Analyses on Generalization Error of Ensemble Kernel Regressors

Akira Tanaka¹, Ichigaku Takigawa², Hideyuki Imai¹, and Mineichi Kudo¹

¹ Division of Computer Science, Hokkaido University,
N14W9, Kita-ku, Sapporo, 060-0814 Japan
`{takira,imai,mine}@main.ist.hokudai.ac.jp`

² Creative Research Institution, Hokkaido University,
N21W10, Kita-ku, Sapporo, 001-0021 Japan
`takigawa@cris.hokudai.ac.jp`

Abstract. Kernel-based learning is widely known as a powerful tool for various fields of information science such as pattern recognition and regression estimation. For the last few decades, a combination of different learning machines so-called ensemble learning, which includes learning with multiple kernels, have attracted much attention in this field. Although its efficacy was revealed numerically in many works, its theoretical grounds are not investigated sufficiently. In this paper, we discuss regression problems with a class of kernels and show that the generalization error by an ensemble kernel regressor with the class of kernels is smaller than the averaged generalization error by kernel regressors with each kernel in the class.

Keywords: kernel regressor, ensemble learning, orthogonal projection, generalization error.

1 Introduction

Kernel-based learning machines [1], represented by the support vector machine [2] and the kernel ridge regressor [3], are widely recognized as powerful tools for various fields of information science such as pattern recognition and regression estimation. In general, an appropriate model selection is required in order to obtain a desirable learning result by kernel machines. Although the model selection in a fixed model space, such as selection of a regularization parameter, is sufficiently investigated in terms of theoretical and practical senses (see [4, 5] for instance), the selection of a model space is not sufficiently investigated in terms of a theoretical sense, while many practical algorithms for selection of a kernel (or its parameters) are proposed. In our previous works [6–9], we discussed the generalization error of a model space specified by a kernel and obtained some theoretical results.

For the last few decades, learning based on multiple kernels have attracted much attention in this field, which can be regarded as one of model selection schemes. The ensemble kernel learning (see [2] for instance), which is a combination of kernel-based learning machines, is representative one, whose theoretical

grounds are not also investigated sufficiently. In this paper, we discuss generalization errors of the ensemble kernel regressor with a class of kernels and the kernel regressor using an individual kernel in the class; and show that the generalization error of the ensemble kernel regressor is smaller than the averaged generalization error of kernel regressors by each kernel.

2 Mathematical Preliminaries for the Theory of Reproducing Kernel Hilbert Spaces

In this section, we give some mathematical preliminaries concerned with the theory of reproducing kernel Hilbert spaces [10, 11].

Definition 1. [10] 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 $K(\mathbf{x}, \tilde{\mathbf{x}})$, ($\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{D}$) is called a reproducing kernel of $\mathcal{H}$, if the following two conditions hold.

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

$$f(\tilde{\mathbf{x}}) = \langle f(\cdot), K(\cdot, \tilde{\mathbf{x}}) \rangle_{\mathcal{H}}, \quad (1)$$

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, abbreviated by RKHS. Eq.(1) is called the reproducing property of a kernel and it enables us to treat a value of a function at a point in $\mathcal{D}$. Note that reproducing kernels are positive definite [10]:

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

for any $N \in \mathbf{N}$, $c_1, \dots, c_N \in \mathbf{R}$, and $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{D}$. In addition, $K(\mathbf{x}, \tilde{\mathbf{x}}) = K(\tilde{\mathbf{x}}, \mathbf{x})$ holds for any $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{D}$ [10]. If a reproducing kernel $K(\mathbf{x}, \tilde{\mathbf{x}})$ exists, it is unique [10]. Conversely, every positive definite function $K(\mathbf{x}, \tilde{\mathbf{x}})$ has the unique corresponding RKHS [10]. Hereafter, the RKHS corresponding to a reproducing kernel $K(\mathbf{x}, \tilde{\mathbf{x}})$ is denoted by $\mathcal{H}_K$. In the following contents, we simply use the symbol K for a kernel by omitting $(\mathbf{x}, \tilde{\mathbf{x}})$ except the cases where it is needed.

Next, we introduce the Schatten product [12] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2. [12] 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

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

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 $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

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

where the superscript $*$ denotes the adjoint operator.

We give the following theorem concerned with the sum of reproducing kernels, which is used in the following contents.

Theorem 1. [10] *If K_i is the reproducing kernel of the class F_i with the norm $\|\cdot\|_i$, then $K = K_1 + K_2$ is the reproducing kernel of the class F of all functions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$, and with the norm defined by*

$$\|f(\cdot)\|^2 = \min [\|f_1(\cdot)\|_1^2 + \|f_2(\cdot)\|_2^2], \quad (5)$$

the minimum taken for all the decompositions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$.

The most important consequence of Theorem 1 is that the RKHS corresponding to $K = K_1 + K_2$ includes $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$.

3 Formulation of Regression Problems

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

$$y_i = f(\mathbf{x}_i) + n_i, \quad (6)$$

where $f(\cdot)$ denotes the unknown true function and n_i denotes zero-mean additive noise. The aim of the regression problem is to estimate the unknown function $f(\cdot)$ by using the given training data set and statistical properties of the noise.

In this paper, we assume that the unknown function $f(\cdot)$ belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel K . If $f(\cdot) \in \mathcal{H}_K$, then Eq.(6) is rewritten as

$$y_i = \langle f(\cdot), K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_K} + n_i, \quad (7)$$

on the basis of the reproducing property of kernels. Let $\mathbf{y} = [y_1, \dots, y_\ell]'$ and $\mathbf{n} = [n_1, \dots, n_\ell]'$ with the superscript $'$ denoting the transposition operator, then applying the Schatten product to Eq.(7) yields

$$\mathbf{y} = \left(\sum_{k=1}^{\ell} [\mathbf{e}_k^{(\ell)} \otimes K(\cdot, \mathbf{x}_k)] \right) f(\cdot) + \mathbf{n}, \quad (8)$$

where $\mathbf{e}_k^{(\ell)}$ denotes the ℓ -dimensional unit vector whose k -th element is unity. For a convenience of description, we write

$$A_{K,X} = \left(\sum_{k=1}^{\ell} [\mathbf{e}_k^{(\ell)} \otimes K(\cdot, \mathbf{x}_k)] \right), \quad (9)$$

where $X = \{\mathbf{x}_1, \dots, \mathbf{x}_\ell\}$. Note that $A_{K,X}$ is a linear operator from $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(8) can be written by

$$\mathbf{y} = A_{K,X} f(\cdot) + \mathbf{n}, \quad (10)$$

which represents the relationship between the unknown true function $f(\cdot)$ and an output vector $\mathbf{y}$. Therefore, a regression problem can be interpreted as an inversion problem of the linear equation Eq.(10) [13].

4 Kernel Specific Generalization Error and Some Known Results

In general, a learning result by kernel machines is represented by a linear combination of $K(\cdot, \mathbf{x}_k)$, which implies that the learning result is an element in the range space of the linear operator $A_{K,X}^*$, written as $\mathcal{R}(A_{K,X}^*)$, since

$$\begin{aligned} \hat{f}(\cdot) &= A_{K,X}^* \boldsymbol{\alpha} = \left(\sum_{k=1}^{\ell} [K(\cdot, \mathbf{x}_k) \otimes \mathbf{e}_k^{(\ell)}] \right) \boldsymbol{\alpha} \\ &= \sum_{k=1}^{\ell} \alpha_k K(\cdot, \mathbf{x}_k) \end{aligned} \quad (11)$$

holds, where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_\ell]'$ denotes an arbitrary vector in $\mathbf{R}^\ell$. The point at issue in this paper is to discuss goodness of a model space, that is, the generalization error of $\mathcal{R}(A_{K,X}^*)$ which is independent from learning criteria. Therefore, we define the generalization error of kernel machines specified by a kernel K and a set of input vectors X as the distance between the unknown true function $f(\cdot)$ and $\mathcal{R}(A_{K,X}^*)$ written as

$$J_{K_e}(f(\cdot); K, X) = \|f(\cdot) - P_{K,X} f(\cdot)\|_{\mathcal{H}_{K_e}}^2, \quad (12)$$

where $P_{K,X}$ denotes the orthogonal projector onto $\mathcal{R}(A_{K,X}^*)$ in $\mathcal{H}_K$ and $\|\cdot\|_{\mathcal{H}_{K_e}}$ denotes the induced norm of $\mathcal{H}_{K_e}$. Note that K_e may or may not be identical to K . Selection of an element in $\mathcal{R}(A_{K,X}^*)$ as a learning result is out of the scope of this paper since the selection depends on learning criteria. We also ignore the observation noise in the following contents since the noise does not affect Eq.(12).

Here, we give some propositions as preparations to evaluate Eq.(12).

Lemma 1. [6]

$$P_{K,X} = \sum_{i,j=1}^{\ell} (G_{K,X}^+)_{ij} [K(\cdot, \mathbf{x}_i) \otimes K(\cdot, \mathbf{x}_j)], \quad (13)$$

where $G_{K,X}$ denotes the Gramian matrix of K with X and the superscript $+$ denotes the Moore-Penrose generalized inverse[14].

From Lemma 1, the orthogonal projection of $f(\cdot) \in \mathcal{H}_K$ onto $\mathcal{R}(A_{K,X}^*)$ is given as

$$P_{K,X}f(\cdot) = \sum_{i,j=1}^{\ell} f(\mathbf{x}_i)(G_{K,X}^+)_{ij}K(\cdot, \mathbf{x}_j). \quad (14)$$

5 Analyses on Generalization Error of Ensemble Kernel Regressors

We consider a class of kernels $\mathcal{K} = \{K_1, \dots, K_n\}$ and corresponding RKHS written as $\mathcal{H}_{K_p}$, ($p \in \{1, \dots, n\}$). We assume that

$$S = \overline{\cap_{p=1}^n \mathcal{H}_{K_p}} \quad (15)$$

forms a non-empty linear class.

In this section, we discuss the generalization error of the ensemble kernel regressor with the kernels in $\mathcal{K}$ and the mean of the generalization errors of the kernel regressors with each kernel in $\mathcal{K}$. For any $f(\cdot) \in S$, the learning result by the kernel regressor with K_p , ($p \in \{1, \dots, n\}$) and the set of input vectors X is written as

$$\hat{f}_p(\cdot) = \sum_{i,j=1}^{\ell} f(\mathbf{x}_i)(G_{K_p,X}^+)_{ij}K_p(\cdot, \mathbf{x}_j) = P_{K_p,X}f(\cdot) \quad (16)$$

from Eq.(14); and its estimation error is reduced to

$$e_p(\cdot) = f(\cdot) - \hat{f}_p(\cdot). \quad (17)$$

Since $e_p(\cdot) \in \mathcal{H}_{K_p}$, we adopt an RKHS including all $\mathcal{H}_{K_p}$, ($p \in \{1, \dots, n\}$) in order to evaluate the norm of $e_p(\cdot)$, ($p \in \{1, \dots, n\}$) uniformly. As such an RKHS, we adopt $\mathcal{H}_{K_e}$ whose corresponding kernel is given as

$$K_e = \sum_{p=1}^n K_p \quad (18)$$

from Theorem 1. Therefore, the expected squared error of the kernel regressors with K_p , ($p \in \{1, \dots, n\}$) and X , evaluated in $\mathcal{H}_{K_e}$, is reduced to

$$J_m = \frac{1}{n} \sum_{p=1}^n \|e_p(\cdot)\|_{\mathcal{H}_{K_e}}^2. \quad (19)$$

On the other hand, the ensemble kernel regressor by the kernels in $\mathcal{K}$ is represented by

$$\hat{f}_e(\cdot) = \frac{1}{n} \sum_{p=1}^n P_{K_p,X}f(\cdot) = \frac{1}{n} \sum_{p=1}^n \hat{f}_p(\cdot), \quad (20)$$

and its generalization error, evaluated in $\mathcal{H}_{K_e}$, is reduced to

$$\begin{aligned} J_e &= \|f(\cdot) - \hat{f}_e(\cdot)\|_{\mathcal{H}_{K_e}}^2 \\ &= \left\| \frac{1}{n} \sum_{p=1}^n (f(\cdot) - \hat{f}_p(\cdot)) \right\|_{\mathcal{H}_{K_e}}^2 = \left\| \frac{1}{n} \sum_{p=1}^n e_p(\cdot) \right\|_{\mathcal{H}_{K_e}}^2. \end{aligned} \quad (21)$$

Note that the adopting the norm of $\mathcal{H}_{K_e}$ in the evaluation J_e is valid since $f(\cdot), \hat{f}_e(\cdot) \in \mathcal{H}_{K_e}$ obviously holds.

The following theorem is the main result of this paper.

Theorem 2

$$J_m - J_e = \frac{1}{2n^2} \sum_{p,q=1}^n \left\| \hat{f}_p(\cdot) - \hat{f}_q(\cdot) \right\|_{\mathcal{H}_{K_e}}^2. \quad (22)$$

Proof. Since $e_p(\cdot) - e_q(\cdot) = \hat{f}_q(\cdot) - \hat{f}_p(\cdot)$ holds, we have

$$\begin{aligned} J_m - \frac{1}{2n^2} \sum_{p,q=1}^n \left\| \hat{f}_p(\cdot) - \hat{f}_q(\cdot) \right\|_{\mathcal{H}_{K_e}}^2 &= \frac{1}{n} \sum_{p=1}^n \|e_p(\cdot)\|_{\mathcal{H}_{K_e}}^2 - \frac{1}{2n^2} \sum_{p,q=1}^n \|e_p(\cdot) - e_q(\cdot)\|_{\mathcal{H}_{K_e}}^2 \\ &= \frac{1}{n} \sum_{p=1}^n \|e_p(\cdot)\|_{\mathcal{H}_{K_e}}^2 - \frac{1}{2n^2} \sum_{p,q=1}^n \left(\|e_p(\cdot)\|_{\mathcal{H}_{K_e}}^2 + \|e_q(\cdot)\|_{\mathcal{H}_{K_e}}^2 - 2\langle e_p(\cdot), e_q(\cdot) \rangle_{\mathcal{H}_{K_e}} \right) \\ &= \frac{1}{n^2} \sum_{p,q=1}^n \langle e_p(\cdot), e_q(\cdot) \rangle_{\mathcal{H}_{K_e}} = \left\| \frac{1}{n} \sum_{p=1}^n e_p(\cdot) \right\|_{\mathcal{H}_{K_e}}^2 = J_e, \end{aligned}$$

which concludes the proof. $\square$

According to Theorem 2, it is concluded that the generalization error of the ensemble kernel regressor is smaller than or equal to the averaged generalization error of the kernel regressors by each kernel since the right-hand side of Eq.(22) is trivially non-negative, which implies that the ensemble kernel regressor gives a better performance than the kernel regressor by each kernel in terms of the averaged performance. Moreover, we can observe from Eq.(22) that when the learning results by the kernel regressors with each kernel are quite similar, the ensemble kernel regressor loses its advantage since the right-hand side of Eq.(22) becomes smaller. On the contrary, when the estimation error by the kernel regressors with each kernel, that is $e_p(\cdot)$, tend to be orthogonal each other, the generalization error of the ensemble kernel regressor J_e tends to be J_m/n from Eq.(21) and the Pythagorean theorem.

6 Example

In this section, we give a toy example confirming Theorem 2. Let $\mathcal{K} = \{K_1, K_2\}$ with

$$K_1(x, y) = 1 + xy + x^2y^2, \quad (23)$$

$$K_2(x, y) = 1 + xy + x^3y^3 \quad (24)$$

be a class of kernels. Note that a basis of $\mathcal{H}_{K_1}$ is $\{1, x, x^2\}$ and that of $\mathcal{H}_{K_2}$ is $\{1, x, x^3\}$. Therefore, we have

$$S = \overline{\mathcal{H}_{K_1} \cap \mathcal{H}_{K_2}} = \overline{\text{span}\{1, x\}}. \quad (25)$$

We adopt $f(x) = 1 + 2x \in S$ as an unknown true function and $X = \{0, 1\}$ as a set of input points. Therefore, we have

$$\hat{f}_1(\cdot) = P_{K_1, X}f(\cdot) = 1 + x + x^2, \quad (26)$$

$$\hat{f}_2(\cdot) = P_{K_2, X}f(\cdot) = 1 + x + x^3 \quad (27)$$

from Eq.(14) and

$$\hat{e}_1(\cdot) = f(x) - \hat{f}_1(x) = (1 + 2x) - (1 + x + x^2) = x - x^2, \quad (28)$$

$$\hat{e}_2(\cdot) = f(x) - \hat{f}_2(x) = (1 + 2x) - (1 + x + x^3) = x - x^3. \quad (29)$$

Let $K_e(x, y) = K_1(x, y) + K_2(x, y) = 2 + 2xy + x^2y^2 + x^3y^3$, then it is trivial that $\dim \mathcal{H}_{K_e} = 4$. Since $e_1(x), e_2(x) \in \mathcal{H}_{K_e}$, they can be expressed by linear combinations of four linearly independent functions in $\mathcal{H}_{K_e}$, such as

$$\begin{aligned} K_e(x, 0) &= 2, \\ K_e(x, 1) &= 2 + 2x + x^2 + x^3, \\ K_e(x, 2) &= 2 + 4x + 4x^2 + 8x^3, \\ K_e(x, 3) &= 2 + 6x + 9x^2 + 27x^3. \end{aligned}$$

In fact, $e_1(x)$ and $e_2(x)$ are represented by

$$e_1(x) = \frac{1}{12} (-23K_e(x, 0) + 48K_e(x, 1) - 33K_e(x, 2) + 8K_e(x, 3)), \quad (30)$$

$$e_2(x) = \frac{1}{4} (-3K_e(x, 0) + 4K_e(x, 1) - 1K_e(x, 2)). \quad (31)$$

By the well known property $\langle K_e(\cdot, x), K_e(\cdot, y) \rangle_{\mathcal{H}_{K_e}} = K_e(x, y)$, we have

$$\|e_1(\cdot)\|_{\mathcal{H}_{K_e}}^2 = \|e_2(\cdot)\|_{\mathcal{H}_{K_e}}^2 = \frac{3}{2}, \quad (32)$$

which immediately yields

$$J_m = \frac{1}{2} \left(\|e_1(\cdot)\|_{\mathcal{H}_{K_e}}^2 + \|e_2(\cdot)\|_{\mathcal{H}_{K_e}}^2 \right) = \frac{3}{2}. \quad (33)$$

On the other hand, the learning result by the ensemble kernel regressor by K_1 and K_2 is reduced to

$$\hat{f}_e(x) = \frac{1}{2}(\hat{f}_1(x) + \hat{f}_2(x)) = 1 + x + \frac{x^2 + x^3}{2} \quad (34)$$

and we have

$$f(x) - \hat{f}_e(x) = x - \frac{x^2 + x^3}{2}, \quad (35)$$

which can be represented by

$$f(x) - \hat{f}_e(x) = \frac{1}{6}(-8K_e(x, 0) + 15K_e(x, 1) - 9K_e(x, 2) + 2K_e(x, 3)), \quad (36)$$

whose squared norm in $\mathcal{H}_{K_e}$ is

$$J_e = \|f(x) - \hat{f}_e(x)\|_{\mathcal{H}_{K_e}}^2 = 1. \quad (37)$$

Therefore, we have

$$J_m - J_e = \frac{1}{2}. \quad (38)$$

Accordingly, it is concluded that the ensemble kernel regressor gives a better result than the kernel regressor by the individual kernel K_1 or K_2 in these settings.

Finally, we evaluate the right-hand side of Eq.(22). Since

$$\begin{aligned} \hat{f}_1(x) - \hat{f}_2(x) &= x^2 - x^3 \\ &= \frac{1}{6}(7K_e(x, 0) - 18K_e(x, 1) + 15K_e(x, 2) - 4K_e(x, 3)) \end{aligned} \quad (39)$$

and $\|\hat{f}_1(x) - \hat{f}_2(x)\|_{\mathcal{H}_{K_e}}^2 = 2$, the right-hand side of Eq.(22) is reduced to

$$\begin{aligned} &\frac{1}{2 \cdot 2^2} \sum_{p,q=1}^2 \|\hat{f}_p(x) - \hat{f}_q(x)\|_{\mathcal{H}_{K_e}}^2 \\ &= \frac{1}{8}(\|\hat{f}_1(x) - \hat{f}_2(x)\|_{\mathcal{H}_{K_e}}^2 + \|\hat{f}_2(x) - \hat{f}_1(x)\|_{\mathcal{H}_{K_e}}^2) \\ &= \frac{1}{2}, \end{aligned} \quad (40)$$

which agrees with Eq.(38). Note that calculation of right-hand side of Eq.(22) does not require the unknown true function $f(\cdot)$, while the left-hand side of Eq.(22) includes $f(\cdot)$, which implies that we can obtain the reduced squared error by the ensemble kernel regressor only from the learning results by kernel regressors with each kernel.

7 Conclusion

In this paper, we discussed the generalization error of the ensemble kernel regressor with a class of kernels; and proved that the generalization error of the ensemble kernel regressor is smaller than the averaged generalization error of the kernel regressors by each kernel in the class. Similar analysis for the noise is one of our future works that should be undertaken.

Acknowledgment. This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500001.

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, 181–201 (2001)
2. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1999)
3. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge (2000)
4. Sugiyama, M., Ogawa, H.: Subspace Information Criterion for Model Selection. *Neural Computation* 13, 1863–1889 (2001)
5. Sugiyama, M., Kawanabe, M., Muller, K.: Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression. *Neural Computation* 16, 1077–1104 (2004)
6. Tanaka, A., Imai, H., Kudo, M., Miyakoshi, M.: Optimal Kernel in a Class of Kernels with an Invariant Metric. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) *S+SSPR 2008*. LNCS, vol. 5342, pp. 530–539. Springer, Heidelberg (2008)
7. Tanaka, A., Miyakoshi, M.: Theoretical Analyses for a Class of Kernels with an Invariant Metric. In: 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), pp. 2074–2077 (2010)
8. Tanaka, A., Imai, H., Kudo, M., Miyakoshi, M.: Theoretical Analyses on a Class of Nested RKHS's. In: 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), pp. 2072–2075 (2011)
9. Tanaka, A., Takigawa, I., Imai, H., Kudo, M.: Extended analyses for an optimal kernel in a class of kernels with an invariant metric. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) *SSPR&SPR 2012*. LNCS, vol. 7626, pp. 345–353. Springer, Heidelberg (2012)
10. Aronszajn, N.: Theory of Reproducing Kernels. *Transactions of the American Mathematical Society* 68, 337–404 (1950)
11. Mercer, J.: Functions of Positive and Negative Type and Their Connection with The Theory of Integral Equations. *Transactions of the London Philosophical Society A*, 415–446 (1909)
12. Schatten, R.: *Norm Ideals of Completely Continuous Operators*. Springer, Berlin (1960)
13. Ogawa, H.: Neural Networks and Generalization Ability. IEICE Technical Report NC95-8, 57–64 (1995)
14. Rao, C.R., Mitra, S.K.: *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons (1971)