Computational Statistics and Data Analysis 53 (2009) 1727–1735

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Informative transcription factor selection using support vector machine-based generalized approximate cross validation criteria

Insuk Sohn [a], Jooyong Shim [b], Changha Hwang [c], Sujong Kim [d], Jae Won Lee [a,*]

[a] Department of Statistics, Korea University, Seoul 136-701, Republic of Korea

[b] Department of Applied Statistics, Catholic University of Daegu, Kyungbuk 712-702, Republic of Korea

[c] Division of Information and Computer Science, Dankook University, Gyeonggido 448-160, Republic of Korea

[d] Skin Research Institute, AmorePacific R&D Center, Kyounggi-do, Republic of Korea

ARTICLE INFO

Article history:
Available online 9 May 2008

ABSTRACT

The genetic regulatory mechanism plays a pivotal role in many biological processes ranging from development to survival. The identification of the common transcription factor binding sites (TFBSs) from a set of known co-regulated gene promoters and the identification of genes that are regulated by the transcription factor (TF) that have important roles in a particular biological function will advance our understanding of the interaction among the co-regulated genes and intricate genetic regulatory mechanism underlying this function. To identify the common TFBSs from a set of known co-regulated gene promoters and classify genes that are regulated by TFs, the new approaches using Support Vector Machine (SVM)-based Generalized Approximate Cross Validation (GACV) criteria are proposed. Two variable selection methods are considered for Recursive Feature Elimination (RFE) and Recursive Feature Addition (RFA). Performances of the proposed methods are compared with the existing SVM-based criteria, Logistic Regression Analysis (LRA), Logic Regression (LR), and Decision Tree (DT) methods by using both two real TF target genes data and the simulated data. In terms of test error rates, the proposed methods perform better than the existing methods.

1. Introduction

The transcriptional regulation's biological mechanism is induced by specific transcription factors (TFs). The TFs are then binding to the transcriptional regulatory elements (TFEs) which located in the *cis*-regulatory region (promoter) of the related genes. The binding sites are present on multiple genes and are sequence specific in their binding sites. By studying these gene promoters, a better understanding of the network interactions will result.

In this paper, we focused on two challenges: (1) the identification of the common transcription factor binding sites (TFBSs) from a set of known co-regulated gene promoters, and (2) the classification of genes that are regulated by the transcription factors (TFs) which have important roles in a particular biological function. These two challenges will advance our understanding of the interaction among the co-regulated genes and intricate genetic regulatory mechanism underlying the particular biological function. The genetic regulatory mechanism plays a pivotal role in many biological processes ranging from development to survival. Many methods have previously been proposed for this objective, including Logistic Regression Analysis model (Wasserman and Fickett, 1998; Krivan and Wasserman, 2001; Liu et al., 2003), Logic Regression

* Corresponding author. Tel.: +82 2 3290 2237; fax: +82 2 924 9895.
E-mail addresses: sis46@korea.ac.kr (I. Sohn), jyshim@cu.ac.kr (J. Shim), chwang@dankook.ac.kr (C. Hwang), sundance@amorepacific.com (S. Kim), jael@korea.ac.kr (J.W. Lee).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.001

analysis (Keles et al., 2004) and Classification and Regression Tree analysis (Jin et al., 2004). Their methods have some advantages of interpretability but need to be improved in terms of their predictive performances.

Support Vector Machines (SVM) has been successfully applied to a number of classification problems (Vapnik, 1998; Scholkopf and Smola, 2002). The drawback of SVM is difficult to interpret (Lee et al., 2006; Koo et al., 2006). A method for improving the interpretability of SVM is the SVM-Recursive Feature Elimination (SVM-RFE) method which has been introduced by Gyuon et al. (2002), and Rakotmamonjy (2003) also derived some ranking criteria relevant to SVM. Recently, borrowing the flexible model building idea of functional ANOVA decomposition, Lee et al. (2006) considered Multicategory Support Vector Machines with ANOVA kernels, and Koo et al. (2006) proposed the Structured Polychotomous Machine (SPM) based on a functional ANOVA decomposition using structured kernels. However, the statistical learning method has been so far largely confined to real valued data. There are many problems in variable selection for discriminating the data with discrete attributes. In this article, we propose the variable selection method of Recursive Feature Elimination (RFE) and Recursive Feature Addition (RFA) using SVM-based Generalized Approximate Cross Validation (GACV) criteria relevant to GACV function proposed by Wahba et al. (1998) to discriminate the data with discrete attributes.

This article is organized as follows. In Section 2, we present materials and methods. In Section 3, we illustrate the performance of the proposed algorithm and those of existing variable selection by using both two real TF target genes data and the simulated data. Finally, conclusion is given in Section 4.

2. Materials and methods

2.1. Materials

We use the library of position weight matrix from TRANSFAC (Matys et al., 2003) and use MatInspector (Quandt et al., 1995) for searching the putative TFBSs. A matrix similarity is calculated by

$$mat\_sim = \frac{\sum_{l=1}^{L} C_j(l) \times score(b, l)}{\sum_{l=1}^{L} C_j(l) \times \max\_score(l)},$$

where L is the length of a matrix, $score(b, l)$ is the value for base b at position l of matrix, $\max\_score\,(l)$ is $\max_{b\in A,C,G,T} score(b, l)$, and $C_j(l)$ is the C_j value of position l of a matrix. $C_j(l)$ is defined as

$$C_j(l) = (100/\ln 5) \times \sum_{b\in A,C,G,T,gap} p(l, b) \times \ln p(l, b) + \ln 5,$$

where $p(l, b)$ is the relative frequency of nucleotide b at position l. A matrix similarity always assumes a value between 0 and 1. For details, see Quandt et al. (1995). We considered the total putative TFBSs with at least one nonzero matrix similarity in promoter sequence of the target genes.

Let x_{ik} denotes a binary indicator for putative TFBS k (=1, ..., p) in gene i (=1, ..., n). A binary indicator is defined as

$$x_{ik} = \begin{cases} 1, & \text{if } mat\_sim_{ik} > mat\_simcut - off \\ 0, & \text{otherwise.} \end{cases}$$

To evaluate the performance of our proposed method in practice, we analyzed two TF target genes data: NF-kB data and Estrogen data.

NF-kB data: We used the promoter sequences of the NF-kB target genes from the work of Liu et al. (2003).The data consist of the promoter sequences of the 58 known NF-kB target genes and 58 mammalian non-immune genes taken by random sampling from the Eukaryotic Promoter Database (EPD) (Perier et al., 2000). The 58 known NF-kB target genes were manually reviewed and selected on the basis of biological literatures. The 58 non-immune genes were not found in any biological literature related to NF-kB family and immune functions. A matrix similarity cut-off is 0.90. We considered a total of 741 putative TFBSs with at least one nonzero PWM in promoter sequence of 58 known NF-kB target genes.

Estrogen data: We used the promoter sequences of the Estrogen target genes from the work of Jin et al. (2004). The data consist of the promoter sequences of the 40 known Estrogen target genes and 40 housekeeping non-target genes (Jin et al., 2004). TFBSs are analyzed by Mat Inspector (Quandt et al., 1995). A matrix similarity cut-off is 0.90. We considered a total of 652 putative TFBSs with at least one nonzero PWM in promoter sequence of the 40 known Estrogen target genes.

2.2. Methods

2.2.1. SVM classifier

In this section, we introduce the basic idea of Support Vector Machines (SVM) (Vapnik, 1995). Consider the training data $\mathbf{D} = \{(x_i, y_i) : i = 1, \ldots, n\}$, where the input $x_i = (x_{i1}, \ldots, x_{ip})$ belongs to some domain $X \subset \Re^p$ and the label y_n, which is the binary class label such that $y_i \in \{-1, 1\}$.

The SVM classifier takes the form

$$f(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b,$$

where the term b is a bias term and $\mathbf{w}$ is a weight vector. Here, the feature mapping function $\phi(\cdot) : \mathbf{R}^d \rightarrow \mathbf{R}^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way.

The optimization problem is defined with a regularization parameter C as fallows:

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

over $\{w, b, \xi\}$ subject to constraints $1 - y_i(\mathbf{w}'\phi(\mathbf{x}) + b) < \xi_i, \xi_i \geq 0,\ i = 1, \ldots, n$. The dual form for the above primal optimization problem (1) is to

$$\text{Maximize } -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_i \alpha_i \tag{2}$$

with constraints $\sum_{i=1}^{n}\alpha_i = 0$ and $\alpha_i \in [0, C]$. The optimal bias and Lagrange multipliers, b and α_i's are obtained by solving the linear equation (2), then the optimal target value for given $\mathbf{x}$ is obtained as

$$f(x) = \sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{3}$$

2.2.2. The variable selection procedure based on SVM

SVM-based procedures for the variable selection are proposed to find a subset of r variables among $p\ (>r)$ variables which maximize the prediction performance. Here, consider the backward sequential selection, in which one starts with all variables and removes one variable at a time until r variables left. At each time, the removed variable is the variable whose removal minimizes the ranking criterion for the given variable k. The ranking criterion of the Support Vector Machines-Recursive Feature Elimination (SVM-RFE), SVM-w2, and SVM-r2w2 are the variations of the leave-one-out error bound L for SVM (Vapnik, 1998). It is known that $L \leq R^2\|\mathbf{w}\|^2$ where R is the radius of the smallest sphere enclosing all the training data.

SVM-RFE: Gyuon et al. (2002) proposed SVM-RFE for the variable selection with ranking criteria as

$$\left|\|\mathbf{w}\|^2 - \left\|\mathbf{w}^{(-k)}\right\|^2\right| = \left|\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i,j}\alpha_i^{(-k)}\alpha_j^{(-k)} y_i y_j K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)})\right|,$$

where $K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)})$ is the (i, j)th element of the Gram matrix of the training data when the kth variable is removed, i.e. $K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)}) = \phi(\mathbf{x}_i^{(-k)})^t\phi(\mathbf{x}_j^{(-k)})$, and α_i is the corresponding solution of Eq. (3). To simplify the complexity of this algorithm, the α_i is supposed to be equal to $\alpha_i^{(-k)}$ even if the kth variable has been removed. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes $\left|\|\mathbf{w}\|^2 - \left\|\mathbf{w}^{(-k)}\right\|^2\right|$.

SVM-w2: Similarly to SVM-RFE, Rakotmamonjy (2003) derived the ranking criteria relevant to weight vector $\|\mathbf{w}\|^2$. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes

$$\left\|\mathbf{w}^{(-k)}\right\|^2 = \sum_{i,j}\alpha_i^{(-k)}\alpha_j^{(-k)} y_i y_j K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)}),$$

where $K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)})$ is again the (i, j)th element of the Gram matrix of the training data when the kth variable is removed. The criterion should be evaluated with the appropriate $\alpha_i^{(-k)}$. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes $\left\|\mathbf{w}^{(-k)}\right\|^2$.

SVM-r2w2: Rakotmamonjy (2003) derived the ranking criteria relevant to the radius/margin bound. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes

$$R^{(-k)^2}\left\|\mathbf{w}^{(-k)}\right\|^2 = R^{(-k)^2}\sum_{i,j}\alpha_i^{(-k)}\alpha_j^{(-k)} y_i y_j K(\mathbf{x}_i^{(-k)}, \mathbf{x}_j^{(-k)}),$$

where $R^{(-k)}$ is the radius of the smallest sphere enclosing the training data when the kth variable is removed. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes $\mathbf{R}^{(-k)^2}\left\|\mathbf{w}^{(-k)}\right\|^2$.

2.2.3. The variable selection using SVM-based GACV criteria

We now propose the variable selection methods of Recursive Feature Elimination (RFE) and Recursive Feature Addition (RFA) using SVM-based Generalized Approximate Cross Validation (GACV) criteria relevant to GACV function proposed

by Wahba et al. (1998). GACV function is believed to be a reasonable estimate for GCKL (Generalized Comparative Kullback Liebler) function (Wahba et al., 1999) which is known as a simple upper bound on the expected misclassification rate.

First of all, we consider the backward sequential selection in which one starts with all the variables and removes one variable at a time until r variables are left. At each time, the removed variable is the variable whose removal minimizes the ranking criterion for the given variable k.

We propose SVM-RFE-GACV for the variable selection with ranking criteria relevant to GACV function

$$\mathrm{GACV}(\mathbf{x}^{(-k)}) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i^{(-k)})_+ + V(\mathbf{x}^{(-k)}),$$

where $V(\mathbf{x}^{(-k)}) = \frac{1}{n}(2\sum_{y_i f_i^{(-k)} < -1} \alpha_i K(\mathbf{x}_i^{(-k)}, \mathbf{x}_i^{(-k)}) + \sum_{-1 \le y_i f_i^{(-k)} \le 1} \alpha_i K(\mathbf{x}_i^{(-k)}, \mathbf{x}_i^{(-k)}))$, $\mathbf{x}^{(-k)}$ is the training data when the kth variable is removed, and $f_i^{(-k)}$ is the estimate with $\{\mathbf{y}, \mathbf{x}^{(-k)}\}$. At each time of the backward sequential selection, the removed variable is the variable whose removal minimizes the GACV function which is the ranking criterion for the given variable k. For simplicity, we call this technique as SVM-RFE-GACV.

We also consider the forward sequential selection in which one starts with one variable and adds one variable at a time until r variables are obtained. We propose SVM-RFA-GACV for the variable selection with ranking criteria relevant to GACV function

$$\mathrm{GACV}(D_0 \cup \mathbf{x}_k) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + V(D_0 \cup \mathbf{x}_k), \quad k = 1, \ldots, p,$$

where D_0 is the empty set and $\mathbf{x}_k$ is the kth variable of training data. We start with the variable whose GACV function is minimal among the p variables. At the $(l+1)$ th step, we consider the GACV functions

$$\mathrm{GACV}(D_l \cup \mathbf{x}_k) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + V(D_l \cup \mathbf{x}_k), \quad k \in \{1, \ldots, p\} \setminus D_l,$$

where D_l is the set consisted of $l-1$ variables selected from the previous steps. Here, one variable which minimizes GACV function is selected. For simplicity, we call this technique as SVM-RFA-GACV.

2.2.4. Other methods

Now we briefly discuss three other methods, the Logistic Regression Analysis (LRA), Decision Tree (DT), and Logistic Regression (LA).

LRA: LRA is used to find the relationship between independent variables and a response variable particularly when the response variable is either binomial or multinomial. For binary response models, suppose $\mathbf{x} = (x_1, \ldots, x_p)$ are the covariates, y is a response variable and $p(y = 1|\mathbf{x})$ is the response probability to be modeled. The logistic model has the form

$$\log\left(\frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})}\right) = \alpha + \boldsymbol{\beta}^t\mathbf{x},$$

where α is the intercept parameter and $\boldsymbol{\beta}$ is the vector of parameters. For details, see Hastie et al. (2002). LRA is implemented using SAS Enterprise Miner. We consider stepwise selection with Akaike's Information Criterion (AIC), as defined by $AIC = -2(\boldsymbol{\beta}) + 2m$, where m is the number of parameters included in the model.

DT: DT has been a quite successful classification tool, for which various algorithms have been proposed for the generation of Decision Trees. DT are widely accepted for predictive modeling because they are easy to interpret and able to model complex input/output associations with an automatic handling of missing values. For details, see Hastie et al. (2002). Many packages are available for implementation of DT which includes CART, C4.5, ID3 and SAS E-Miner. SAS E-Miner is employed for our work. We performs splitting criterion of the Gini Reduction. This is a binary tree with a minimum of 5 observations, 10 observations required for a split search and a maximum of 6 depths of the tree.

LR: LR was introduced by Ruczinski et al. (2003) to construct predictors as Boolean combinations of binary covariates. Let $\mathbf{x} = (x_1, \ldots, x_p)$ be the binary covariates, and let y be a response variable. The logistic logic model has the form

$$\log\left(\frac{E[y]}{(1 - E[y])}\right) = b_0 + \sum_{j=1}^{t} b_j L_j,$$

where L_j is a Boolean expression of covariates $\mathbf{X}$ like $L_j = [(X_1 \wedge X_3^c) \vee X_5]$. The L_j and b_j are estimated simultaneously using a simulated annealing algorithm. For details, see Ruczinski et al. (2003). LR is implemented using the LogicReg package.

Table 1
Mean error rate and one-sided p-value of Wilcoxon test against SVM-RFA-GACV for NF-kB data

Method	Mean error rate	p-value
SVM-RFE	0.1575	0.0069
SVM-RFE-GACV	0.1295	0.5233
SVM-RFA-GACV	0.1281	*
DT	0.3020	0.0000
LRA	0.2880	0.0000
LR	0.3108	0.0000
SVM	0.3724	0.0000

SVM indicates SVM without TFBS selection.

3. Results

We apply our proposed methods to both real TF target genes and simulated data. We selected the optimal TFBS by minimizing the leave-one-out-cross-validation (LOOCV) as

$$\mathrm{LOOCV}(\mathbf{x}^{(-k)}) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f^{(-i)})_+,$$

where $\mathbf{x}^{(-k)}$ is the training data when the kth variable is removed, $f^{(-i)}$ is the estimate of y_i without the ith observation of $\{\mathbf{y}, \mathbf{x}^{(-k)}\}$. We use the diffusion kernel (Kondor and Lafferty, 2002). Let $\mathrm{K}(\mathbf{x}_l, \mathbf{x}_m)$ stands for the (l, m) th entry of a $n \times n$ matrix.

$$K(\mathbf{x}_l, \mathbf{x}_m) = \prod_{i=1}^{n}\left(\frac{1 - \mathrm{e}^{-(x_{il}+x_{im})\beta}}{1 + ((x_{il} + x_{im}) - 1)\mathrm{e}^{-(x_{il}+x_{im})\beta}}\right)^{\delta(x_{il},x_{im})},$$

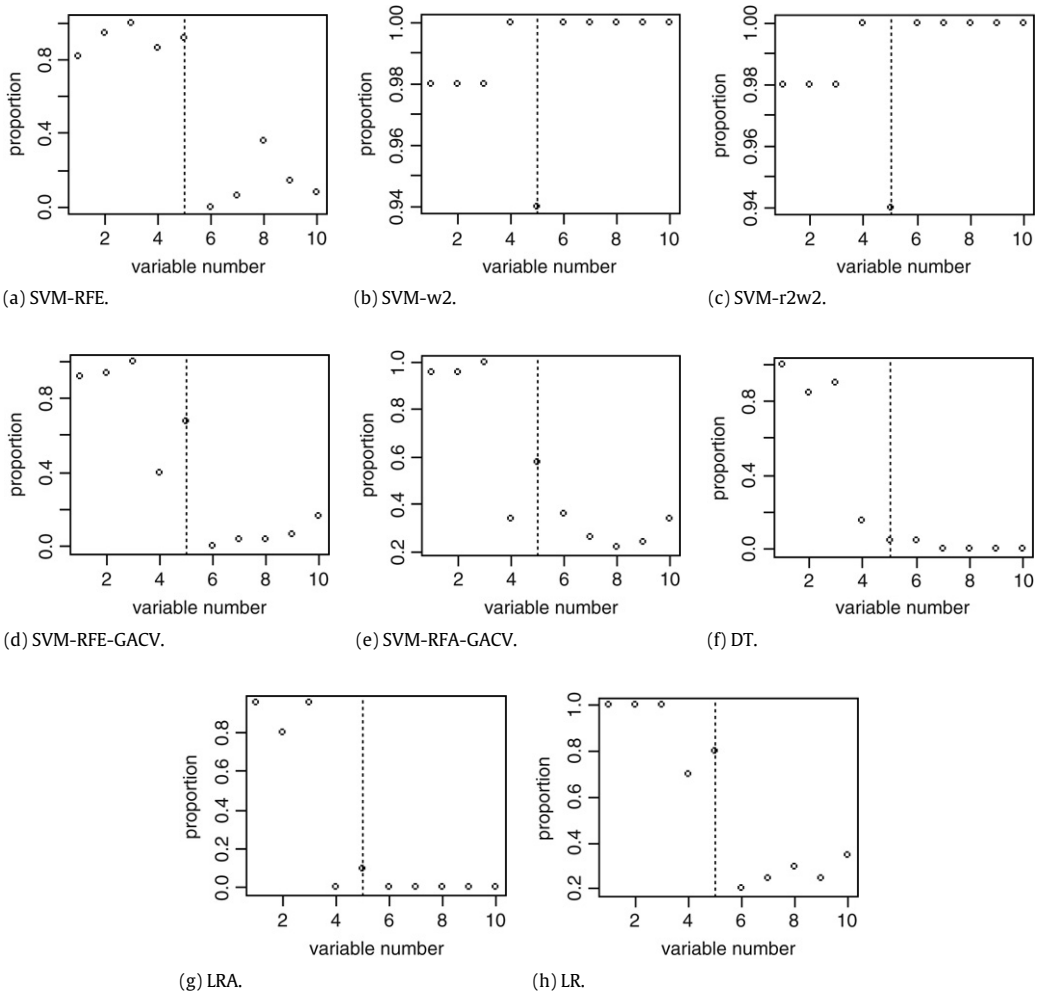
where β is the bandwidth parameter. The bandwidth parameter β of diffusion kernel was tuned by searching over $\{0, 0.1, 0.2, \ldots, 1.9, 2\}$.

3.1. NF-kB data

Before applying the proposed methods to the data, its effectiveness was tested on a miniature data derived from the original data to validate how existing SVM-based criteria, LRA, LR, and DT methods choose the variables for discriminating the data with discrete attributes as in Lee et al. (2006). The miniature dataset of 10 TFBSs, with 116 observations, was constructed as follows. First, using the Chi-square statistic as a measure of marginal association from the original data, we ranked the TFBSs and selected the top 5 TFBSs. The five irrelevant TFBSs were randomly selected among the bottom 5% TFBSs based on Chi-square statistic computed from randomly permuted data. 100 bootstrap samples are drawn from this miniature data. As in Figure 2(a) of Lee et al. (2006), Fig. 1 shows the proportion of selecting each TFBS in 100 replicated classifiers based on the bootstrap samples. TFBSs are labeled as 1–5 for the top 5 TFBSs and 6–10 for the irrelevant 5 TFBSs. In the SVM-RFE method, all the 5 informative TFBSs were selected >82% of the runs while 4 non-informative TFBSs out of the 5 TFBSs were picked up <14% of the runs and 1 non-informative TFBS was picked up 36% of the runs. In the SVM-RFE-GACV method, 4 informative TFBSs out of the 5 TFBSs were selected >68% of the runs while all the 5 non-informative TFBSs were picked up <16% of the runs. In the SVM-RFA-GACV method, 3 informative TFBSs out of the 5 TFBSs were selected >96% of the runs while all the 5 non-informative TFBSs were picked up 36% of the runs. Both SVM-w2 and SVM-r2w2 methods consistently selected the 5 non-informative TFBSs. This result implies that SVM-RFE, SVM-RFE-GACV, and SVM-RFA-GACV methods work well for the informative TFBSs selection, while SVM-w2 and SVM-r2w2 methods work poorly. LRA, DT, and LR methods also work well for the informative TFBS selection.

We compared the number of TFs selected by each method from NF-kB data. Our methods take a considerable amount computing time for TFBS selection because they selected optimal TFBS by minimizing LOOCV, and thus we only used the TFBSs whose chi-square p-value <0.1 from the original NF-kB data. Using LOOCV, the optimal number of TFBS were 5, 7, and 9 for SVM-RFE, SVM-RFE-GACV, and SVM-RFA-GACV methods, respectively, as shown in Fig. 2. The estimated misclassification error rates are 0.190 (22/116 genes), 0.155 (18/116 genes), and 0.1644 (19/116 genes) for SVM- RFE, SVM-RFE-GACV, and SVM-RFA-GACV methods, respectively. SVM-RFE-GACV and SVM-RFA-GACV methods yield slightly smaller error rates than SVM-RFE method. The number of TFBSs selected by LRA, DT, and LR were 6, 4, and 8, respectively.

To investigate relative performances of SVM-RFE, SVM-RFE-GACV, SVM-RFA-GACV, LRA, DT, LR, and SVM without TFBS selection methods for NF-kB data, 116 genes in NF-kB data were randomly divided into 77 training genes and 39 test genes. 39 training genes were used to fit model and the test error rates were computed by 39 test genes. This procedure was repeated 100 times. As in Table 2 of Park et al. (2008), we display mean error rates and one-sided p-value of Wilcoxon test against SVM-RFA-GACV for NF-kB data (see Table 1). SVM-RFE-GACV and SVM-RFA-GACV methods gave smaller mean error rate and SVM-RFE method also gave small mean error rates.



(a) SVM-RFE. (b) SVM-w2. (c) SVM-r2w2.

(d) SVM-RFE-GACV. (e) SVM-RFA-GACV. (f) DT.

(g) LRA. (h) LR.

Fig. 1. The proportion of selecting each TFBS for 100 bootstrap samples. The dotted line delimits informative TFBSs from non-informative ones.

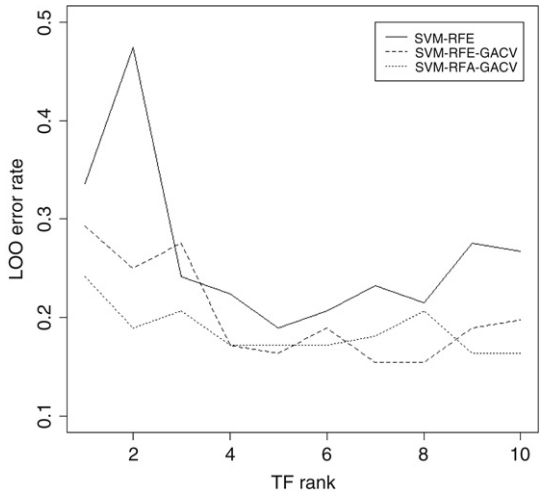


Fig. 2. The choice of the number of TFBSs from NF-kB data. The estimated misclassification error rates by LOOCV against TFBS rank.

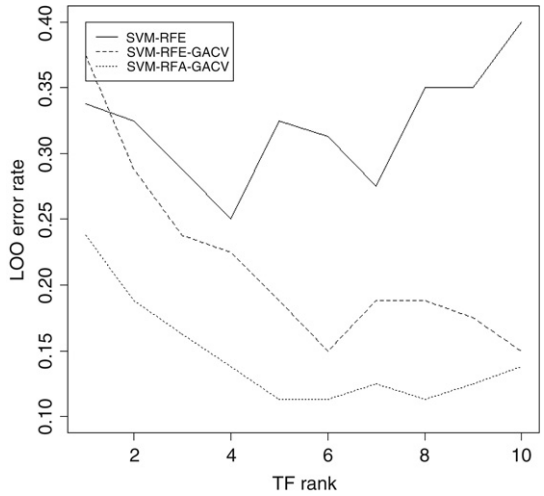


Fig. 3. The choice of the number of TFBSs from Estrogen data. The estimated misclassification error rates by LOOCV against TFBS rank.

Table 2
Mean error rate, standard error and one-sided *p*-value of Wilcoxon test against SVM-RFA-GACV for Estrogen data

Method	Mean error rate	*p*-value
SVM-RFE	0.1302	0.0004
SVM-RFE-GACV	0.0932	0.4407
SVM-RFA-GACV	0.0925	*
DT	0.3384	0.0000
LRA	0.3852	0.0000
LR	0.5133	0.0000
SVM	0.3687	0.0000

SVM indicates SVM without TFBS selection.

3.2. Estrogen data

We compared the number of TFs selected by each method form Estrogen data. Our methods take a considerable amount computing time for TFBS selection because they selected optimal TFBS by minimizing LOOCV, and thus we only used TFBSs whose chi-square p-value <0.1 from the original Estrogen data. Using LOOCV, the optimal number of TFBS were 4, 6, and 5 for SVM-RFE, SVM-RFE-GACV, and SVM-RFA-GACV method, respectively, as shown in Fig. 3. The estimated misclassification error rates are 0.25 (20/80 genes), 0.15 (12/80 genes), and 0.1125 (9/80 genes) for SVM-RFE, SVM-RFE-GACV, and SVM-RFA-GACV methods, respectively. Thus, SVM-RFE-GACV and SVM-RFA-GACV methods yield smaller error rates than SVM-RFE method. The number of TFBSs selected by LRA, DT, and LR methods were 3, 4, and 8, respectively.

To investigate relative performances of SVM-RFE, SVM-RFE-GACV, SVM-RFA-GACV, LRA, DT, LR, and SVM without TFBS selection methods for Estrogen data, 80 genes in Estrogen data were randomly divided into 53 training genes and 27 test genes. 53 training genes were used to fit model and the test error rates were computed by 27 test genes. This procedure was repeated 100 times. As in Table 2 of Park et al. (2008), we display mean error rate and one-sided p-value of Wilcoxon test against SVM-RFA-GACV for Estrogen data (see Table 2). SVM-RFE-GACV and SVM-RFA-GACV methods gave smaller mean error rate and SVM-RFE method also gave small mean error rates.

3.3. Simulation study

We carried out a simulation study to evaluate the proposed methods. We consider normal a matrix similarity of a block compound symmetry (CS) correlation structure as in Jung (2005). A matrix similarity of each TFBS k in the promoter sequences of the ith gene is generated as

$$mat\_sim_{ik} = \phi^{-1}(\delta_{ik} + \varepsilon_{ik}),$$

where $\phi^{-1}(\cdot)$ is an inverse of standard normal distribution function, δ_{ik} is a target effect, and ε_{ik} is a normal random variable with mean 0 and variance 1. Let x_{ik} denotes a binary indicator for putative TFBS k $(=1, \ldots, p)$ in gene i $(=1, \ldots, n)$. A binary indicator is defined as

$$x_{ik} = \begin{cases} 1, & \text{if } mat\_sim_{ik} > mat\_simcut - off \\ 0, & \text{otherwise.} \end{cases}$$

Table 3
The average number of TFBSs selected and the average number of true TFBSs selected from the simulated data

Number of non-target genes	Methods	SVM-RFE	SVM-RFE-GACV	SVM-RFA-GACV	DT	LRA	LR
58	Average number of TFBSs selected	16.25	12.7	13.65	4.00	6.65	7.80
	Average number of true TFBSs selected	7.92	7.65	8.00	2.20	2.60	2.70
	Recovery rate (%)	0.48	0.60	0.58	0.55	0.39	0.34
116	Average number of TFBSs selected	13.55	12.25	13.45	5.60	9.25	7.75
	Average number of true TFBSs selected	7.20	7.10	7.25	2.95	3.15	2.80
	Recovery rate (%)	0.53	0.57	0.53	0.52	0.34	0.36
580	Average number of TFBSs selected	11.20	10.00	7.00	5.70	8.55	7.80
	Average number of true TFBSs selected	6.80	5.80	4.80	3.25	3.90	2.35
	Recovery rate (%)	0.60	0.58	0.68	0.57	0.45	0.30

The recovery rate is defined as the percentage of the number of true TFBSs out of the selected TFBSs.

Table 4
Mean error rates, standard error and one-sided p-value of Wilcoxon test against best method from the simulated data

Number of non-target genes	Method	Mean error rate	p-value
58	SVM-RFE	0.1627	0.0064
	SVM-RFE-GACV	0.1128	*
	SVM-RFA-GACV	0.1333	0.0756
	DT	0.3118	0.0000
	LRA	0.3289	0.0000
	LR	0.3230	0.0006
	SVM	0.4012	0.0000
116	SVM-RFE	0.2086	0.4892
	SVM-RFE-GACV	0.2051	*
	SVM-RFA-GACV	0.2069	0.4675
	DT	0.3061	0.0000
	LRA	0.2535	0.0065
	LR	0.2681	0.0065
	SVM	0.4654	0.0000
580	SVM-RFE	0.0966	0.4162
	SVM-RFE-GACV	0.0947	*
	SVM-RFA-GACV	0.0957	0.3334
	DT	0.1315	0.0000
	LRA	0.1187	0.0002
	LR	0.1053	0.1412
	SVM	0.5526	0.0000

SVM indicates SVM without TFBS selection.

We consider 58 target genes and 58, 112 or 580 non-target genes, which are the same as in the real NF-kB dataset. Target genes are generated by $\delta_{ik} = 1$ and non-target genes are generated by $\delta_{ik} = 0$. We generated 1000 TFBSs, which contained 10% significant TFBSs. A block exchangeable correlation structure was assumed with the correlation coefficient $\rho = 0.3$ and block size 10, i.e. TFBSs are correlated within blocks and uncorrelated between blocks and a matrix similarity cut-off is 0.90, which is the same as in the real NF-kB dataset.

To investigate the relative performances of SVM-RFE, SVM-RFE-GACV, SVM-RFA-GACV, LRA, DT, LR, and SVM without TFBS selection methods for the simulated data, the simulated data are randomly partitioned into the training data (2/3) and the test data (1/3). The training data was used to fit the model and the test error rates were computed by the test data. This procedure was repeated 100 times. We compared the average number of TFBSs selected and the average number of true TFBSs selected from the simulated data, and misclassification rate as in Koo et al. (2006). As in Table 2 of Koo et al. (2006), Table 3 shows the average number of TFBSs selected and the average number of true TFBSs selected from the simulated data. SVM-RFE, SVM-RFE-GACV, SVM-RFA-GACV, and DT methods provide high recovery rate and LRA and LR methods provide low recovery rate. As in Table 4 of Park et al. (2008), we also display mean error rates and one-sided p-value of Wilcoxon test against the best method from the simulated data (see Table 4). SVM-RFE-GACV, SVM-RFA-GACV and SVM-RFE methods gave smaller mean error rates than the other methods.

4. Conclusion

In this paper, we propose the new approaches using SVM-based GACV criteria to identify the common TFBSs from a set of known co-regulated gene promoters and classify genes that are regulated by TF that have important roles in a particular biological function. Our proposed methods using SVM-RFE-GACV and SVM-RFA-GACV gave satisfactory classification performances for both real two TF target genes and simulated data.

In the further study, we will apply the proposed methods to the promoter sequences of both the genes previously identified in the human genome and the published expression profiling experiment data to identify novel NF-kB-responsive immune genes and Estrogen target genes.

Acknowledgements

We are grateful to two anonymous reviewers for their valuable comments. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-352-C00008). The work of Jae Won Lee was supported by the Korea Research Foundation Grant funded by Korean Government (MOEHRD) (R14-2003-002-01002-0).

References

Gyuon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.
Hastie, T., Tibshirani, R., Friedman, J., 2002. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Verlag.
Jin, V.X., Leu, Y.W., Liyanarachchi, S., Sun, H., Fan, M., Nephew, K.P., Huang, T.H.M., Davuluri, R.V., 2004. Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray. Nucleic Acids Research 32, 6627–6635.
Jung, S.H., 2005. Sample size for FDR-control in microarray data analysis. Bioinformatics 21, 3097–3104.
Keles, S., van der Laan, M.J., Vulpe, C., 2004. Regulatory motif finding by logic regression. Bioinformatics 20, 2799–2811.
Kondor, R.S., Lafferty, J., 2002. Diffusion kernels on graphs and other discrete structures. In: Proceedings of the ICML.
Koo, J.-Y., Sohn, I., Kim, S., Lee, J.W., 2006. Structured polychotomous machine diagnosis of multiple cancer types using gene expression. Bioinformatics 22 (8), 950–990.
Krivan, W., Wasserman, W.W., 2001. A predictive model for regulatory sequences directing liver-specific transcription. Genome Research 278, 167–181.
Lee, Y., Kim, Y., Lee, S., Koo, J.-Y., 2006. Structured multicategory support vector machines with analysis of variance decomposition. Biometrika 93 (3), 555–571.
Liu, R., McEachin, R.C., States, D.J., 2003. Computationally identifying novel NF-kB-regulated immune genes in the humana genome. Genome Research 13, 654–661.
Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al., 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Research 31, 374–378.
Park, C.Y., Koo, J.-Y., Kim, S., Sohn, I., Lee, J.W., 2008. Classification of gene functions using SVM for time-course gene expression data. Computational Statistics and Data Analysis 52, 2578–2587.
Perier, R.C., Praz, V., Junier, T., Bonnard, C., Bucher, P., 2000. The Eukaryotic Promoter Database (EPD). Nucleic Acids Res. 28, 302–303.
Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T., 1995. MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23, 4878–4884.
Rakotmamonjy, A., 2003. Variable selection using SVM-based criteria. Journal of Machine Learning Research 3, 1357–1370.
Ruczinski, I., Kooperberg, C., Leblanc, M., 2003. Logic regression. Journal of Computational and Graphical Statistics 12, 475–511.
Scholkopf, B., Smola, J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Massachusetts.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Wahba, G., Lin, X., Gao, F., Klein, R., Klein, B., 1998. The bias-variance tradeoff and the randomized GACV. Advances in Neural Information Processing Systems 11, 620–626.
Wahba, G., Lin, Y., Zhang, H., 1999. Generalized approximate cross validation for support vector machines, or, another way to look at margin-like quantities. Technical Report 1006, University of Wisconsin.
Wasserman, W.W., Fickett, J.W., 1998. Identification of regulatory regions which confer muscle-specific gene expression. Journal of Molecular Biology 278, 167–181.