VARIABLE SELECTION FOR SUPPORT VECTOR MACHINES USING THE GBIC OF THE KERNEL LOGISTIC REGRESSION

Jiang, Ying
The Graduate School of Medicine, Kurume University

Yanagawa, Takashi
The Biostatistics Center, Kurume University

https://doi.org/10.5109/25905

出版情報：Bulletin of informatics and cybernetics. 42, pp.45-54, 2010-12. Research Association of Statistical Sciences
バージョン：
権利関係：

VARIABLE SELECTION FOR SUPPORT VECTOR MACHINES USING THE GBIC OF THE KERNEL LOGISTIC REGRESSION

by

Ying JIANG and **Takashi YANAGAWA**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.42

FUKUOKA, JAPAN
2010

Bulletin of Informatics and Cybernetics, Vol. 42, 2010

VARIABLE SELECTION FOR SUPPORT VECTOR MACHINES USING THE GBIC OF THE KERNEL LOGISTIC REGRESSION

By

Ying Jiang[*] and **Takashi Yanagawa**[†]

Abstract

In this paper we propose a method of variable selection for support vector machines based on the approximate relationship between a support vector machine and kernel logistic regression. First, we derive the generalized Bayesian information criterion (GBIC) of a kernel logistic regression. Then we select variables that minimize the GBIC, and propose to use them for a support vector machine. Finally, we apply the proposed method to identify peptides that could be related to pancreatic cancer.

Key Words and Phrases: variable selection, support vector machines, kernel logistic regression, generalized Bayesian information criterion, classification.

1. Introduction

The support vector machine is a two-class classification method developed by Vapnik (1995) as a machine learning method. It uses the idea of kernel substitution to classify data in a high-dimensional feature space, and has demonstrated excellent performance in fields as diverse as handwritten digit recognition (Scholkopf et al. (1997)), object recognition (Pontil and Verri (1998)), text categorization (Joachims (1999)), speaker identification (Wan and Campbell (2000)), and face recognition (Guo et al. (2000)). These authors applied the support vector machine to predict group labels using all available features in the data.

In recent years, applications of support vector machines with selected features in data are increasing (see, for example Weston et al. (2000)). One of the popular selection methods for linear support vector machines is Recursive Feature Elimination (RFE) (Guyon et al. (2002)). In RFE, one starts by training support vector machine using all features, then eliminate the feature with the smallest square weights in the result classifier, and repeat the same procedure with the remaining features until the set of selected features is small enough. Krupka et al. (2008) proposed a Meta-Feature based Predictive Feature Selection (MF-PFS) to improve the RFE.

Feature selection algorithms for support vector machines discussed above are based on optimization of machine learning. We consider the feature selection problem in this paper from statistical point of view, and call it the variable selection instead of feature

[*] The Graduate School of Medicine, Kurume University, 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan. tel +81–942–31–7835 a208gm008k@std.kurume-u.ac.jp

[†] The Biostatistics Center, Kurume University 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan. tel +81–942–31–7835 yanagawa __ takashi@kurume-u.ac.jp

selection. In statistical science, information criteria such as AIC and BIC are standard tools for variable selection. However, the support vector machine is not a stochastic classifier so it is difficult to construct those criteria directly. In this paper we develop a variable selection method suitable for support vector machine based on a relationship between support vector machine and kernel logistic regression discovered previously by Zhu and Hastie (2005). Kobayashi and Komaki (2006) used the same relationship to choose the tuning parameters of a support vector machine. For the information criterion of kernel logistic regression, Ando et al. (2004) developed a method of choosing a subset of training data by means of the BIC.

The remainder of this paper is organized as follows. In Section 2, we review the theory of support vector machines and kernel logistic regression. In Section 3, we derive the generalized Bayesian information criterion (GBIC) (Konishi et al. (2004)) for kernel logistic regression, and propose a new variable selection method for support vector machines. Finally, in Section 4, we apply the proposed method for identifying peptides that could be related to pancreatic cancer.

2. Support Vector Machines and Kernel Logistic Regression

Let D be the set of training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. Function $K(\mathbf{x}, \mathbf{x}_i) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, which assigns a weight to $\mathbf{x}_i$ based on its distance from $\mathbf{x}$, is called the kernel function. For any kernel function K which is symmetric and positive-semi definite, let $\mathcal{H}_k$ be the reproducing kernel Hilbert space defined by

$$\mathcal{H}_k \stackrel{\text{def}}{=} \left\{ f \middle| f(\mathbf{x}) = \sum_{i=0}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i),\ K(\mathbf{x}, \mathbf{x}_0) = 1,\ \alpha_i \in \mathbb{R} \right\},$$

where norm $\| \cdot \|_{\mathcal{H}_k}$ is given by

$$\| f \|^2_{\mathcal{H}_k} = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j),$$

for

$$f(\mathbf{x}) = \sum_{i=0}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i). \tag{1}$$

Hastie et al. (2001) showed that the support vector machine was equivalent to minimize the following equation with respect to $f \in \mathcal{H}_k$:

$$\frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2} \| f \|^2_{\mathcal{H}_k},$$

where $\lambda (> 0)$ is a tuning parameter, and $[t]_+$ is defined as $[t]_+ = \max(0, t)$.

Now suppose that $y_i \in \{-1, 1\}$ is an observed value of random variable Y_i, and that its conditional distribution given $\mathbf{x}_i$ is $p(\mathbf{x}_i) = P(Y_i = 1 | \mathbf{x}_i)$. For the f defined in (1), put

$$\log \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = f(\mathbf{x}).$$

Then the log likelihood function of y_i given $\mathbf{x}_i$ $(i = 1, 2, \cdots, n)$ is

$$\ell = -\sum_{i=1}^{n} \log\left[1 + \exp(-y_i f(\mathbf{x}_i))\right]. \tag{2}$$

A negative penalized log likelihood function is

$$\ell_\lambda = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i f(\mathbf{x}_i))\right] + \frac{\lambda}{2} \parallel f \parallel^2_{\mathcal{H}_k}.$$

The problem of minimizing ℓ_λ with respect to $f \in \mathcal{H}_k$ is called the Kernel Logistic Regression (KLR).

Zhu and Hastie (2005) found that $\log[1+\exp(-yf)]$ approximates $[1-yf]_+$ well. In this paper we develop a method of variable selection based on the information criterion for the KLR and propose to use the selected variables for a support vector machine.

Now, for the f given in (1), the negative penalized log likelihood function ℓ_λ is

$$\ell_\lambda(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp\left(-y_i \sum_{j=0}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)\right)\right] + \frac{\lambda}{2}\sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_n)'$. Thus, the KLR problem reduces to finding the $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ that minimizes $\ell_\lambda(\boldsymbol{\alpha})$.

Let $\mathbf{K}$ be an $n \times n$ matrix with (i, j)-th element $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, and put

$$\Phi = \begin{pmatrix} \mathbf{1}_n & \mathbf{K} \end{pmatrix} \text{ and } R = \begin{pmatrix} 1 & \mathbf{1}_n' \\ \mathbf{1}_n & \mathbf{K} \end{pmatrix}. \tag{4}$$

Let H be an $n \times n$ diagonal matrix whose i^{th} element is $p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))$. If the matrix $\Phi' H \Phi + n\lambda R$ is nonsingular, as seen in Ando et al. (2004), it follows that the solution $\hat{\boldsymbol{\alpha}}$ of the KLR problem (3) is also a solution of the following iteration:

$$\boldsymbol{\alpha}^{new} = (\Phi' H \Phi + n\lambda R)^{-1} \Phi' H (\Phi \boldsymbol{\alpha}^{old} + H^{-1}\boldsymbol{\delta}). \tag{5}$$

Here $\boldsymbol{\delta}$ is an n-dimensional vector whose i^{th} element is $(y_i + 1)/2 - p(\mathbf{x}_i)$.

3. Variable Selection

3.1. The generalized Bayesian information criterion for KLR

We use a radial basis function for K, i.e.,

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\sigma \parallel \mathbf{x} - \mathbf{y} \parallel^2). \tag{6}$$

σ is an unknown positive real value called the kernel parameter. Note that the solution $\hat{\boldsymbol{\alpha}}$ of the KLR problem (3) cannot be obtained unless λ and σ are specified.

Let x_i be the i^{th} component of $\mathbf{x} \in \mathbb{R}^p$. The variable selection problem can be phrased as follows: choose a subset $\{x_{j_1}, \cdots, x_{j_m}\}$ from $\{x_1, \cdots, x_p\}$, or equivalently a subset of indices $\{j_1, \cdots, j_m\}$ from $\{1, 2, \cdots, p\}$, such that the $\hat{\boldsymbol{\alpha}}$ obtained from

$(x_{j_1 i}, \cdots, x_{j_m i})$, $i = 1, \cdots, n$ is just as efficient as the $\hat{\boldsymbol{\alpha}}$ obtained from $(x_{1i}, \cdots, x_{pi})$, $i = 1, \cdots, n$. We explore the generalized Bayesian information criterion for variable selection.

Let $\Lambda = \{\lambda_1, \cdots, \lambda_{n_1}\}$ and $\Sigma = \{\sigma_1, \cdots, \sigma_{n_2}\}$ be sets of candidate tuning parameters and kernel parameters respectively. Let $S_c \subset \{1, 2, \cdots, p\}$ be a set of candidates of variables to be selected. There are $2^p - 1$ possible subsets S_c for each $\lambda \in \Lambda$ and $\sigma \in \Sigma$. We refer to the combinations of $\lambda \in \Lambda$, $\sigma \in \Sigma$ and S_c as models. There are $m = n_1 n_2 (2^p - 1)$ models altogether. After placing them in some order, we denote the j^{th} combination model by M_j.

The posterior probability $P(M_j|D)$ of model M_j given observed data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ may be represented by

$$P(M_j|D) = \frac{P\left(\{y_i\}_{i=1}^n | M_j, \{\mathbf{x}_i\}_{i=1}^n\right) P\left(M_j | \{\mathbf{x}_i\}_{i=1}^n\right)}{\sum_{k=1}^m P\left(\{y_i\}_{i=1}^n | M_k, \{\mathbf{x}_i\}_{i=1}^n\right) P\left(M_k | \{\mathbf{x}_i\}_{i=1}^n\right)}, \qquad j = 1, \cdots, m. \tag{7}$$

The model that attains $\max_{j=1,\cdots,m} P(M_j|D)$ is called the optimum model. When we suppose $P\left(M_j | \{\mathbf{x}_i\}_{i=1}^n\right) = 1/m$, it follows from (7) that the optimum model is given by the model M_j that minimizes the following GBIC.

$$\text{GBIC} = -2 \log P\left(\{y_i\}_{i=1}^n | M_j, \{\mathbf{x}_i\}_{i=1}^n\right). \tag{8}$$

GBIC is called the generalized Bayesian information criterion (Konishi et al. (2004)).

Let $\pi\left(\boldsymbol{\alpha} | M_j, \{\mathbf{x}_i\}_{i=1}^n\right)$ be the prior density for the $(n+1)$-dimensional parameter vector $\boldsymbol{\alpha}$ in KLR problem (3). We consider the following singular multivariate normal density for $\pi(\cdot)$:

$$\pi\left(\boldsymbol{\alpha} | M_j, \{\mathbf{x}_i\}_{i=1}^n\right) = (2\pi)^{-r/2} (n\lambda)^{r/2} |R|_+^{1/2} \exp\left(-\frac{n\lambda}{2} \boldsymbol{\alpha}' R \boldsymbol{\alpha}\right). \tag{9}$$

Here R is defined by equation (4), and $|R|_+$ is the product of r nonzero eigenvalues of R. Then we have the following theorem.

THEOREM 3.1. *Suppose that the prior density of $\boldsymbol{\alpha}$ is given by equation (9). Then the GBIC is represented as follows:*

$$GBIC = 2n\ell_\lambda(\hat{\boldsymbol{\alpha}}) - (n + 1 - r) \log(2\pi/n) - r \log \lambda - \log |R|_+ + \log |J(\hat{\boldsymbol{\alpha}})|,$$

where $J(\hat{\boldsymbol{\alpha}}) = \Phi' H \Phi / n + \lambda R$.

PROOF. It holds that

$$P\left(\{y_i\}_{i=1}^n | M_j, \{\mathbf{x}_i\}_{i=1}^n\right) = \int P\left(\{y_i\}_{i=1}^n | \boldsymbol{\alpha}, M_j, \{\mathbf{x}_i\}_{i=1}^n\right) \pi\left(\boldsymbol{\alpha} | M_j, \{\mathbf{x}_i\}_{i=1}^n\right) d\boldsymbol{\alpha}, \tag{10}$$

where $P\left(\{y_i\}_{i=1}^n | \boldsymbol{\alpha}, M_j, \{\mathbf{x}_i\}_{i=1}^n\right)$ is the conditional likelihood function given $\boldsymbol{\alpha}$, M_j and $\{\mathbf{x}_i\}_{i=1}^n$ that is identical to e^ℓ where ℓ is the log likelihood function given in (2) under M_j. Put

$$\tau(\boldsymbol{\alpha} | M_j, \{\mathbf{x}_i\}_{i=1}^n) = \frac{1}{n} \log\left[P\left(\{y_i\}_{i=1}^n | \boldsymbol{\alpha}, M_j, \{\mathbf{x}_i\}_{i=1}^n\right) \pi\left(\boldsymbol{\alpha} | M_j, \{\mathbf{x}_i\}_{i=1}^n\right)\right].$$

Then from (9), we have

$$\begin{aligned}
\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n) &= \frac{1}{n}\log P\left(\{y_i\}_{i=1}^n|\boldsymbol{\alpha},M_j,\{\mathbf{x}_i\}_{i=1}^n\right)\\
&\quad+\frac{1}{n}\left(-\frac{r}{2}\log 2\pi+\frac{r}{2}\log n\lambda+\frac{1}{2}\log|R|_+-\frac{n\lambda}{2}\boldsymbol{\alpha}'R\boldsymbol{\alpha}\right)\\
&= \frac{1}{n}\ell-\frac{\lambda}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)-\frac{r}{2n}\log 2\pi+\frac{r}{2n}\log n\lambda\\
&\quad+\frac{1}{2n}\log|R|_+\\
&= -\ell_\lambda(\boldsymbol{\alpha})-\frac{r}{2n}\log(2\pi/n)+\frac{r}{2n}\log\lambda+\frac{1}{2n}\log|R|_+, \qquad (11)
\end{aligned}$$

where $\ell_\lambda(\boldsymbol{\alpha})$ is the negative penalized log likelihood function given in (3) and λ is the parameter specified by M_j. This equation shows that $\hat{\boldsymbol{\alpha}}$, the solution of the KLR problem (3), maximizes $\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n)$. Thus, by applying the Laplace approximation for integrals developed by Davison (1986) and Tierney and Kadane (1986), we have

$$\begin{aligned}
&\int P\left(\{y_i\}_{i=1}^n|\boldsymbol{\alpha},M_j,\{\mathbf{x}_i\}_{i=1}^n\right)\pi\left(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n\right)d\boldsymbol{\alpha}\\
&=\int\exp(n\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n))d\boldsymbol{\alpha}\\
&\approx\frac{(2\pi)^{(n+1)/2}}{n^{(n+1)/2}\mid J(\hat{\boldsymbol{\alpha}})\mid^{1/2}}\exp\{n\tau(\hat{\boldsymbol{\alpha}}|M_j,\{\mathbf{x}_i\}_{i=1}^n)\}, \qquad (12)
\end{aligned}$$

where

$$J(\hat{\boldsymbol{\alpha}})=-\frac{\partial^2\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\bigg|_{\hat{\boldsymbol{\alpha}}}.$$

Therefore, from (8), (10), (11) and (12) we have

$$\begin{aligned}
\text{GBIC} &= -2\left\{\frac{n+1}{2}\log 2\pi-\frac{n+1}{2}\log n-\frac{1}{2}\log|J(\hat{\boldsymbol{\alpha}})|\right\}-2n\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n)\\
&= -(n+1)\log(2\pi/n)+\log|J(\hat{\boldsymbol{\alpha}})|+2n\ell_\lambda(\hat{\boldsymbol{\alpha}})+r\log(2\pi/n)-r\log\lambda-\log|R|_+\\
&= 2n\ell_\lambda(\hat{\boldsymbol{\alpha}})-(n+1-r)\log(2\pi/n)-r\log\lambda-\log|R|_++\log|J(\hat{\boldsymbol{\alpha}})|.
\end{aligned}$$

Now we compute $J(\hat{\boldsymbol{\alpha}})$. It follows from (11) that the second derivative of $\tau(\boldsymbol{\alpha}|M_j,\{\mathbf{x}_i\}_{i=1}^n)$ with respect to $\boldsymbol{\alpha}$ is identical to the second derivative of the negative penalized log likelihood function $\ell_\lambda(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, which is involved implicitly in the iteration algorithm given in (5). Thus, $J(\boldsymbol{\alpha})=\Phi'H\Phi/n+\lambda R$, and $J(\hat{\boldsymbol{\alpha}})$ is the value of $J(\boldsymbol{\alpha})$ at $\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}$. This completes the proof of the theorem.

3.2. Variable Selection for Support Vector Machines

Note that the variable selection method does not work if $\Phi'H\Phi+n\lambda R$ degenerates. Also note that if σ is specified inappropriately, many values of $K(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_t)$ will be extremely close to zero or one, making $\Phi'H\Phi+n\lambda R$ degenerates. Putting $\tilde{\mathbf{x}}_i=(x_{j_1i},\cdots,x_{j_mi})'$,

$z_1 = \| \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 \|^2$, $z_2 = \| \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_3 \|^2$, $\cdots, z_N = \| \tilde{\mathbf{x}}_n - \tilde{\mathbf{x}}_{n-1} \|^2$ $(N = n(n-1)/2)$, $z_m = \min\{z_r\}_{r=1}^N$, $z_M = \max\{z_r\}_{r=1}^N$, and $\sigma_0 = N \big/ \sum_{r=1}^N z_r$, we may show that

$$\exp\left(-\frac{z_M}{z_m}a\right) \leq K(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_t) \leq \exp\left(-\frac{z_m}{z_M}a\right)$$

for $a = \sigma/\sigma_0$ and any $\tilde{\mathbf{x}}_i$, $\tilde{\mathbf{x}}_t$. Thus, by selecting a approprately, we can always obtain $K(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_t)$ values that are far from zero and one. From these considerations, we propose the following variable selection method for the support vector machine.

First, decide Λ^*, a subset of Λ, such that inverse matrix of $\Phi' H \Phi + n\lambda R$ exists for any $\lambda \in \Lambda^*$.

Second, set the class Σ as $\Sigma^* = \{a\sigma_0,\ 5a\sigma_0, 10a\sigma_0,\ 50a\sigma_0,\ 100a\sigma_0\}$ for $\sigma_0 = N \big/ \sum_{r=1}^N z_r$ and for some selected candidates of a.

Third, compute the GBIC for all elements of $\Lambda^* \times \Sigma^* \times S_c$, where $S_c \subset \{1, 2, \cdots, p\}$.

Fourth, denoting by S_c^* the S_c that attains the minimum GBIC, and employ variables whose suffixes belong to S_c^* as selected variables for the support vector machine.

Fifth, compute the misclassification error rate of the support vector machine with those variables by using leave-one-out cross validation for several combinations of λ and σ, and select λ and σ that attain the minimum misclassification error rate. We have no ideas about the candidates of λ and σ, but selected $\lambda \in \{0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 5,\ 10\}$ and $\sigma \in \{0.1,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}$ by trial and errors in the application below.

Finally, apply the support vector machine with the selected variables and the selected parameters λ, σ to new data.

4. Application

Prostate Stem Cell Antigen (PSCA) is found mainly on the surfaces of prostate cancer cells. Recently it has been shown that PSCA can be also found in pancreatic cancer. PSCA consists of 123 amino acids, whose fragments are called peptides. In order to study which peptides might be related to pancreatic cancer, the antibodies for 58 peptides were measured in 40 patients with pancreatic cancer and 29 patients without pancreatic cancer. Rather than discriminating between patients with and without pancreatic cancer using all peptides, it was considered more important to discriminate between patients using just a few peptides that are strongly related to pancreatic cancer. This motivated us to develop the method of variable selection for support vector machines discussed in the present paper.

We simply denote the names " X1 ",“ X2 ", $\cdots$,“ X58 " for the measurements of antipeptide antibodies of 58 peptides. We anticipated that there were less than five peptides which would be strongly related to pancreatic cancer and we first select 10 candidate peptides from 58 peptides by applying the Lepage test (Lepage (1971)), which is a two-sample non-parametric test for detecting the location and scale differences. We picked up top 10 peptides from the largest inspecting the values of the test statistics. The p-values of these peptides were all less than or equal to 0.0001.

We applied the variable selection method developed in preceding sections to the data with those 10 variables, thus $p = 10$ in this application. Candidate of Λ we considered was $\Lambda^* = \{0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 5,\ 10,\ 50,\ 100,\ 500,\ 1000,\ 5000,\ 10000,\ 50000\}$ and of Σ was the intersection of $\Sigma_a^* = \{a\sigma_0,\ 5a\sigma_0, 10a\sigma_0,\ 50a\sigma_0,\ 100a\sigma_0\}$ for $a \in \{0.001,$

0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50}. We need not compute the GBIC for all combinations of $\lambda \in \Lambda^*$ and $\sigma \in \Sigma^*$. The subsets that give smaller values of the GBIC may be selected by trial and errors. Those subsets that we selected are given in Table 1.

Table 1: Λ^* and a used in the computation

number of variables	Λ^*
$\leq$5	$\{50, 100, 500, 1000, 5000, 10000, 50000\}$
6, 7, 8, 9	$\{0.5, 1, 5, 10, 50, 100, 500\}$
10	$\{0.005, 0.01, 0.05, 0.1, 0.5, 1, 5\}$

number of variables	a
1	50
2	1
3, 4, 5	0.01
$\geq$6	0.001

Figure 1 plots the minimum GBIC and its value for each number of variables. Those values were computed according to the method described in the text. The figure shows that the smallest value among those values of minimum GBIC, i.e. GBIC $= -36.65$, is attained when the number of variables is five. This value is attained when $\lambda = 500$ and $\sigma = 0.00515$, and by {“ X2 ”,“ X14 ”,“ X21 ”,“ X23 ”,“ X29 ”}.

Using these five variables, we next computed the misclassification rates of the support vector machine to decide the optimum λ and σ. As mentioned above, we employed the leave-one-out cross validation technique for combinations of $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ and $\sigma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ and obtained the optimum value as $\lambda = 1$, $\sigma = 0.4$. Table 2 shows the result of classification obtained by leave-one-out cross validation using those optimum values, giving the misclassification error rate 0.217. For the sake of comparison we applied the support vector machine to the same data using all 58 variables. Table 3 shows the result that gives the minimum misclassification rates obtained by leave-one-out cross validation over the combinations of $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ and $\sigma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$, the minimum rate is attained when $\lambda = 0.1$ and $\sigma = 0.1$. From the table the error rate is computed to be 0.232. This shows that we may achieve better discrimination with selected variables than we do using all variables, provided variables are selected appropriately.

Next we apply the RFE, the method of selecting variables for linear support vector machines discussed in the introduction of this paper, to the same data. The RFE ranks all variables in descending order based on some criterion. However, no method for deciding the number of variables is suggested in Guyon et al. (2002). So we performed the leave-one-out cross validation using the same sets for λ and σ as above and evaluated the minimum misclassification error rates of the combinations of variables of size two, three, $\cdots$ by combining the top rank and the second top rank when the size is two, the top rank, the second top rank and the third top rank when the size is three, and so on. Figure 2 plots those minimum misclassification error rates obtained by this method when the size of combination is 2, 3, $\cdots$, 10. The minimum misclassification rate when the size of variable is one is also plotted in the figure. The figure shows that the smallest error rate is 0.246 attained by one variable {“ X30 ”}, and also by six variables {“ X30 ”,

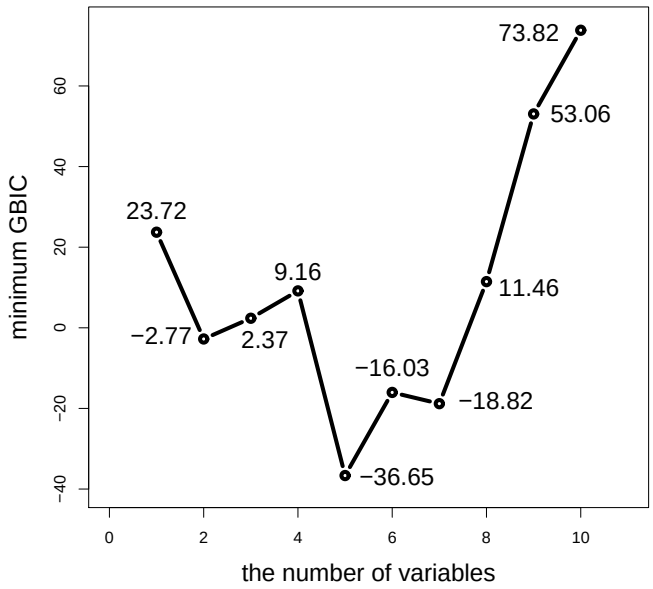

Figure 1: The smallest GBIC value obtained for each number of variables.

Table 2: The results of leave-one-out cross validation of the support vector machine with selected variables

		Results of discrimination		
		non-cancer	pancreatic cancer	total
Data	non-cancer	19	10	29
Label	pancreatic cancer	5	35	40
	total	24	45	69

Table 3: The results of leave-one-out cross validation of the support vector machine with all variables

		Results of discrimination		
		non-cancer	pancreatic cancer	total
Data	non-cancer	19	10	29
Label	pancreatic cancer	6	34	40
	total	25	44	69

“ X39 ”,“ X24 ”,“ X8 ”,“ X41 ”,“ X53 ”}. The smallest error rate (0.246) is larger than the smallest misclassification error rate (0.217) by our method.

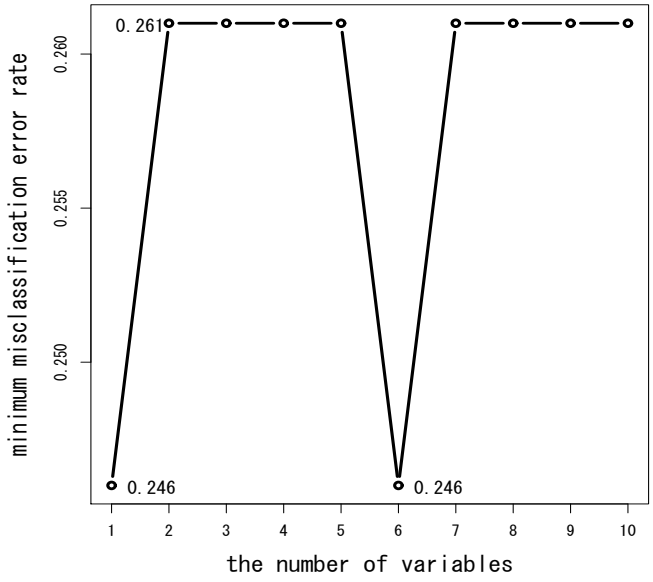

Figure 2: The minimum misclassification error rate of leave-one-out cross validation for each number of variables.

The variables selected by the proposed method and the RFE are exclusive, in particular, the variables selected by our method do not include“ X30 ”, the top ranked variable selected by the RFE. The reason might be considered as follows. It is well known in statistics that the combination of the most effective and the second most effective variables is not necessary efficient for discrimination, in particular, if they are strongly correlated; combination of two weakly correlated variables could often provide better discrimination. Little biological knowledge is available yet on which peptides are related to pancreatic cancer, and it is not easy to give any conclusive results, but we believe that the selection of variables by statistical methods based on Bayes approach such as the one developed in the present paper would provide better results than those methods such as RFE developed in machine learning.

Acknowledgement

The authors would like to thank the reviewers of this paper for their constructive comments that have substantially strengthened this paper.

References

Ando, T., Imoto, S. and Konishi, S. (2004). Adaptive learning machines for nonlinear classification and bayesian information criteria. *Bulletin of Informatics and Cybernetics*, **36**, 147-162.

Davison, A.C. (1986). Approximate predictive likelihood. *Biometrika*, **73**, 323-332.

Guo, G., Li, S.Z., and Chan, K. (2000). Face Recognition by Support Vector Machines. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 196-201.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, **46**, 389-422.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*. Springer.

Joachims, T. (1999). Transductive Inference for text classification using support vector machines. *Proceedings of the International Conference on Machine Learning*, 200-209.

Kobayashi, K. and Komaki, F. (2006). Information criteria for support vector machines. *IEEE Transactions on Neural Networks*, **17**(3), 571-577.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**(1), 27-43.

Krupka, E., Navot, A. and Tishby, N. (2008). Learning to select features using their properties. *Journal of Machine Learning Research*, 9, 2349-2376.

Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika*, **58**, 213-217.

Pontil, M. and Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(6), 637-646.

Scholkopf, B., Sung, K., Burges, C.J., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, **45**(11), 2758-2765.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Wan, V. and Campbell, W.M. (2000). Support vector machines for speaker verification and identification. *Proceedings of the IEEE Workshop Neural Networks for Signal Processing*, 775-784.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. (2000). Feature Selection for SVMs. *In Advances in Neural Information Processing Systems(NIPS)*, **13**.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, **14**(1), 185-205.

Received November 4, 2009
Revised September 30, 2010