Extended Analyses for an Optimal Kernel in a Class of Kernels with an Invariant Metric

Akira Tanaka[1], Ichigaku Takigawa[2], Hideyuki Imai[1], and Mineichi Kudo[1]

[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. 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. An appropriate model selection is required in order to obtain desirable learning results. In our previous work, we discussed a class of kernels forming a nested class of reproducing kernel Hilbert spaces with an invariant metric and proved that the kernel corresponding to the smallest reproducing kernel Hilbert space, including an unknown true function, gives the best model. In this paper, we relax the invariant metric condition and show that a similar result is obtained when a subspace with an invariant metric exists.

Keywords: kernel regressor, reproducing kernel Hilbert space, orthogonal projection, invariant metric.

1 Introduction

Learning based on kernel machines [1], represented by the support vector machine (SVM) [2] and the kernel ridge regression [3,4], is widely known as a powerful tool for various fields of information science such as pattern recognition, regression estimation, and density 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 [5,6] for instance), the selection of a model space itself is not sufficiently investigated in terms of a theoretical sense, while practical algorithms for selection of a kernel (or its parameters) such as the cross-validation are revealed. The difficulty of the theoretical analyses for selection of a kernel (or its parameters) lies on the fact that the metrics of two reproducing kernel Hilbert spaces (RKHS)[7,8] corresponding to two different kernels may differ in general, which means that we do not have a unified framework to evaluate learning results obtained by different kernels. In order to avoid this difficulty, we considered a class of kernels whose corresponding RKHS's have an invariant

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 345–353, 2012.

metric and proved that the kernel corresponding to the smallest reproducing kernel Hilbert space, including an unknown true function, gives the best model in [9].

In this paper, we relax the invariant metric condition for the whole space and prove that a similar result is obtained when a subspace with an invariant metric exists in the smallest reproducing kernel Hilbert space.

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 [7,8].

Definition 1. *[7] 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 $\tilde{\boldsymbol{x}} \in \mathcal{D}$, $K(\cdot, \tilde{\boldsymbol{x}})$ is a function belonging to $\mathcal{H}$.*
2. *For every $\tilde{\boldsymbol{x}} \in \mathcal{D}$ and every $f \in \mathcal{H}$,*

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

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 [7]:

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

for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ holds for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ [7]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [7]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [7]. Hereafter, the RKHS corresponding to a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is denoted by $\mathcal{H}_K$.

Next, we introduce the Schatten product [10] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2. *[10] 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 superscript * denotes the adjoint operator.

We give some theorems concerned with sum and difference of reproducing kernels used in the following contents.

Theorem 1. *[7] 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 $f = f_1 + f_2$ with $f_i \in F_i$, and with the norm defined by*

$$||f||^2 = \min \left[||f_1||_1^2 + ||f_2||_2^2 \right], \tag{5}$$

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

Theorem 2. *[7] If K is the reproducing kernel of the class F with the norm $||\cdot||$, and if the linear class $F_1 \subset F$ forms a Hilbert space with the norm $||\cdot||_1$, such that $||f||_1 \geq ||f||$ for any $f \in F_1$, then the class F_1 possesses a reproducing kernel K_1 such that $K^c = K - K_1$ is also a reproducing kernel.*

Theorem 3. *[7] If K and K_1 are the reproducing kernels of the classes of F and F_1 with the norms $||\cdot||$, $||\cdot||_1$, and if $K - K_1$ is a reproducing kernel, then $F_1 \subset F$ and $||f_1||_1 \geq ||f_1||$ for every $f_1 \in F_1$.*

3 Formulation of Regression 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

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{6}$$

where f denotes the unknown true function and n_i denotes a zero-mean additive noise. The aim of regression problems is to estimate the unknown function f by using the given training data set and statistical properties of the 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.(6) is rewritten as

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

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 superscript $'$ denoting the transposition operator, then applying the Schatten product to Eq.(7) yields

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

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

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

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

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

which represents the relation between the unknown true function f and an output vector $\boldsymbol{y}$. Therefore, a regression problem can be interpreted as an inversion problem of the linear equation Eq.(10) [11]. In general, an estimated function $\hat{f}(\boldsymbol{x})$ is represented as

$$\hat{f}(\cdot) = L\boldsymbol{y}, \tag{11}$$

where L denotes a learning operator such as the support vector machine and the kernel ridge regressor.

4 Kernel Specific Generalization Ability and Some Known Results

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

$$\hat{f}(\cdot) = A_{K,X}^* \boldsymbol{\alpha} = \left(\sum_{k=1}^{\ell} [K(\cdot, \boldsymbol{x}_k) \otimes \boldsymbol{e}_k^{(\ell)}] \right) \boldsymbol{\alpha} = \sum_{k=1}^{\ell} \alpha_k K(\cdot, \boldsymbol{x}_k) \tag{12}$$

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

$$J(f; K, X) = ||f - P_{K,X} f||_{\mathcal{H}_K}^2, \tag{13}$$

where $P_{K,X}$ denotes the orthogonal projector onto $\mathcal{R}(A_{K,X}^*)$ and $||\cdot||_{\mathcal{H}_K}$ denotes the induced norm of $\mathcal{H}_K$. Note that the orthogonality of $P_{K,X}$ is also defined by the metric of $\mathcal{H}_K$. 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.(13).

Here, we give some propositions as preparations to evaluate Eq.(13).

Lemma 1. *[9]*

$$P_{K,X} = \sum_{i=1}^{\ell}\sum_{j=1}^{\ell}(G_{K,X}^{+})_{i,j}\left[K(\cdot,\boldsymbol{x}_i)\otimes K(\cdot,\boldsymbol{x}_j)\right], \tag{14}$$

where $G_{K,X}$ denotes the Gramian matrix of K with X and the superscript $^{+}$ denotes the Moore-Penrose generalized inverse[12].

Lemma 2. *[9] For any $f \in \mathcal{H}_K$ and X,*

$$||P_{K,X}f||_{\mathcal{H}_K}^2 = \boldsymbol{f}'G_{K,X}^{+}\boldsymbol{f} \tag{15}$$

holds, where $\boldsymbol{f} = [f(\boldsymbol{x}_1),\ldots,f(\boldsymbol{x}_\ell)]'$.

Since $P_{K,X} = A_{K,X}^{+}A_{K,X}$, the learning result

$$\hat{f}(\cdot) = A_{K,X}^{+}\boldsymbol{y} \tag{16}$$

gives the minimum norm least-squares solution of f and it gives the orthogonal projection of f onto $\mathcal{R}(A_{K,X}^{*})$ in noise free cases.

In [9], we discussed a class of nested RKHS's with an invariant metric. We review important results given in [9].

Let K_1 and K_2 be reproducing kernels satisfying

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2} \tag{17}$$

and

$$||f||_{\mathcal{H}_{K_1}}^2 = ||f||_{\mathcal{H}_{K_2}}^2 \quad \textit{for any } f \textit{ in } \mathcal{H}_{K_1}. \tag{18}$$

Then we have the following theorem.

Theorem 4. *[9] Let K_1 and K_2 be kernels. If Eqs.(17) and (18) hold, then for any $f \in \mathcal{H}_{K_1}$ and X,*

$$||f - P_{K_1,X}f||_{\mathcal{H}_{K_2}}^2 \leq ||f - P_{K_2,X}f||_{\mathcal{H}_{K_2}}^2 \tag{19}$$

holds.

This theorem claims that the kernel corresponding to the smallest RKHS in the class of RKHS's with an invariant metric gives the best model space if an unknown function f belongs to the smallest one. Note that Theorem 4 does not hold in general without the invariant metric condition. In fact, we gave an example that Theorem 4 does not hold without the invariant metric condition in [13].

From Eq.(18) and Theorem 2, there exists the reproducing kernel K^c satisfying

$$K_2 = K_1 + K^c. \tag{20}$$

In [14], we discussed the relationship between these kernels (or corresponding RKHS's) and the condition of the invariant metric; and obtained the following theorem.

Theorem 5. *Let K_1 and $K_2 = K_1 + K^c$ be kernels whose corresponding RKHS's satisfy $\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2}$. The following three statements are equivalent each other.*

1) For any $f \in \mathcal{H}_{K_1}$, $||f||^2_{\mathcal{H}_{K_1}} = ||f||^2_{\mathcal{H}_{K_2}}$,

2) $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$,

3) For any $f_1 \in \mathcal{H}_{K_1}$ and $f_2 \in \mathcal{H}_{K^c}$, $\langle f_1, f_2 \rangle_{\mathcal{H}_{K_2}} = 0$.

5 Analyses under Relaxed Invariant Metric Condition

Let K_1 and $K_2 = K_1 + K^c$ be kernels where K^c is also a kernel, then corresponding RKHS's satisfy

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2} \tag{21}$$

from Theorem 1; and we have

$$||f||_{\mathcal{H}_{K_1}} \geq ||f||_{\mathcal{H}_{K_2}} \tag{22}$$

for any $f \in \mathcal{H}_{K_1}$ from Theorem 3.

We assume that a linear class $F \subset \mathcal{H}_{K_1}$ forms a Hilbert space with the norm $|| \cdot ||_F$ and assume that

$$||f||_F = ||f||_{\mathcal{H}_{K_i}} \quad \textit{for any } f \in F, \ (i \in \{1, 2\}). \tag{23}$$

Then on the basis of Theorems 2, there exists a kernel K_F such that

$$K_i^c = K_i - K_F, \ (i \in \{1, 2\}) \tag{24}$$

is also a kernel and

$$\mathcal{H}_{K_F} \cap \mathcal{H}_{K_i^c} = \{0\} \tag{25}$$

holds from Theorem 5. Note that it is trivial that Eq.(24) can be rewritten as

$$K_i = K_i^c + K_F, \ (i \in \{1, 2\}). \tag{26}$$

Since K_F is guaranteed to be a kernel, we use $\mathcal{H}_{K_F}$ instead of F, hereafter. Note that we can also represent K_2 as

$$K_2 = K_F + K_1^c + K^c. \tag{27}$$

If we have an explicit form of K_F, then $\mathcal{R}(A^*_{K_F,X})$ gives a better model than $\mathcal{R}(A^*_{K_1,X})$ and $\mathcal{R}(A^*_{K_2,X})$ for any $f \in \mathcal{H}_{K_F}$ according to Theorem 4. However in general, we can not always obtain K_F from K_1 and K_2 (or K^c). When $\mathcal{H}_{K_1^c} \cap \mathcal{H}_{K^c} = \{0\}$, $\mathcal{R}(A^*_{K_1,X})$ gives a better model than $\mathcal{R}(A^*_{K_2,X})$ for any $f \in \mathcal{H}_{K_1}$ according to Theorems 4 and 5. However, when $\mathcal{H}_{K_1^c} \cap \mathcal{H}_{K^c} \neq \{0\}$, $\mathcal{R}(A^*_{K_2,X})$ may be a better model than $\mathcal{R}(A^*_{K_1,X})$ for some $f \in \mathcal{H}_{K_1}$ since the metrics of $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ may differ.

The aim of this paper is to show that for any $f \in \mathcal{H}_{K_F}$, $\mathcal{R}(A^*_{K_1,X})$ gives a better model than $\mathcal{R}(A^*_{K_2,X})$ even if $\mathcal{H}_{K_1^c} \cap \mathcal{H}_{K^c} \neq \{0\}$.

Lemma 3. *[9] Let H_1 and H_2 be n.n.d. Hermitian matrices and let $\boldsymbol{y} \in \mathcal{R}(H_1)$, then*

$$J = \boldsymbol{y}^*(H_1^+ - (H_1 + H_2)^+)\boldsymbol{y} \geq 0 \tag{28}$$

holds.

The following theorem is the main result of this paper.

Theorem 6. *Let K_1 and $K_2 = K_1 + K^c$ be kernels where K^c is also a kernel. If Eq.(23) holds, then for any $f \in \mathcal{H}_{K_F}$ and X,*

$$\|f - P_{K_1,X}f\|^2_{\mathcal{H}_{K_2}} \leq \|f - P_{K_2,X}f\|^2_{\mathcal{H}_{K_2}} \tag{29}$$

holds.

Proof. From Eq.(22), we have

$$\begin{aligned} Z &= \|f - P_{K_2,X}f\|^2_{\mathcal{H}_{K_2,X}} - \|f - P_{K_1,X}f\|^2_{\mathcal{H}_{K_2,X}} \\ &\geq \|f - P_{K_2,X}f\|^2_{\mathcal{H}_{K_2,X}} - \|f - P_{K_1,X}f\|^2_{\mathcal{H}_{K_1,X}} = Z_1 \end{aligned}$$

and from Lemma 2, the Pythagorean theorem, the invariant metric condition,

$$\begin{aligned} Z_1 &= ||f||^2_{\mathcal{H}_{K_2}} - \boldsymbol{f}'G^+_{K_2,X}\boldsymbol{f} - (||f||^2_{\mathcal{H}_{K_1}} - \boldsymbol{f}'G^+_{K_1,X}\boldsymbol{f}) \\ &= ||f||^2_{\mathcal{H}_{K_2}} - \boldsymbol{f}'G^+_{K_2,X}\boldsymbol{f} - (||f||^2_{\mathcal{H}_{K_2}} - \boldsymbol{f}'G^+_{K_1,X}\boldsymbol{f}) \\ &= \boldsymbol{f}'G^{-1}_{K_1,X}\boldsymbol{f} - \boldsymbol{f}'G^{-1}_{K_2,X}\boldsymbol{f} \\ &= \boldsymbol{f}'\left(G^{-1}_{K_1,X} - G^{-1}_{K_2,X}\right)\boldsymbol{f} = Z_2 \end{aligned}$$

is obtained. Since

$$G_{K_1,X} = G_{K_F,X} + G_{K_1^c,X} \tag{30}$$

and

$$G_{K_2,X} = G_{K_F,X} + G_{K_1^c,X} + G_{K^c,X} \tag{31}$$

hold, the fact that

$$\boldsymbol{f} \in \mathcal{R}(A_{K_F,X}) = \mathcal{R}(A_{K_F,X}A^*_{K_F,X}) = \mathcal{R}(G_{K_F,X}) \subset \mathcal{R}(G_{K_F,X} + G_{K_1^c,X})$$

and Lemma 3 yield $Z_2 \geq 0$ which concludes the proof. □

According to Theorem 6, it is concluded that if the smaller RKHS include a subspace with an invariant metric, the kernel corresponding to the smaller RKHS gives a better model than larger one for any function in the subspace, which is an extension of the results obtained in [9].

6 Conclusion

In this paper, we discussed a class of kernels forming a nested class of RKHS's; and proved that if the smallest RKHS in the class has a subspace with an invariant metric, the kernel corresponding to the smallest RKHS gives the best model for any function in the subspace, which is a direct extension of our previous result obtained in [9]. Drastic relaxation of the invariant metric condition and extending the obtained results to practical learning machines such as the SVM and the kernel ridge regressor are ones of our future works.

Acknowledgement. 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. 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. Sugiyama, M., Ogawa, H.: Subspace Information Criterion for Model Selection. Neural Computation 13, 1863–1889 (2001)
6. 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)
7. Aronszajn, N.: Theory of Reproducing Kernels. Transactions of the American Mathematical Society 68, 337–404 (1950)
8. 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)
9. 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)
10. Schatten, R.: Norm Ideals of Completely Continuous Operators. Springer, Berlin (1960)
11. Ogawa, H.: Neural Networks and Generalization Ability. IEICE Technical Report NC95-8, 57–64 (1995)

12. Rao, C.R., Mitra, S.K.: Generalized Inverse of Matrices and its Applications. John Wiley & Sons (1971)
13. 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)
14. 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, pp. 2074–2077 (2010)