Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 340678, 8 pages
http://dx.doi.org/10.1155/2013/340678

Research Article

SNP Selection in Genome-Wide Association Studies via Penalized Support Vector Machine with MAX Test

Jinseog Kim,[1] Insuk Sohn,[2] Dennis (Dong Hwan) Kim,[3] and Sin-Ho Jung[2,4]

[1] *Department of Statistics and Information Science, Dongguk University, Gyeongju 780-714, Republic of Korea*
[2] *Samsung Cancer Research Institute, Samsung Medical Center, Seoul 137-710, Republic of Korea*
[3] *Department of Medical Oncology and Hematology, Princess Margaret Hospital, University of Toronto, Toronto, ON, Canada M5G 2M9*
[4] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA*

Correspondence should be addressed to Sin-Ho Jung; sinho.jung@duke.edu

Received 22 May 2013; Revised 14 August 2013; Accepted 22 August 2013

Academic Editor: Wenqing He

One of main objectives of a genome-wide association study (GWAS) is to develop a prediction model for a binary clinical outcome using single-nucleotide polymorphisms (SNPs) which can be used for diagnostic and prognostic purposes and for better understanding of the relationship between the disease and SNPs. Penalized support vector machine (SVM) methods have been widely used toward this end. However, since investigators often ignore the genetic models of SNPs, a final model results in a loss of efficiency in prediction of the clinical outcome. In order to overcome this problem, we propose a two-stage method such that the the genetic models of each SNP are identified using the MAX test and then a prediction model is fitted using a penalized SVM method. We apply the proposed method to various penalized SVMs and compare the performance of SVMs using various penalty functions. The results from simulations and real GWAS data analysis show that the proposed method performs better than the prediction methods ignoring the genetic models in terms of prediction power and selectivity.

1. Introduction

We consider a genome-wide association study (GWAS) on a complex disease. One of the popular study objectives of such study is to predict a binary clinical outcome, such as benign versus malignant and response versus no response with respect to a specific regimen, based on single-nucleotide polymorphisms (SNPs) data. A fitted prediction model will be used to predict the diagnostic or prognostic outcomes of future patients. Recently, penalization approaches incorporating logistic model or support vector machines have been actively proposed to fit prediction models with binary outcomes. These are well known to achieve both predictive accuracy and variable selection simultaneously.

By introducing shrinkage priors of the normal exponential-gamma (NEG) distribution family, Hoggart et al. [1] suggested a stochastic search method for penalized logistic regression models with SNPs. Ayers and Cordell [2] showed that the NEG priors have better performance than other competing penalized methods using simulations, while it is very computing intensive to produce the results. Wu et al. [3] considered lasso-penalized logistic regression [4] with a large number of SNPs and proposed a cyclic coordinate descent algorithm [5] to implement the computation. Kooperberg et al. [6] removed SNPs that had a Hardy-Weinberg P value smaller than 10^{-5} and applied logistic regression models with lasso and Elastic net [7] penalties using a set of SNPs preselected by a cross-validation procedure. On the other hand, Wei et al. [8] proposed selecting SNPs using EigenStrat algorithm [9] and applying the SVM and logistic regression as predictive models. Abraham et al. [10] showed that the two penalized methods, l_1 and Elastic-net SVM, were robust in case/control predictive performance based on simulation studies and real data analyses. These simultaneous analysis methods ignored the genetic models of SNPs [6] or assumed the additive model for all SNPs [6, 8, 10].

The statistical tests such as the Pearson's chi-squared test or the Cochran-Armitage trend test (CATT) are frequently used to test if an SNP is associated with a binary outcome by assuming a specific genetic model. Oftentimes, however, the true genetic model is unknown. We can improve the testing power if we know the true genetic model of an SNP [11]. Toward this end, the test based on the maximum over the three CATT statistics (MAX test) has been presented by several authors [12, 13]. Kim et al. [14] recently proposed a prediction method for time-to-event traits using SNPs and showed that a prediction model based on the best fitting genetic models of SNPs can improve the prediction efficiency. We extend their approach to the prediction of binary outcomes using SVMs.

In this paper, we propose a prediction method combining the MAX test and penalized SVM to predict binary outcome using SNPs. The proposed method consists of two phase procedures: (i) to select candidate prognostic SNPs and identify their genetic models using MAX test, and (ii) to fit a prediction model using the penalized SVM with appropriate scores for the selected SNPs based on their genetic types. We compare the performance of the proposed method using a different penalized SVM method through simulations and a real GWAS data analysis. Each SVM method is combined with MAX test or the general practice ignoring the genetic types of the SNPs.

To facilitate and enable MAX test, we provide the R package called `SNPselect` in http://datamining.dongguk.ac.kr/Rlib/SNPselect which uses the penalized SVM R package [15] to implement SVM with SCAD, l_1, and Elastic Net penalties.

2. Methods

2.1. Penalized Support Vector Machine. Suppose that there are n subjects. For the subject i $(= 1, \ldots, n)$, we have an input vector $x_i \in R^p$ and a class label $y_i \in \{-1, 1\}$. The SVM [16, 17] is to find the optimal hyperplane which separates data points into two classes with the largest margin.

Wahba et al. [18] and Hastie et al. [19] found that the optimization problem of the SVM can be represented as a penalized optimization problem:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)\right]_+ + p_\lambda \left(\boldsymbol{\beta}\right), \tag{1}$$

where $[1 - yf]_+ = \max(1 - yf, 0)$ is called the hinge loss and p_λ is a penality function with regularization parameter λ. The SVM using an l_2-norm, $p_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$, as a penalty function is called the standard SVM or l_2-SVM.

The l_2-SVM has been successfully applied to classification with high-dimensional data such as gene microarrays and SNPs, but it does not select the variables affecting the response class label. For feature selection with l_2-SVM, Guyon et al. [20] proposed the SVM-REF procedure which combines the recursive feature elimination (RFE) with the l_2-SVM. This procedure consists of a two-step procedure using an external gene selection method.

In order to achieve classification accuracy and feature selection simultaneously, variants of SVM have been proposed by replacing the penalty function in (1) with other types of penalty functions, for example, SVM with 1-norm [21, 22], adaptive lasso [23], or smoothly clipped and absolute deviation (SCAD) [24, 25] penalties. The SVM with 1-norm (or l_1-SVM) adapts the lasso (or l_1) penalty, $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, originally proposed by Tibshirani [4] as a practical alternative to l_2 penalty. Due to the l_1 penalty, the l_1-SVM automatically selects variables by shrinking the small coefficients of the hyperplane to exactly zero.

One of major drawbacks of the l_1 penalty is that it tends to select only one variable when there are many correlated input variables in data. To overcome this limitation of LASSO, Zou and Hastie [7] proposed the Elastic Net penalty by combineing l_1 and l_2 penalties:

$$p_\lambda \left(\boldsymbol{\beta}\right) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2. \tag{2}$$

The Elastic Net penalty provides variable selection owing to l_1 penalty, while finding highly correlated variables, called grouping effect. Wang et al. [26] applied the Elastic Net penalty to SVM classification problems.

Fan and Li [24] proposed the smoothly clipped absolute deviation (SCAD) penalty given as

$$p_\lambda \left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} p_\lambda \left(\beta_j; a\right), \tag{3}$$

where

$$p_\lambda \left(\beta; a\right) = \begin{cases} \lambda \left|\beta\right| & \text{if } \left|\beta\right| < \lambda \\ -\dfrac{\left|\beta\right|^2 - 2a\lambda \left|\beta\right| + \lambda^2}{2\left(a - 1\right)} & \text{if } \lambda \le \left|\beta\right| \le a\lambda \\ \dfrac{\left(a + 1\right)\lambda^2}{2} & \text{if } \left|\beta\right| \ge a\lambda. \end{cases} \tag{4}$$

Here, a (>2) and λ (>0) are tuning parameters. Fan and Li [24] showed that the prediction with SCAD penalty is not sensitive to the tuning parameter a and recommended to use $a = 3.7$.

The SCAD yields the same behavior as l_1 for small coefficients β_j, $j = 1, \ldots, p$, but assigns a constant penalty for large coefficients. This property reduces the estimation bias. Fan and Li [24] demonstrate more desirable theoretical properties of the SCAD penalty compared to the l_1 penalty. Later, Zhang et al. [25] proposed the SVM with the SCAD penalty for feature selection.

2.2. Genetic Models for SNPs. Let AA, AB, and BB be three possible genotypes where B is a risk allele for a given SNP. We denote the number of B alleles in a genotype by k; that is, $k = 0, 1$, or 2 if the genotype is AA, AB, or BB, respectively. For a given SNP, the data from n patients are summarized in Table 1.

Let p_k denote the response probability given a genotype $k = 0, 1, 2$. If B is the response allele, the response probability increases as the number of B alleles in the SNP increases; that is, $p_0 \le p_1 \le p_2$. In this paper, we will consider three popular

Table 1: Genotype frequencies.

	AA	AB	BB	Total
Response	r_0	r_1	r_2	r
No response	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

genetic models satisfying this assumption:

(i) *recessive* model: $p_0 = p_1 < p_2$;

(ii) *dominant* model: $p_0 < p_1 = p_2$;

(iii) *additive* model: $p_0 < p_1 = (p_0 + p_2)/2$.

2.3. Trend Test and MAX Test. For testing association between an SNP and a clinical outcome in case-control studies, the statistical tests such as the Pearson's chi-squared test or CATT are frequently used when the true genetic model is known. In this case, the CATT is usually more powerful than Pearson's chi-squared test when $p_0 \le p_1 \le p_2$ [12]. For a single SNP, borrowing the notations of Table 1, the CATT statistic can be written as

$$T_c = \frac{n^{1/2} \sum_{k=0}^{2} c_k \left(sr_k - rs_k\right)}{\sqrt{rs\left\{n \sum_{k=0}^{2} c_k^2 n_k - \left(\sum_{k=0}^{2} x_k n_k\right)^2\right\}}}, \tag{5}$$

where (c_0, c_1, and c_2) is a set of scores assigned to genotypes (AA, AB, and BB) with respect to a specific genotype. The trend test is invariant under a linear transformation with $c_0 \le c_1 \le c_2$, so that the typical choice of these scores is $c_0 = 0$ and $c_2 = 1$, but c_1 can take a different value according to a specific genetic model. From the results of Sasieni [27] and Zheng et al. [12, 28], the optimal choices of c_1 are 0, 1/2 and 1 for the recessive, additive, and dominant models, respectively. Let p_k denote the response probability for genotype group $k = 0, 1, 2$. Under the null hypothesis of no association, $H_0 : p_0 = p_1 = p_2$, T_c approximately follows $N(0, 1)$ for large n.

When the true genetic model is unknown, the test based on multiple CATTs for different genetic models can lead to substantial reduction in statistical power [11] or inflated type I error rate. To address this issue, the test based on the maximum over the three CATT statistics (MAX test) has been proposed by several authors [12, 13]. Let T_R, T_A, and T_D denote the CATT statistics using the scores for recessive, additive, and dominant models, respectively. Based on the three CATT statistics, the MAX test statistic is defined as

$$T_{\max} = \max\left(\left|T_R\right|, \left|T_A\right|, \left|T_D\right|\right). \tag{6}$$

The MAX test has robust properties [29] and is more powerful than the Pearson's chi-squared test [12] when the underlying genetic model is unknown.

Even though one can easily calculate the MAX test statistic from (5) and (6), it is not simple to compute its P value. One approach of obtaining the P value is based on a Monte-Carlo simulation. Under H_0, Zheng et al. [12] showed that (T_R, T_D, T_A) is asymptotically normal with covariances

$$\begin{aligned}
\operatorname{cov}\left(T_R, T_A\right) &= \frac{f_2\left(f_1 + 2f_0\right)}{\sqrt{f_2\left(1 - f_2\right)}\sqrt{f_0\left(f_1 + 2f_2\right) + f_2\left(f_1 + 2f_0\right)}}, \\
\operatorname{cov}\left(T_R, T_D\right) &= \frac{f_0 f_2}{\sqrt{f_0\left(1 - f_0\right)}\sqrt{f_2\left(1 - f_2\right)}}, \\
\operatorname{cov}\left(T_A, T_D\right) &= \frac{f_0\left(f_1 + 2f_2\right)}{\sqrt{f_0\left(1 - f_0\right)}\sqrt{f_0\left(f_1 + 2f_2\right) + f_2\left(f_1 + 2f_0\right)}},
\end{aligned} \tag{7}$$

where f_k denotes the relative frequency of genotype $k = 0, 1, 2$. Thus we can approximate the P value of MAX test based on Monte-Carlo samples from multivariate normal distribution with estimated variance-covariance matrix $\widehat{\Sigma}$ which is obtained by replacing f_k in the above covariances with $\widehat{f}_k = r_k/n_k$ for $k = 0, 1, 2$ ($f_0 + f_1 + f_2 = 1$).

There have been some studies on variants of MAX test for binary clinical outcomes. Zheng et al. [12] developed a robust ranking method, called MAX-rank test. Conneely et al. [30] proposed an efficient P value computation method that is shown to be more accurate than that using permutations by adjusting for correlated test statistics. Li et al. [31] proposed the P-rank test approximating the P value for the MAX test with or without covariate adjustment. Li et al. [32] compared the performance of the MAX-rank and P-rank tests. For more detailed discussions on MAX test, see [11] or [32].

2.4. Classification via SVM with MAX Test. For patient $i = 1, \ldots, n$, let y_i denote the binary clinical outcome taking 1 if responded or -1 if not responded and $(k_{i1}, \ldots, k_{im})$ the encoded data on m SNPs, that is, $k_{ij} = 0, 1, 2$, the number of the risk allele for SNP j $(= 1, \ldots, m)$. To build a classification model with this data set, we propose a method combining a penalized SVM and the MAX test. Our method consists of two-phase procedures: (i) prescreening SNPs and identifying the genetic models for the selected SNPs using the MAX test and (ii) applying the penalized SVM to fit a classification model. Our method can be summarized as follows.

(1) Read in the clinical outcomes $(y_1, \ldots, y_n)$ and SNP data $\{(k_{i1}, \ldots, k_{im}),\ i = 1, \ldots, n\}$.

(2) For SNP j $(= 1, \ldots, m)$,

(a) using the original data, calculate test statistics $(T_{j,R}, T_{j,A}, T_{j,D})$ and their two-sided P values $(p_{j,R}, p_{j,A}, p_{j,D})$ and MAX test statistic $T_{j,\max} = \max(|T_{j,R}|, |T_{j,A}|, |T_{j,D}|)$.

(b) compute the approximate P value of MAX test by Monte-Carlo simulation:

(i) estimate the variance-covarince matrix $\widehat{\Sigma}_j$;

(ii) generate $(t_{j,R}^{(b)}, t_{j,A}^{(b)}, t_{j,D}^{(b)})$ from $N(0, \widehat{\Sigma}_j)$ for $b = 1, \ldots, B$ (=100,000, say);

(iii) approximate the P value for MAX test by

$$p_j = B^{-1}\sum_{b=1}^{B} I\left(t_{j,\max}^{(b)} \geq T_{j,\max}\right), \quad (8)$$

$$\text{where } t_{j,\max}^{(b)} = \max(|t_{j,R}^{(b)}|, |t_{j,A}^{(b)}|, |t_{j,D}^{(b)}|).$$

(3) SNP screening: select J ($\ll m$) SNPs with $p_j < \alpha$ for a prespecified α value, such as 0.01.

(4) For SNP j, identify the genetic model by the smallest P value among $p_{j,R}$, $p_{j,A}$, and $p_{j,D}$.

(5) Assign covariate values $(z_{i1}, \ldots, z_{iJ})$ using the score corresponding to the identified genetic model.

(6) Standardize the covariates; that is,

$$z'_{ij} = \frac{z_{ij} - \overline{z}_j}{s_j}, \quad (9)$$

where $\overline{z}_j = n^{-1}\sum_{i=1}^{n} z_{ij}$ and $s_j^2 = n^{-1}\sum_{i=1}^{n}(z_{ij} - \overline{z}_j)^2$.

(7) Apply the penalized SVM to the response data $(y_1, \ldots, y_n)$ and the standardized covariates $\{(z'_{i1}, \ldots, z'_{iJ}),\ i = 1, \ldots, n\}$.

3. Results

3.1. Simulation Studies. At first, we generate IID $N(0,1)$ random variables $\epsilon_{i1}, \ldots, \epsilon_{im}$ and, for $\rho \in (0,1)$, set

$$x_{ij} = \begin{cases} \epsilon_{ij}, & j = 1 \\ \rho x_{i,j-1} + \sqrt{1-\rho^2}\,\epsilon_{ij}, & j = 2, \ldots, m. \end{cases} \quad (10)$$

Note that $x_{i1}, \ldots, x_{im}$ have an AR(1) correlation structure with autocorrelation coefficient ρ as in [14]. Correlated SNP data are generated by

$$z_{ij} = \begin{cases} 0, & x_{ij} < u_{f_0} \\ 1, & u_{f_0} \leq x_{ij} < u_{(f_0+f_1)} \\ 2, & \text{otherwise}, \end{cases} \quad (11)$$

where u_q denotes the qth quantile of the standard normal distribution. The binary clinical outcome of patient i is generated using response probability p_i which is related to the covariates by

$$\text{logit}\left(p_i\right) = \sum_{j=1}^{m} \beta_j z_{ij}. \quad (12)$$

To consider the cases of uncorrelated or moderately correlated SNPs in our experiment, we set $\rho = 0$ or 0.3. We generate $m = 1000$ encoded SNPs with $(f_0, f_1) = (1/4, 1/2)$ for $j = 1, \ldots, 6$ and $(f_0, f_1) = (1/3, 1/3)$ for $j = 7, \ldots, 1000$. SNPs 1 and 2 have recessive models; SNPs 3 and 4 have dominant models, and SNPs 5 and 6 have additive models, the regression coefficients for these six prognostic SNPs are set at $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0.8$. According to the above data generation scheme, we have generated simulation data sets of size 200, and each data set is partitioned into 2/3 training set and 1/3 test set. For a classification model fitting, the SVM with one of the three penalty functions, SCAD (SCAD-SVM), l_1 (l_1-SVM), and Elastic Net (Enet-SVM), is applied to the SNPs selected using $\alpha = 0.01$. To choose a final classification model, we use 5-fold cross-validation for selecting the tuning parameters. One of the standard practice in the classification model fitting using SNP data will be assuming an equal genetic model for all SNPs. In order to evaluate the performance of the model fitting methods combined with the MAX test, we also have fitted a classification model by assuming one genetic model for all SNPs.

For each model fitting method, we calculate three performance measures such as the number of the selected SNPs, the number of the selected prognostic SNPs by the penalized SVM, and the misclassification error. Here, the selected SNPs are selected by penalized SVM among SNPs after a prescreening step, and the selected prognostic SNPs are the prognostic ones included in the selected SNPs. The misclassification errors are estimated using test data set; that is,

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \text{sign}\,\widehat{f}\left(\mathbf{z}_i\right)\right), \quad (13)$$

where $I(x)$ is an indicator function, $\widehat{f}(\mathbf{z}) = \widehat{\beta}_1 z'_1 + \cdots + \widehat{\beta}_J z'_J$ denote the predicted response score predicted for the test set, and z'_js are standardized covariates in the test set using the means and standard errors calculated from the training set. In order to assess the variability of the experiments, we replicate the whole process 100 times. Table 2 summarizes the three averaged performance measures from our simulations.

When comparing the number of selected SNPs in Table 2, we observe that Enet-SVM tends to select more SNPs but SCAD-SVM selects lower SNPs except for the case of $\rho = 0.3$ and dominant model. In view of different genetic models, the proposed method selects more SNPs when applying l_1-SVM or Enet-SVM. However, the combination of proposed method and SCAD-SVM selects much less SNPs than other combinations. Comparing the numbers of prognostic SNPs, Enet-SVM or l_1-SVM performs better than SCAD-SVM and assuming the proposed method or additive model has good selectivities of the true prognostic SNPs. In results with correlated SNPs ($\rho = 0.3$), Enet-SVM and l_1-SVM with the proposed method result in better selectivities for true prognostic SNPs than those with the additive model. However, the proposed methods can be the worst when SCAD-SVM is used for uncorrelated SNP data. We also compare the misclassification errors. Even if there are a little differences between Enet-SVM and l_1-SVM, Enet-SVM performs better than other penalized methods. SCAD-SVM produces the worst misclassification errors for all cases. We also find that the proposed method has the lowest misclassification errors whatever the penalized SVM method used except the case of applying SCAD-SVM for $\rho = 0.3$. Based on the discussions on the simulation results so far,

Table 2: The result of simulations with 100 replications: selected SNPs and prognostic SNPs indicate the averaged numbers of the selected SNPs and the selected prognostic SNPs, respectively, in the fitted models; standard error is reported in the parentheses.

ρ	Genetic model	Selected SNPs			Prognostic SNPs			Misclassification error		
		l_1	Enet	SCAD	l_1	Enet	SCAD	l_1	Enet	SCAD
0	Proposed	43.10	48.66	20.31	5.11	5.46	3.31	0.1766	0.1567	0.2736
		(0.54)	(0.70)	(0.65)	(0.09)	(0.07)	(0.14)	(0.0048)	(0.0054)	(0.0062)
	Recessive	40.50	41.38	25.28	4.62	4.74	3.66	0.2408	0.2518	0.2912
		(0.56)	(1.11)	(0.66)	(0.10)	(0.10)	(0.14)	(0.0054)	(0.0065)	(0.0047)
	Additive	42.71	45.58	24.12	5.23	5.35	4.07	0.2161	0.2118	0.3272
		(0.49)	(0.87)	(0.91)	(0.08)	(0.08)	(0.18)	(0.0048)	(0.0064)	(0.0076)
	Dominant	41.35	43.46	23.99	4.70	4.86	3.38	0.2457	0.2347	0.2995
		(0.58)	(0.98)	(0.70)	(0.10)	(0.09)	(0.12)	(0.0056)	(0.0063)	(0.0042)
0.3	Proposed	42.92	45.68	19.56	5.12	5.20	3.40	0.1690	0.1541	0.2833
		(0.50)	(0.80)	(0.48)	(0.08)	(0.08)	(0.10)	(0.0049)	(0.0047)	(0.0060)
	Recessive	39.49	41.09	27.23	4.34	4.47	3.03	0.2383	0.2368	0.2741
		(0.58)	(0.87)	(0.59)	(0.10)	(0.11)	(0.03)	(0.0057)	(0.0057)	(0.0019)
	Additive	42.07	43.90	21.89	5.06	5.04	3.74	0.2126	0.2074	0.3338
		(0.52)	(0.96)	(0.67)	(0.08)	(0.08)	(0.08)	(0.0052)	(0.0057)	(0.0056)
	Dominant	39.97	38.56	24.97	4.56	4.29	2.04	0.2502	0.2338	0.2607
		(0.62)	(1.03)	(0.53)	(0.10)	(0.11)	(0.03)	(0.0065)	(0.0059)	(0.0039)

the proposed method combined with Enet-SVM or l_1-SVM could improve the selectivity for true prognostic SNPs and the ablility of precdiction than other methods using a prefixed genetic model.

3.2. Real Data Analysis Example. Kim et al. [33] performed a GWAS using Affymetic Genome-wide Human SNP Arrays 6.0 (San Diego, CA, USA) on 190 patients with chronic myelogenous leukemia (CML). After excluding the SNPs with one missing case and those with the same genotype for all 190 patients, we use 330,353 autosomal SNPs in the further data analysis. The clinical endpoint is the achievement of major molecular response by 18 months to an induction chemotherapy. BCR/ABL transcript levels were measured to determine molecular response to imatinib therapy as described before by Kim et al. [34] and presented using the international scale. Major molecular response (MMR) was defined as <0.1% of the BCR/ABL fusion gene transcript level on an international scale by quantitative PCR. Among the 190 patients, 115 responded.

We randomly partition the CML data into 126 training samples and 64 test samples and then calculate the predictive performance measures for the methods over 100 random partitions. Table 3 summarizes the number of selected SNPs and the mean misclassification errors with their standard errors in parentheses over 100 random partitions. Similar to the simulation results, l_1-SVM and Enet-SVM using the MAX test slightly increase the number of selections, but produce lower misclassification errorr. Among the three penalized methods, Enet-SVM selects the largest number of SNPs but has the lowest misclassification error regardless of the use of the MAX test. However, SCAD-SVM selects the lowest SNPs, while it has poor prediction performances for any assumption for genetic models, which is the same observation in the simulation results.

Table 4 shows the list of 51 SNPs selected commonly by three penalized methods from 126 training samples of one of 100 random partitions. TGFBR1 gene (rs420549, located in 3′UTR region) among 51 SNPs, transforming growth factor beta receptor 1, interacts with TGF beta 1 [35, 36] and TGF beta receptor 2 [37, 38] and is located in 9q22. TGF beta is playing an important role of maintaining the growth and differentiation balance of hematopoietic cells [39, 40] and is known to have bidirectional properties of tumor suppressing and promoting function [41]. TGF-β-FOXO signaling pathway is involved in the maintenance of leukemia-initiating cells in CML, contributing to intrinsic resistance of CML LSCs to tyrosine kinase inhibitor [42, 43]. Accordingly, intrinsic trait of receptor affinity on TGF-β might contribute to different sensitivities to TGF-β; thus, it is potentially explainable that the response to imatinib therapy is dependent on the TGFBR1 genotype.

4. Conclusions

Although the penalized methods have been considered as successful ones for prediction in GWAS, they are still subject to high misclassification error by ignoring the genetic models of prognostic SNPs. In this paper, we proposed a two-phase procedure: (i) carrying out the MAX test for screening out noncandidate SNPs and identifying the genetic models of the selected SNPs at the first stage and then (ii) applying a penalized SVM to the selected SNPs for fitting a classification model at the second stage. We have compared the performances of the proposed method with the conventional methods ignoring the genetic type of prognostic

Table 3: The results of CML data: number of selected SNPs and misclassification error are calculated on average over 100 random partitions; standard error is reported in the parentheses.

Genetic model	Average number of selected SNPs			Misclassification error		
	l_1-SVM	Enet-SVM	SCAD-SVM	l_1-SVM	Enet-SVM	SCAD-SVM
Proposed	70.38	99.80	55.90	0.0737	0.0590	0.1098
	(1.29)	(4.10)	(0.52)	(0.0036)	(0.0062)	(0.0013)
Recessive	55.24	120.46	27.82	0.1184	0.0562	0.2003
	(1.19)	(4.73)	(2.33)	(0.0048)	(0.0051)	(0.0044)
Additive	66.32	120.76	43.50	0.1063	0.0667	0.1530
	(1.12)	(5.00)	(0.27)	(0.0051)	(0.0061)	(0.0026)
Dominant	51.90	91.92	50.90	0.1013	0.0702	0.1663
	(0.89)	(4.81)	(1.30)	(0.0062)	(0.0069)	(0.0044)

Table 4: List of SNPs selected commonly by three penalized methods.

RS ID	Genetic model	P value	RS ID	Genetic model	P value	RS ID	Genetic model	P value
rs3750551	D	0.000510	rs9289221	R	0.000160	rs6621316	A	0.000890
rs3886721	A	0.000040	rs16972014	A	0.000170	rs9890262	R	0.000210
rs2938451	A	0.000000	rs3013492	R	0.000760	rs6779769	A	0.000510
rs6429646	R	0.000050	rs7095688	A	0.000920	rs9502826	D	0.000690
rs6426870	R	0.000230	rs1439691	R	0.000100	rs9896683	R	0.000850
rs4784924	R	0.000100	rs7123207	R	0.000490	rs12907966	D	0.000220
rs8075266	R	0.000190	rs16830058	A	0.000830	rs5979009	D	0.000150
rs4851920	R	0.000130	rs10484180	R	0.000930	rs17157980	D	0.000730
rs9809817	R	0.000190	rs1952096	A	0.000250	rs2865510	R	0.000160
rs342735	A	0.000180	rs2842068	D	0.000600	rs12457620	D	0.000810
rs17066311	D	0.000790	rs420549	D	0.000440	rs4510937	R	0.000390
rs6627852	A	0.000470	rs16822723	A	0.000590	rs8073928	R	0.000510
rs11841074	D	0.000130	rs2492664	A	0.000270	rs10409991	R	0.000290
rs9447907	R	0.000650	rs2029866	R	0.000730	rs1871332	A	0.000150
rs16873423	D	0.000360	rs764515	A	0.000030	rs1264547	D	0.000670
rs315025	A	0.000390	rs11197596	A	0.000240	rs2016016	A	0.000360
rs2355615	A	0.000130	rs9344734	D	0.000690	rs6605081	R	0.000150

SNPs through simulations and real data example. In the simulations, we observed that Enet-SVM and l_1-SVM select more SNPs but have higher selectivities for true prognostic SNPs and lower misclassification errors among the three penalized SVM methods. Combining the proposed method which selects candidate SNPs and estimates their genetic models, we observed that the penalized SVMs except for SCAD-SVM could improve the performances in terms of the selection of the true prognostic SNPs and misclassification errors. Furthermore, the differences of misclassification errors among the three methods with the proposed method become much smaller. Hence, whichever a penalized SVM for model fitting we use, combining it with the MAX test to identify the genetic models of candidate prognostic SNPs could help to improve its performances. We made similar observations from a real data example. Even so, the selection of candidate SNPs could vary according to the choice of a prespecified α; thus, the prescreening by the MAX test could not select a part of true prognostic SNPs. We will consider this point in future work.

Authors' Contribution

Jinseog Kim and Insuk Sohn contributed equally to this work.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2010-0023302).

References

[1] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding, "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000130, 2008.

[2] K. L. Ayers and H. J. Cordell, "SNP Selection in genome-wide and candidate gene studies via penalized logistic regression," *Genetic Epidemiology*, vol. 34, no. 8, pp. 879–891, 2010.

[3] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B*, vol. 73, no. 3, pp. 273–282, 2011.

[5] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[6] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 7, pp. 643–652, 2010.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, vol. 67, no. 2, pp. 301–320, 2005.

[8] Z. Wei, K. Wang, H.-Q. Qu et al., "From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes," *PLoS Genetics*, vol. 5, no. 10, Article ID e1000678, 2009.

[9] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[10] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2, pp. 184–195, 2013.

[11] B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth, "Trend tests for case-control studies of genetic markers: power, sample size and robustness," *Human Heredity*, vol. 53, no. 3, pp. 146–152, 2002.

[12] G. Zheng, B. Freidlin, and J. L. Gastwirth, "Comparison of robust tests for genetic association using case-control studies," in *IMS Lecture Notes-Monograph Series 2nd Lehmann Symposium—Optimality*, vol. 49, pp. 253–265, 2006.

[13] R. Sladek, G. Rocheleau, J. Rung et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.

[14] J. Kim, I. Sohn, D. Son, D. H. Kim, T. Ahn, and S. Jung, "Prediction of a time-to-event trait using genome wide SNP data," *BMC Bioinformatics*, vol. 14, p. 58, 2013.

[15] N. Becker, W. Werft, G. Toedt, P. Lichter, and A. Benner, "PenalizedSVM: a R-package for feature selection SVM classification," *Bioinformatics*, vol. 25, no. 13, pp. 1711–1712, 2009.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1996.

[17] B. Scholkopf and A. Smola, *Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2002.

[18] G. Wahba, Y. Lin, and H. Zhang, "Gacv for support vector machines," in *Advances in Large Margin Classiers*, A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, Eds., pp. 297–211, 2000.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York, NY, USA, 2001.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[21] P. Bradley and O. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Morgan Kaufmann (ICML '98)*, 1998.

[22] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Neural Information Processing Systems 16. Massachusettes*, MIT Press, 2003.

[23] H. Zou, "An improved 1-norm SVM for simultaneous classification and variable selection," in *Proceedings of the Eleventh International Conference on Articial Intelligence and Statistics*, vol. 2, pp. 675–681, 2007.

[24] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[25] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[26] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, 2006.

[27] P. D. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, vol. 53, no. 4, pp. 1253–1261, 1997.

[28] G. Zheng, B. Freidlin, Z. Li, and J. L. Gastwirth, "Choice of scores in trend tests for case-control studies of candidate-gene associations," *Biometrical Journal*, vol. 45, no. 3, pp. 335–348, 2003.

[29] J. L. Gastwirth, "The use of maximin efficiency robust tests in combining contingency tables and survival analysis," *Journal of the American Statistical Association*, vol. 80, no. 390, pp. 380–384, 1985.

[30] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests," *American Journal of Human Genetics*, vol. 81, no. 6, pp. 1158–1168, 2007.

[31] Q. Li, K. Yu, Z. Li, and G. Zheng, "MAX-rank: a simple and robust genome-wide scan for case-control association studies," *Human Genetics*, vol. 123, no. 6, pp. 617–623, 2008.

[32] Q. Li, G. Zheng, Z. Li, and K. Yu, "Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies," *Annals of Human Genetics*, vol. 72, no. 3, pp. 397–406, 2008.

[33] D. Kim et al., "Genome-wide genotype-based prognostic stratification of treatment outcomes following Imatinib therapy in chronic myeloid leukemia in chronic phase," In submission, 2013.

[34] D. H. Kim, J. H. Kong, J. Y. Byeun et al., "The IFNG (IFN-γ) genotype predicts cytogenetic and molecular response to imatinib therapy in chronic myeloid leukemia," *Clinical Cancer Research*, vol. 16, no. 21, pp. 5339–5350, 2010.

[35] R. Ebner, R.-H. Chen, S. Lawler, T. Zioncheck, and R. Derynck, "Determination of type I receptor specificity by the type II receptors for TGF-β or activin," *Science*, vol. 262, no. 5135, pp. 900–902, 1993.

[36] S. P. Oh, T. Seki, K. A. Goss et al., "Activin receptor-like kinase 1 modulates transforming growth factor-β1 signaling in the regulation of angiogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 6, pp. 2626–2631, 2000.

[37] B. Razani, X. L. Zhang, M. Bitzer et al., "Caveolin-1 regulates transforming growth factor (TGF)-β/SMAD signaling through an interaction with the TGF-β type I receptor," *The Journal of Biological Chemistry*, vol. 276, no. 9, pp. 6727–6738, 2001.

[38] M. Kawabata, A. Chytil, and H. L. Moses, "Cloning of a novel type II serine/threonine kinase receptor through interaction with the type I transforming growth factor-β receptor," *The Journal of Biological Chemistry*, vol. 270, no. 10, pp. 5625–5630, 1995.

[39] S.-J. Kim and J. Lettirio, "Transforming growth factor-β signaling in normal and malignant hematopoiesis," *Leukemia*, vol. 17, no. 9, pp. 1731–1737, 2003.

[40] N. O. Fortunel, J. A. Hatzfeld, M.-N. Monier, and A. Hatzfeld, "Control of hematopoietic stem/progenitor cell fate by transforming growth factor-β," *Oncology Research*, vol. 13, no. 6–10, pp. 445–453, 2002.

[41] B. Bierie and H. L. Moses, "Tumour microenvironment—TGFB: the molecular Jekyll and Hyde of cancer," *Nature Reviews Cancer*, vol. 6, no. 7, pp. 506–520, 2006.

[42] K. Naka, T. Hoshii, T. Muraguchi et al., "TGF-B-FOXO signalling maintains leukaemia-initiating cells in chronic myeloid leukaemia," *Nature*, vol. 463, no. 7281, pp. 676–680, 2010.

[43] K. Miyazono, "Tumour promoting functions of TGF-B in CML-initiating cells," *The Journal of Biochemistry*, vol. 152, no. 5, pp. 383–385, 2012.