^{\Phi} \rangle}{||\mathbf{m}_p^{\Phi}||^2 ||\mathbf{m}_q^{\Phi}||^2}, \tag{15}$$

where $\mathbf{m}_p^{\Phi}$ is the mean vector in the Hilbert feature space with respect to the Φ-mapped data drawn from $p(\mathbf{x})$, and $\mathbf{m}_q^{\Phi}$ is the feature space mean vector with respect to the Φ-mapped data drawn from $q(\mathbf{x})$. Hence, quite interestingly, it turns out that the CS information theoretic pdf divergence measure has a dual interpretation as a measure of the cosine of the angle between the unit norm mean vectors in the kernel Hilbert feature space.

The CS divergence measure has recently been utilized as a cost function for clustering by the current authors [12], where the optimization was carried out using the Lagrange multiplier formalism.

Finally, a second pdf divergence measure for ITL was also proposed in [8] based on the Euclidean difference of vectors inequality. The Euclidean distance (ED) is defined as [8]

$$\begin{aligned} D_{ED}(p,q) &= \int \{p(\mathbf{x}) - q(\mathbf{x})\}^2 \, d\mathbf{x} \\ &= \int p^2(\mathbf{x}) d\mathbf{x} + \int q^2(\mathbf{x}) d\mathbf{x} - 2\int p(\mathbf{x})q(\mathbf{x}) d\mathbf{x} \geq 0. \end{aligned} \tag{16}$$

Performing a similar analysis as above by estimating the densities by the Parzen method, the D_{ED} can be expressed as follows

$$\begin{aligned} D_{ED}(p,q) &= \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1} k(\mathbf{x}_i,\mathbf{x}_{i'}) + \frac{1}{N_2^2}\sum_{j=1}^{N_2}\sum_{j'=1}^{N_2} k(\mathbf{x}_j,\mathbf{x}_{j'}) \\ &- 2\frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} k(\mathbf{x}_i,\mathbf{x}_j) \\ &= ||\mathbf{m}_p^{\Phi} - \mathbf{m}_q^{\Phi}||^2. \end{aligned} \tag{17}$$

Hence, the ED information theoretic pdf divergence measure can be seen to also have a geometric interpretation in the Hilbert feature space. The D_{ED} measures the square of the norm of the difference vector between the two means $\mathbf{m}_p^{\Phi}$ and $\mathbf{m}_q^{\Phi}$. If the norm of the difference vector goes to zero, the corresponding continuous pdfs in the input space are maximally aligned with each other, that is, having a maximum amount of overlap.

Our discussion in this section clearly shows that each of the ITL cost functions has a dual representation in the Hilbert kernel feature space. As such, the ITL learning algorithms are also Mercer kernel-based learning algorithms. Since the ITL algorithms are by definition linked to pdf estimation, the kernel versions of these algorithms are fundamentally linked to pdf estimation too. This gives ***a theoretical criterion for selecting the Mercer kernel***, namely the Mercer kernel that would lead to a relatively accurate density estimate if used as the Parzen window in density estimation.

We can also draw the following conclusion: Whenever we encounter an expression like

$$\sum_i \sum_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{18}$$

where the RBF kernel function k satisfies Mercer's conditions, it has a dual expression as an integral over a product of pdfs, i.e. $\int f^2(\mathbf{x})d\mathbf{x}$, where the density is estimated by the Parzen window method (up to a constant, if the kernel does not integrate to one).

In this exposition, we have estimated the information theoretic cost functions using Parzen window estimators with uniform weighting on each window function. In the next section, we briefly examine the non-linear SVM classifier, the arguably most well-known Mercer kernel-based learning algorithm. We show that it can in fact be expressed in terms of the ED information theoretic pdf divergence measure that we discussed above, only that the pdfs are estimated by a weighted Parzen window estimator, instead of a uniform, as a consequence of the maximum margin metric that constitutes the basis of the SVM theory.

THE SVM AS AN INFORMATION THEORETIC COST FUNCTION

In this section we assume that the SVM theory is familiar to the reader. We refer to [1, 2] for details. We are given the training set $\{\mathbf{x}_i, d_i\}$, $i = 1, \ldots, N$, $d_i \in \{-1, 1\}$, $\mathbf{x}_i \in R^d$. The task is to train a SVM classifier, such that it creates a maximum margin linear classifier in the kernel feature space. The problem can be formulated in the Lagrange formalism by introducing positive Lagrange multipliers α_i, $i = 1, \ldots, N$. After expressing the problem by the Wolfe dual Lagrange function, the function to be maximized is the following

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{19}$$

subject to the constraints

$$\sum_{i=1}^{N} \alpha_i d_i = 0, \tag{20}$$

$$\alpha_i \geq 0, \quad \forall i. \tag{21}$$

The hyperplane weight vector in the kernel feature space can be shown to be given by [2]

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \mathbf{\Phi}(\mathbf{x}_i). \tag{22}$$

The maximum margin metric, which is the basis for the SVM theory, specifies the form of the Karush-Kuhn-Tucker (KKT) conditions [2], which have to be satisfied at the solution (22). These conditions imply that only those α_i which correspond to a $\mathbf{\Phi}(\mathbf{x}_i)$ which lies on the margin of the hyperplane, will have a value other than zero. This sparseness condition in the kernel feature space, which follows from the maximum margin metric, is crucial for the superior generalization ability of the SVM classifier.

Now, we rewrite (19). Associate the Lagrange multipliers α_i with the first class, and α_j^* with the second class. Note that $\sum_{i=1}^{N_1} \alpha_i = \sum_{j=1}^{N_2} \alpha_j^* = A$ by (20). Now, (19) can be rewritten as follows

$$\begin{aligned} L_D &= 2A - \frac{1}{2}\{\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1} \alpha_i \alpha_{i'} k(\mathbf{x}_i, \mathbf{x}_{i'}) \\ &+ \sum_{j=1}^{N_2}\sum_{j'=1}^{N_2} \alpha_j^* \alpha_{j'}^* k(\mathbf{x}_j, \mathbf{x}_{j'}) - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \alpha_i \alpha_j^* k(\mathbf{x}_i, \mathbf{x}_j)\} \\ &= 2A - \frac{A^2}{2}\{\frac{1}{A^2}\sum_{i=1}^{N_1}\sum_{i'=1}^{N_1} \alpha_i \alpha_{i'} k(\mathbf{x}_i, \mathbf{x}_{i'}) \\ &+ \frac{1}{A^2}\sum_{j=1}^{N_2}\sum_{j'=1}^{N_2} \alpha_j^* \alpha_{j'}^* k(\mathbf{x}_j, \mathbf{x}_{j'}) - 2\frac{1}{A^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \alpha_i \alpha_j^* k(\mathbf{x}_i, \mathbf{x}_j)\}. \end{aligned} \tag{23}$$

This expression can in fact be seen to depend on the ITL quantity $D_{ED}(p, q)$, examined in the previous section. To see this, estimate the density $p(\mathbf{x})$ nonparametrically based on the data from the class corresponding to $d_i = 1$, and estimate the density $q(\mathbf{x})$ based on the data for which $d_i = -1$. However, now the densities are estimated using *weighted Parzen estimators*, defined as

$$\hat{p}(\mathbf{x}) = \frac{1}{A}\sum_{i=1}^{N_1} \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{24}$$

$$\hat{q}(\mathbf{x}) = \frac{1}{A}\sum_{j=1}^{N_2} \alpha_j^* k(\mathbf{x}, \mathbf{x}_j). \tag{25}$$

That is, while training a SVM, one effectively adjusts the weighting components on the Parzen windows, such that $D_{ED}(p,q)$ is minimized, while at the same time keeping the weighting components, α_i, α_j^*, from going to zero. Hence, (19), which is to be maximized, can also be written as

$$L_D = 2A - \frac{A^2}{2} D_{ED}(p,q). \tag{26}$$

It is the minimization of $D_{ED}(p,q)$ that guarantees the selection of the support vectors as those points on the boundary. This example shows that the SVM can be related to ITL and non-parametric pdf estimation via weighted Parzen windowing.

CONCLUDING REMARKS

In this paper, our aim was to provide a first step towards unifying the two research areas known as information theoretic learning and kernel-based learning, respectively. We have shown that the most widely used ITL cost functions, when estimated non-parametrically using the Parzen window density estimator, can be expressed in terms of inner products in a Hilbert kernel feature space defined by a Mercer kernel, where the Mercer kernel is in fact the Parzen window. Our discussion reveals an intriguing duality between the Mercer kernel and the Parzen window, which provides a theoretical criterion for the selection of the Mercer kernel, namely as the kernel that would lead to a relatively accurate pdf estimate if used as the Parzen window in density estimation. Moreover, we have argued that the kernel-based methods that can be expressed only in terms of inner products in the Hilbert feature space, likewise have a dual expression as an ITL cost function, dependent on integrals over products of pdfs. This link will be further pursued in our future work.

Acknowledgments

This work was partially supported by grants ECS-0300340 and EIA-0135946. Robert Jenssen would like to thank the Department of Mathematical Sciences at the University of Tromsø, Norway, for granting two research scholarships in order for him to visit the Computational NeuroEngineering Laboratory at the University of Florida, USA, the academic year 2002/2003 and spring 2004, for a total of 13 months.

REFERENCES

[1] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[2] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp. 121–167, 1998.

[3] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller, "Fisher Discriminant Aanalysis with Kernels," in *IEEE Workshop on Neural Networks for Signal Processing*, Piscataway, NJ, USA, 1999, pp. 41–48.

[4] B. Schölkopf, A. J. Smola, and K. R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[5] Y. A. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U. A. Müller, E. Säckinger, P. Y. Simard, and V. N. Vapnik, "Learning Algorithms for Classification: A Comparison on Handwritten Digit Reconstruction," *Neural Networks*, pp. 261–276, 1995.

[6] K. R. Müller, A. J. Smola, G. Rätsch B. Schölkopf, J. Kohlmorgen, and V. N. Vapnik, "Predicting Time Series with Support Vector Machines," in *Artificial Neural Networks - Springer Lecture Notes in Computer Science, W. Gerstner and A. Germond and M. Hasler and J.-D. Nicoud (Eds.), Springer-Verlag*, Berlin, Germany, 1997, vol. 1327, pp. 999–1004.

[7] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K. R. Müller, "Engineering Support Vector Machine Kernels that Recognize Translation Invariant Sites in DNA," *Bioinformatics*, vol. 16, pp. 906–914, 2000.

[8] J. Principe, D. Xu, and J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, S. Haykin (Ed.), John Wiley & Sons, 2000, vol. I, Chapter 7.

[9] D. Erdogmus and J. C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1035–1044, 2002.

[10] M. Lazaro, I. Santamaria, D. Erdogmus, K. E. Hild II, C. Pantaleon, and J. C. Principe, "Stochastic Blind Equalization Based on PDF Fitting using Parzen Estimator," *To appear in IEEE Transactions on Signal Processing*, 2004.

[11] K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind Source Separation using Renyi's Mutual Information," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 174–176, 2001.

[12] R. Jenssen, D. Erdogmus, K. E. Hild II, J. C. Principe, and T. Eltoft, "Efficient Information Theoretic Clustering using Stochastic Approximation," in *Submitted to IEEE International Workshop on Machine Learning for Signal Processing*, Sao Luis, Brazil, 2004.

[13] J. Mercer, "Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations," *Philos. Trans. Roy. Soc. London*, vol. A, pp. 415–446, 1909.

[14] E. Parzen, "On the Estimation of a Probability Density Function and the Mode," *Ann. Math. Stat.*, vol. 32, pp. 1065–1076, 1962.

[15] M. Girolami, "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem," *Neural Computation*, vol. 14, no. 3, pp. 669–688, 2002.

[16] R. Jenssen, J. C. Principe, and T. Eltoft, "Information Cut and Information Forces for Clustering," in *IEEE International Workshop on Neural Networks for Signal Processing*, Toulouse, France, 2003, pp. 459–468.