Model Selection Using a Class of Kernels with an Invariant Metric

Akira Tanaka[1], Masashi Sugiyama[2], Hideyuki Imai[1], Mineichi Kudo[1], and Masaaki Miyakoshi[1]

[1] Division of Computer Science,
Graduate School of Information Science and Technology, Hokkaido University,
Sapporo, 060-0814, Japan,
{takira,imai,mine,miyakosi}@main.eng.hokudai.ac.jp
[2] Department of Computer Science, Tokyo Institute of Technology,
Meguro-ku, Tokyo, 152-8552, Japan,
sugi@cs.titech.ac.jp

Abstract. Learning based on kernel machines is widely known as a powerful tool for various fields of information science such as pattern recognition and regression estimation. The efficacy of the model in kernel machines depends on the distance between the unknown true function and the linear subspace, specified by the training data set, of the reproducing kernel Hilbert space corresponding to an adopted kernel. In this paper, we propose a framework for the model selection of kernel-based learning machines, incorporating a class of kernels with an invariant metric.

1 Introduction

Learning based on kernel machines[1] is widely known as a powerful tool for various fields of information science such as pattern recognition and regression estimation. Many kernel machines, represented by the support vector machines[2] and the kernel ridge regression[3, 4], are proposed. In these methods, kernels are recognized as useful tools to calculate the inner product in high-dimensional feature spaces[3, 4].

On the other hand, according to the theory of reproducing kernel Hilbert spaces[5, 6], the essence of using kernels in learning problems is that the unknown target (classifiers in pattern recognition problems, unknown true functions in regression estimation problems, and so on) belongs to the reproducing kernel Hilbert space corresponding to the adopted kernel. On the basis of this essence, Ogawa formulated a learning problem as an inversion problem of a linear operator from the reproducing kernel Hilbert space corresponding to the adopted kernel onto a certain vector space concerned with the given training data set and constructed a series of learning machines, named "(parametric) projection learning", that gives a good approximation of the orthogonal projector of the unknown true function onto the linear subspace, specified by the given training

data set, of the reproducing kernel Hilbert space corresponding to the adopted kernel[7].

In the field of machine learning based on kernel machines, the model selection, that is, the selection of a kernel (or its parameters) is one of the most important problems. In this paper, we construct a framework of the kernel selection on the basis of the projection-learning-based interpretation of learning problems, incorporating a class of kernels with an invariant metric.

2 Mathematical Preliminaries for The Theory of Reproducing Kernel Hilbert Spaces

In this section, we prepare some mathematical tools concerned with the theory of reproducing kernel Hilbert spaces.

Definition 1 *[5] Let $\mathbf{R}^n$ be an n-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on $\mathcal{D} \subset \mathbf{R}^n$, forming a Hilbert space of real-valued functions. The function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, $(\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D})$ is called a reproducing kernel of $\mathcal{H}$, if*

1. *For every $\tilde{\boldsymbol{x}} \in \mathcal{D}$, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is a function of $\boldsymbol{x}$ belonging to $\mathcal{H}$.*
2. *For every $\tilde{\boldsymbol{x}} \in \mathcal{D}$ and every $f \in \mathcal{H}$,*

$$f(\tilde{\boldsymbol{x}}) = \langle f(\boldsymbol{x}), K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}, \tag{1}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of the Hilbert space $\mathcal{H}$.

The Hilbert space $\mathcal{H}$ that has a reproducing kernel is called a reproducing kernel Hilbert space (RKHS). The reproducing property Eq.(1) enables us to treat a value of a function at a point in $\mathcal{D}$. Note that reproducing kernels are positive definite [5]:

$$\sum_{i,j=1}^{N} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \tag{2}$$

for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ is followed[5]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique[5]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [5].

Next, we introduce the Schatten product [8] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2 *[8] Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. The Schatten product of $g \in \mathcal{H}_2$ and $h \in \mathcal{H}_1$ is defined by*

$$(g \otimes h) f = \langle f, h \rangle_{\mathcal{H}_1} g, \quad f \in \mathcal{H}_1. \tag{3}$$

Note that $(g \otimes h)$ is a linear operator from $\mathcal{H}_1$ onto $\mathcal{H}_2$. It is easy to show that the following relations hold for $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

$$(h \otimes g)^* = (g \otimes h), \quad (h \otimes g)(u \otimes v) = \langle u, g \rangle_{\mathcal{H}_2} (h \otimes v), \tag{4}$$

where the super script * denotes the adjoint operator.

3 Formulation of Learning as Linear Inverse Problems

Let $\{(y_i, \boldsymbol{x}_i)|i = 1, \ldots, \ell\}$ be a given training data set with $y_i \in \mathbf{R},\ \boldsymbol{x}_i \in \mathbf{R}^n$, satisfying

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{5}$$

where f denotes the unknown true function and n_i denotes a zero-mean additive noise. The aim of machine learning is to estimate the unknown function f by using the given training data set and statistical properties of noise.

In this paper, we assume that the unknown function f belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel function K. If $f \in \mathcal{H}_K$, then Eq.(5) is rewritten by

$$y_i = \langle f(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} + n_i, \tag{6}$$

on the basis of the reproducing property of kernels. Let $\boldsymbol{y} = [y_1, \ldots, y_\ell]'$ and $\boldsymbol{n} = [n_1, \ldots, n_\ell]'$ with the super script $'$ denoting the transposed matrix (or vector), then applying the Schatten product to Eq.(6) yields

$$\boldsymbol{y} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\boldsymbol{x}, \boldsymbol{x}_k)]\right) f(\boldsymbol{x}) + \boldsymbol{n}, \tag{7}$$

where $\boldsymbol{e}_k^{(\ell)}$ denotes the k-th vector of the canonical basis of $\mathbf{R}^\ell$. For a convenience of description, we write

$$A_K = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\boldsymbol{x}, \boldsymbol{x_k})]\right). \tag{8}$$

The operator A_K is linear one that maps an element of $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(7) can be written by

$$\boldsymbol{y} = A_K f + \boldsymbol{n}, \tag{9}$$

which represents the relation between the unknown true function f and an output vector $\boldsymbol{y}$. The information of input vectors is integrated in the operator A_K. Therefore, a machine learning problem can be interpreted as an inversion problem of Eq.(9) [7].

Based on the model Eq.(9), a novel learning framework named "(parametric) projection learning" was proposed[7, 9–11]. The projection learning gives the minimum variance unbiased estimator of the orthogonal projection of the unknown true function f onto $\mathcal{R}(A_K^*)$ (the range of A_K^*), and the parametric projection learning gives its improvement, incorporating a relaxation of the unbiasedness of the projection learning. The parametric projection learning includes the projection learning as a special case. The parametric projection learning is defined as follows:

Definition 3 *[10, 11] The parametric projection learning B_{PPL} is defined by*

$$\begin{aligned} B_{PPL}(\gamma) = \mathrm{argmin}_B[\mathrm{tr}[(BA_K - P_{\mathcal{R}(A_K^*)})(BA_K - P_{\mathcal{R}(A_K^*)})^*] \\ + \gamma E_{\boldsymbol{n}}||B\boldsymbol{n}||^2], \end{aligned} \tag{10}$$

where $P_{\mathcal{R}(A_K^*)}$ and γ denote the orthogonal projector onto $\mathcal{R}(A_K^*)$ and a real positive parameter that controls the trade-off of the two terms, which works as a relaxation of the unbiasedness, respectively.

One of the solutions of the parametric projection learning is given by

$$B_{PPL}(\gamma) = A_K^*(A_K A_K^* + \gamma Q)^+ \tag{11}$$

as shown in [10, 11], where the super script $^+$ denotes the Moore-Penrose generalized inverse [12] and Q denotes the noise correlation matrix defined by

$$Q = E_{\boldsymbol{n}}[\boldsymbol{n}\boldsymbol{n}'].$$

Finally, the solution of the parametric projection learning is given by

$$\hat{f}(\boldsymbol{x}) = B_{PPL}\boldsymbol{y},$$

and the concrete form of it is written by

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \left(\sum_{i=1}^{\ell}\left[K(\boldsymbol{x},\boldsymbol{x_i})\otimes \boldsymbol{e}_i^{(\ell)}\right]\right)(G+\gamma Q)^+\boldsymbol{y} \\ &= \sum_{i=1}^{\ell}\boldsymbol{y}'(G+\gamma Q)^+\boldsymbol{e}_i^{(\ell)}K(\boldsymbol{x},\boldsymbol{x_i}),\end{aligned} \tag{12}$$

where $G = A_K A_K^*$ is the Gram's matrix of K written by $G = (g_{ij})$, $g_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which is easily confirmed by using the properties Eq.(4) of the Schatten product. Note that the assumption $Q = O$ yields the solution based on the Moore-Penrose generalized inverse of A_K.

4 Model Selection Using a Class of Kernels with an Invariant Metric

In general, the solution of kernel-based learning machines is given by a linear combination of $K(\boldsymbol{x}, \boldsymbol{x}_i)$ that spans $\mathcal{R}(A_K^*)$. Thus, the validity of the model depends on $||f - P_{\mathcal{R}(A_K^*)}f||^2_{\mathcal{H}_K}$. However, we can not directly evaluate it, since f in unknown. In this section, we construct a framework of selection of a good kernel that minimizes $||f - P_{\mathcal{R}(A_K^*)}f||^2_{\mathcal{H}_K}$ by incorporating a class of kernels with an invariant metric.

Let K_0 be a specific kernel and let $\mathcal{K}$ be a class of kernels satisfying

$$\mathcal{H}_K \subset \mathcal{H}_{K_0} \tag{13}$$

and

$$\langle f, g\rangle_{\mathcal{H}_K} = \langle f, g\rangle_{\mathcal{H}_{K_0}}, \tag{14}$$

for any $K \in \mathcal{K}$ and any functions $f, g \in \mathcal{H}_K$. Let

$$\mathcal{S}_{\mathcal{K}} = \{f | f \in \mathcal{H}_K \textit{ for all } K \in \mathcal{K}\}. \tag{15}$$

We assume that $\mathcal{S}_{\mathcal{K}} \neq \phi$. Thus, $\langle f, g \rangle_{\mathcal{H}_K}$ is invariant for any $K \in \mathcal{K}$ and any $f, g \in \mathcal{S}_{\mathcal{K}}$, which means that $K \in \mathcal{K}$ has the invariant metric that is the same with that of $\mathcal{H}_{K_0}$ for any $f \in \mathcal{S}_{\mathcal{K}}$. Note that $||f||^2_{\mathcal{H}_K}$ is also invariant for any $K \in \mathcal{K}$ and any $f \in \mathcal{S}_{\mathcal{K}}$.

We assume that $f \in \mathcal{S}_{\mathcal{K}}$ and let

$$f = P_{\mathcal{R}(A^*_K)} f + (I - P_{\mathcal{R}(A^*_K)}) f \tag{16}$$

be a decomposition of f with $K \in \mathcal{K}$, then

$$\begin{aligned} ||f||^2_{\mathcal{H}_K} &= ||P_{\mathcal{R}(A^*_K)} f||^2_{\mathcal{H}_K} + ||(I - P_{\mathcal{R}(A^*_K)}) f||^2_{\mathcal{H}_K} \\ &= ||P_{\mathcal{R}(A^*_K)} f||^2_{\mathcal{H}_{K_0}} + ||(I - P_{\mathcal{R}(A^*_K)}) f||^2_{\mathcal{H}_{K_0}} \end{aligned} \tag{17}$$

holds and it immediately follows that

$$||f||^2_{\mathcal{H}_K} \geq ||P_{\mathcal{R}(A^*_K)} f||^2_{\mathcal{H}_{K_0}}. \tag{18}$$

Thus, it is guaranteed that $||(I - P_{\mathcal{R}(A^*_K)}) f||^2_{\mathcal{H}_K} (= ||(I - P_{\mathcal{R}(A^*_K)}) f||^2_{\mathcal{H}_{K_0}})$ is minimized by

$$K_{opt} = \mathrm{argmax}_{K \in \mathcal{K}} ||P_{\mathcal{R}(A^*_K)} f||^2_{\mathcal{H}_{K_0}}, \tag{19}$$

which means that the selection of the best kernel from $\mathcal{K}$ is achieved.

As is mentioned in the previous section, a minimum variance unbiased estimator of $P_{\mathcal{R}(A^*_K)} f$ is given by the projection learning. However, its variance may be too large to use the solution as an approximation of $P_{\mathcal{R}(A^*_K)} f$. Thus, we may have to use an another solution, such as that based on a regularization scheme, as an approximation of $P_{\mathcal{R}(A^*_K)} f$, for instance.

5 Numerical Examples

In this section, we show a numerical example of a regression estimation of a one-dimensional function in order to investigate the properties of the proposed framework of a kernel selection.

We adopt L^2 as $\mathcal{H}_{K_0}$ and the sinc kernel defined by

$$K^{\alpha}_S(x, \tilde{x}) = \frac{\sin \alpha(x - \tilde{x})}{\pi(x - \tilde{x})}, \ \alpha \in [\alpha_s, \alpha_e], \ 0 < \alpha_s < \alpha_e. \tag{20}$$

as a class of kernels with an invariant metric. In fact, the sinc kernel has the same metric with L^2 as shown in [13]. Moreover,

$$\mathcal{H}_{K^{\alpha_1}_S} \subset \mathcal{H}_{K^{\alpha_2}_S} \tag{21}$$

holds for any $\alpha_1 \leq \alpha_2$, since the RKHS corresponding to K^{α}_S is the space of band-limited functions in $[-\alpha, \alpha]$ in the Fourier domain. According to the monotonicity of the RKHSs corresponding to the sinc kernels,

$$\mathcal{S}_{\mathcal{K}} = \{f | f \in \mathcal{H}_{K^{\alpha}_S} \ for \ all \ \alpha \in [\alpha_s, \alpha_e]\} = \{f | f \in \mathcal{H}_{K^{\alpha_s}_S}\}. \tag{22}$$

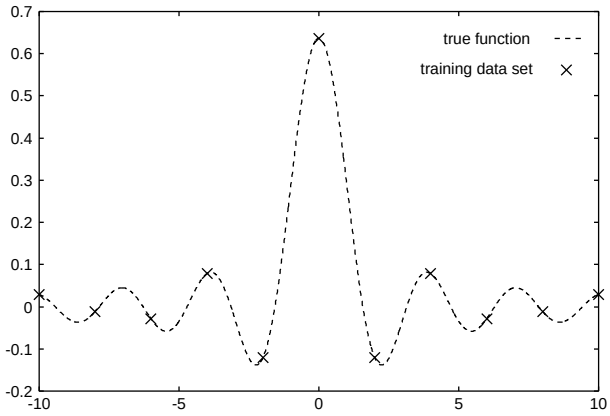

Fig. 1. The relation of the training data set and the unknown true function.

Thus, the unknown true function f must belong to $\mathcal{H}_{K_S^{\alpha_s}}$ to make our framework to be consistent for any $\alpha \in [\alpha_s, \alpha_e]$.

We use

$$f(x) = \frac{\sin 2x}{\pi x} \tag{23}$$

as the unknown true function f and

$$\{(f(x_i), x_i) | x_i \in \{-10, -8, \ldots, -2, 0, 2, \ldots, 8, 10\}\} \tag{24}$$

as the given training data set. Figure 1 shows the relation of the training data set and the unknown true function. We adopt A_K^+ as a learning machine, since $Q = O$ in this case.

We dare to adopt $[1.5, 2.5]$ for the interval of the parameter searching. Note that when $\alpha \in [1.5, 2)$, the condition $f \in \mathcal{H}_{K_S^\alpha}$ is broken, that is, the estimated function obtained by A_K^+ is no longer the orthogonal projection of f. The result with the condition $\alpha \in [1.5, 2)$ could reveal the importance of the condition $f \in \mathcal{H}_K$ in machine learning problem. On the other hand, when $\alpha \in [2, 2.5]$, the consistency of our framework is guaranteed and the result based on it could reveal the validity of our framework. Figure 2 shows the transitions of $||\hat{f}||_{L^2}^2$, $||f - \hat{f}||_{L^2}^2$, and the sum of them with respect to α. Figures $3 \sim 5$ show the learning results with the parameters $\alpha = 1.5,\ 2.0,\ 2.5$, respectively.

According to the result shown in Fig.2 with $\alpha \in [1.5, 2)$, it is confirmed that $f \neq \mathcal{H}_{K_S^\alpha}$ causes the fail of estimation of the orthogonal projection of f. In fact, the norm of $\hat{f}$ is larger than that of f. Thus, it is concluded that adopting the kernel whose RKHS does not include f makes no sense for learning.

On the other hand, when $\alpha \in [2, 2.5]$ is satisfied, that is, $f \in \mathcal{H}_{K_S^\alpha}$ holds, it is confirmed that $\hat{f}$ is the orthogonal projection of f, since the sum of $||\hat{f}||_{L^2}^2$ and

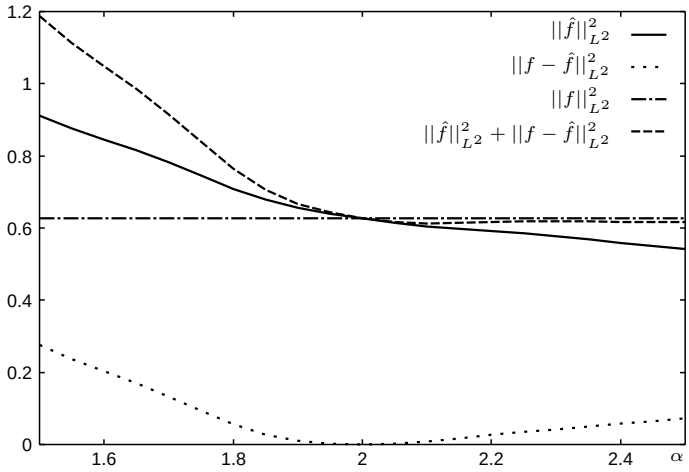

Fig. 2. Transitions of the squared norm of the estimated function , that of the error, and the sum of them with respect to α.

$||f-\hat{f}||^2_{L^2}$ is nearly equal to $||f||^2_{L^2}$. Moreover, it is confirmed that the maximizer of $||\hat{f}||^2_{L^2}$, satisfying $f \in \mathcal{H}_{K_S^\alpha}$, actually catches the best parameter $\alpha = 2$, which supports the validity of our framework.

Remarks

We used a noise-free case in the numerical example. However, it is inevitable to consider the noise in practical cases.

As mentioned in the previous section, when the noise exists, the solution based on the projection learning is not robust in general. Thus, we may have to use a regularization scheme such as parametric projection learning with the optimal parameter chosen by a parameter selection criterion such as the SIC[14].

Although we adopted the sinc kernel as a class of kernels with an invariant metric in the numerical example, the sinc kernel is not so useful, since the intersection of the corresponding RKHSs is reduced to the RKHS corresponding to the minimum parameter of the interval for the parameter searching due to the monotonicity of the corresponding RKHSs, which means that we can not adopt the interval that includes the unknown true parameter. Therefore, it is one of very important problems to construct a wide class of kernels with an invariant metric whose intersection includes a wide class of functions.

6 Conclusion

In this paper, we constructed a framework of a kernel selection on the basis of the projection-learning-based interpretation of learning problems, incorporating

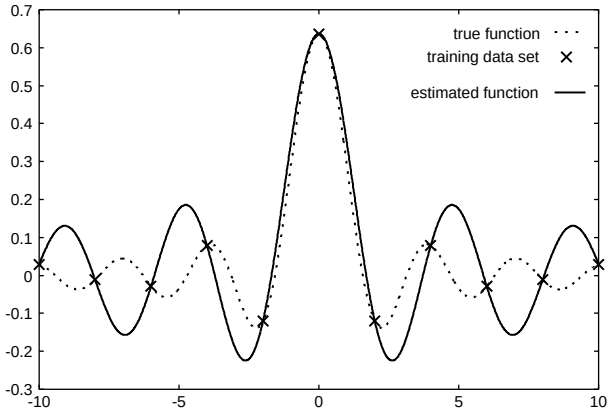

Fig. 3. The learning result with $\alpha = 1.5$.

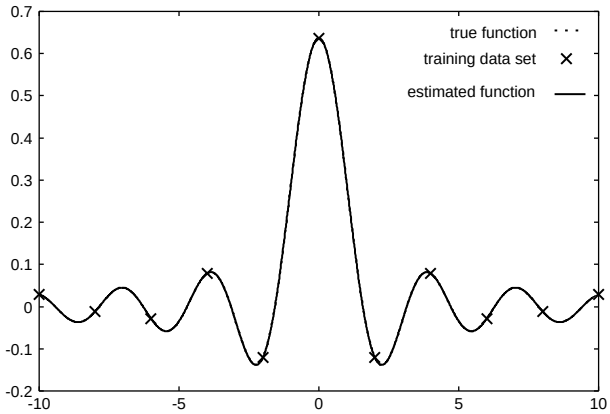

Fig. 4. The learning result with $\alpha = 2.0$.

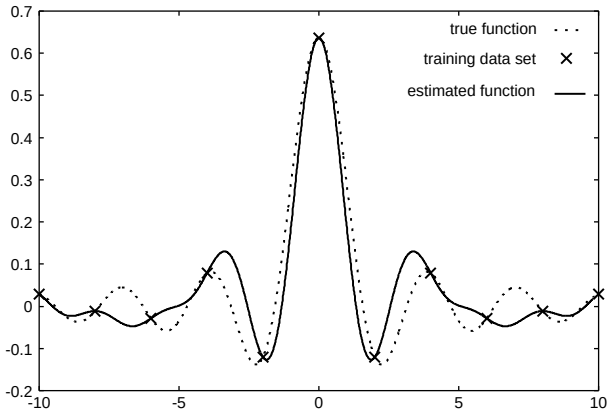

Fig. 5. The learning result with $\alpha = 2.5$.

a class of kernels with an invariant metric. Coping with the noise and construction of a class of kernels with an invariant metric that is suitable for practical problems are future works.

References

1. Muller, K., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks **12** (2001) 181–201
2. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1999)
3. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Recognition. Cambridge University Press, Cambridge (2004)
4. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
5. Aronszajn, N.: Theory of Reproducing Kernels. Transactions of the American Mathematical Society **68** (1950) 337–404
6. Mercer, J.: Functions of Positive and Negative Type and Their Connection with The Theory of Integral Equations. Transactions of the London Philosophical Society **A** (1909) 415–446
7. Ogawa, H.: Neural Networks and Generalization Ability. IEICE Technical Report **NC95-8** (1995) 57–64
8. Schatten, R.: Norm Ideals of Completely Continuous Operators. Springer-Verlag, Berlin (1960)
9. Sugiyama, M., Ogawa, H.: Incremental Projection Learning for Optimal Generalization. Neural Networks **14** (2001) 53–66
10. Imai, H., Tanaka, A., Miyakoshi, M.: The family of parametric projection filters and its properties for perturbation. The IEICE Transactions on Information and Systems **E80–D** (1997) 788–794
11. Oja, E., Ogawa, H.: Parametric Projection Filter for Image and Signal Restoration. IEEE Transactions on Acoustics, Speech and Signal Processing **ASSP–34** (1986) 1643–1653
12. Rao, C.R., Mitra, S.K.: Generalized Inverse of Matrices and its Applications. John Wiley & Sons (1971)
13. Saitoh, S.: Integral Transforms, Reproducing Kernels and Their Applications. Addison Wesley Longman Ltd, UK (1997)
14. Sugiyama, M., Ogawa, H.: Subspace Information Criterion for Model Selection. Neural Computation **13** (2001) 1863–1889