Published in final edited form as:
Stat Anal Data Min. 2011 June 1; 4(3): 301–312. doi:10.1002/sam.10110.

Sequential Support Vector Regression with Embedded Entropy for SNP Selection and Disease Classification

Yulan Liang[1] and **Arpad Kelemen**[2]

Yulan Liang: ylian001@umaryland.edu

[1]Department of Family and Community Health, University of Maryland, Baltimore 655 W. Lombard Street, Baltimore, MD 21201-1579

[2]Department of Organizational Systems and Adult Health, University of Maryland, Baltimore 655 W. Lombard Street, Baltimore, MD 21201-1579

SUMMARY

Comprehensive evaluation of common genetic variations through association of SNP structure with common diseases on the genome-wide scale is currently a hot area in human genome research. For less costly and faster diagnostics, advanced computational approaches are needed to select the minimum SNPs with the highest prediction accuracy for common complex diseases. In this paper, we present a sequential support vector regression model with embedded entropy algorithm to deal with the redundancy for the selection of the SNPs that have best prediction performance of diseases. We implemented our proposed method for both SNP selection and disease classification, and applied it to simulation data sets and two real disease data sets. Results show that on the average, our proposed method outperforms the well known methods of Support Vector Machine Recursive Feature Elimination, logistic regression, CART, and logic regression based SNP selections for disease classification.

Keywords

Support Vector Regression; Sequential Algorithm; Entropy Measures; Embedded Methods; Sliding Window; Single Nucleotide Polymorphism; Common Complex Disease

1. Introduction

DNA sequencing, the process of determining the exact order of the 3 billion chemical building blocks (called DNA base pairs or nucleotides and abbreviated as A, T, C, and G) that make up the DNA of the 24 different human chromosomes, is the greatest technical challenge in the Human Genome Project (HGP). Achieving this goal has helped to reveal the estimated 20,000–25,000 human genes and the resulting DNA sequence maps are being used to explore human biology and other complex phenomena. Many small regions of DNA that vary among individuals (called polymorphisms) are identified during the HGP, mostly Single Nucleotide Polymorphisms (SNPs). While millions of SNPs have been identified, there is a great need, conceptually as well as computationally, to develop advanced robust algorithms and analytical methods to characterize genetic variations that lean toward more parsimonious solutions with non-redundancy and to identify the target SNPs that are most likely to affect and predict the phenotype traits and ultimately contribute to disease development.

Exploiting information redundancy due to associations among SNPs potentially reduces the efforts in terms of time and cost for genomic association studies. To investigate the correlation structures among SNPs for SNP-disease association analysis, HapMap has

documented tagSNPs for common variation using pairwise r^2 method, which has largely reduced the redundancy between SNPs [1–3]. Linkage Disequilibrium (LD) based methods for selecting a maximally informative set of SNPs for association analyses are now widely being used [4]. For example, Zhang and Jin [5] introduced a tagSNPs criterion based on pair-wise LD and haplotype r^2 measure for case control association studies. Without pre-known disease status or outcome, several unsupervised methods that are based on LD and haplotype block concept are developed for SNP redundancy, which include clustering and graph methods [6,7], principal component analysis [8,9], and haplotype pattern mining [10].

Comparably, correlating SNPs with phenotypic differences or known disease outcomes has been also one of the major efforts for discovering and selecting the disease associated SNPs. In this case, the SNP selection problem has been addressed via both parametric and nonparametric methods. Within the parametric framework, non-model based methods either directly define a test statistics or statistic measures including Mantel statistic [11]; Score statistic [12,13]; Scan statistic [14]; Sliding window approaches [15]; Weighted-average statistic [16]; Minimum description length [17,18]; and Entropy-based measure [19,20].

For example, a sliding window approach developed by Neale and Sham [15] incorporates the ordering of SNPs on the chromosome and combines p-values from m independent tests using $\chi^2 = -2\sum_{i=1}^{m}\log(p_i) \sim \chi^2_{2m}$ in each sliding window, where p_i is the p-value of association between SNP_i and disease outcome variable, and m is the number of SNPs in the sliding window. The test statistic χ^2 has a chi-square distribution with 2m degrees of freedom, and is used to decide if this window is significantly associated with the disease. The implicit assumption of the sliding window approach is that the SNPs are equally spaced and the number of SNPs (m) in each window is predefined. Furthermore, Sun, et al. [21] developed a chromosomal scan statistic approach, which not only considers the order of SNPs on the chromosome but also the distance between them. This approach assumes that SNPs are randomly distributed on the chromosome with a Poisson distribution. The lengths between two adjacent SNPs have an exponential distribution and the sum-of-lengths between SNPs follows a Gamma distribution. This approach includes (i) Identifying SNP clusters; (ii) Identifying SNP clusters with significant disease association by scan test.

Compared to the above non-model based approaches, model based approaches within the parametric framework for SNP selections, e.g. via testing the significance of model coefficients, etc. have advantages of incorporating the environmental factors and adjusting for potential covariates in the modeling for the associations. They are more precise and powerful when models and distribution assumptions are correctly specified, but suffer from bias when they are not. For example, Logistic Regression with Likelihood Ratio test (LRLR) is a commonly used approach for SNP selection in case control studies. Recently, Schwender and Ickstadt [22] proposed logic regression based identification of SNP interactions for the disease outcome in case-control study and proposed two measures for quantifying the importance of SNP interactions for classification. This method is called LOGICFS. In comparison with some well-known classification methods, such as CART [23] and Random Forests [24], LOGICFS has shown a good classification performance when applied to SNP data. However, since it focuses on the categorical SNP variable's interactions, similar to multidimensional data reduction methods, there is a drawback of such method when one intends to incorporate the continuous environmental variables.

Compared to parametric approaches, nonparametric approaches in the statistical learning framework have advantage of robustness to the model and distribution misspecification. They do not require any a priori assumptions about the genetic inheritance model of the disease. Furthermore, multiple testing and "curse of dimensionality" issues of SNP selection

could be transformed into embedded optimization problems by minimizing optimization functions or information criteria, which are more computationally efficient for high dimensional data [25–27]. The variable sets (e.g., SNPs) selected within the minimum redundancy - maximum relevance nonparametric framework based on information gain may represent broader characteristics of outcomes than those obtained through parametric statistical testing methods [28]. They are more robust, generalize well to test and unseen data, and lead to significantly improved classifications and predictions [29].

However, searching for an optimal variable subset in high dimensional space based on the definitions of variable relevance and redundancy is combinatorial in nature, and is a challenge since it is obvious that exhaustive or complete search is prohibitively expensive (with a large number of SNPs and environmental variables). Therefore, it is an appealing and challenging new area to develop nonparametric embedded learning methodology within predictive models to search for an optimal SNP set iteratively with minimum redundancy [30]. At the current stage, most SNP-disease studies focus on the identifications of SNPs through association analysis.

Support Vector (SV) regression models belong to a family of generalized linear classifiers that simultaneously minimize the empirical classification error and maximize the geometric margin between classes [31–34]. Support vector models map predictor vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Supervised support vector regression models [33] have shown good generalization performance for prediction and classification in many applications [35,36]. However, due to the complexities, SV models can be abysmally slow, inherently unstable, tend to overfit the data and therefore the results may not generalize well to cross validation or unseen data.

Recent works have shown that some of these drawbacks can be overcome by incorporating a wide range of learning algorithms into SV models, e.g. by connecting regularization theory [37]. For example, Liang and Kelemen [38] presented Bayesian regularization algorithm with Automatic Relevance Determination in neural network to exclude a large number of irrelevant variables for predictions. Wang, et al. [26] designed the elastic-net penalty, a mixture of the L2-norm and the L1-norm penalties to perform automatic variable selection in a doubly regularized support vector machine. Guyon, et al. [39] proposed the Support Vector Machine Recursive Feature Elimination (SVMRFE) algorithm to recursively classify the samples with SVM, and they applied it to gene expression data analysis for the selections of the genes according to their weights in the SVM classifiers.

Therefore, the overall goal of this paper is to develop a novel nonparametric embedded learning method within predictive models, called “Sequential Support Vector Regression with Embedded Entropy” for the selection of a minimal optimum subset of SNPs and environmental variables that will have the highest predictive accuracy. The proposed method can detect the disease susceptibility effects while retaining most of the information. The paper is organized as follows. In section 2 we provide the details of the proposed models and algorithms. In sections 3, we illustrate our method with a small scale working example, a simulation study, and an application to complex SNP-disease data sets. We compare the performances of the proposed methods and some popular SNP selection models discussed earlier, including sliding window approach, scan test, logistic regression, logic regression, Naive Bayesian classifier, Classification and Regression Tree (CART), and Support Vector Machine Recursive Feature Elimination. We end our paper by discussions of the proposed methods and point to future work in section 4.

2. Methods

2.1 Support Vector Regression Models with Regularization for SNP Selection and Disease Prediction

We formulate our SV model for SNP selection and disease prediction as follows. Assume we have identical independently distributed (i.i.d.) samples { (x_i, y_i) i = 1, 2, . . ., n}. The *i*'th observed binary response $y_i \in \{-1,1\}$ follows binomial distribution, which can be extended to multinomial distribution if y_i includes three or more categories. Let x_i be a p-dimensional vector with G predictors, $x_i = (x_{i,1}, x_{i,2},\ldots, x_{i,G})$. Both categorical and continuous predictors are allowed in the model. We intend to train a SV regression model using samples (x_i, y_i) and to find an optimal subset of SNPs that can predict the disease outcomes (such as case and control) while minimizing the total prediction errors. Therefore, we need to find a hyper-plane that can maximize the margin between two classes. In a SV model, the predictor x_i is first mapped onto an m-dimensional variable space using a linear or nonlinear mapping function; for instance $g_j(x_i)$, and then a linear model is constructed in this variable space, which can be given as follows

$$f(x_i)=\sum_{j=1}^{J} w_j g_i(x_i)+b \tag{1}$$

where $g_j(x_i)$, j =1,…, J denotes a set of linear or nonlinear transformations, and b is a scalar offset, $w = (w_1, w_2,\ldots w_J)$ is regression coefficient or weight vector.

In eq. (1) only the samples (x_i, y_i) which are closest to the separating boundary or decision surface have non-zero *w*'s (called Support Vectors: SV's). Therefore, the decision function $f(x_i)$ is a linear or nonlinear combination of only the SV's. In this study, we focus on linear support vector regression models. Therefore, the decision function $f(x_i)$ is a linear function of these non-zero support vectors.

In an ε-SV model, the goal is to find a decision function $f(x_i)$ that has at most ε deviation from the actually obtained target y_i for all the training data, and in the meanwhile is as flat as possible to reduce the model complexity and to avoid overfitting. To control the model complexity and to achieve efficient automatic variable selection for high dimensional data, we design the following optimization function by including L1-norm penalty in Lagrange multipliers for regularization:

$$\min\{\frac{1}{2}\parallel w \parallel +\lambda\sum_{i=1}^{l}(\xi_i+\xi_i^*)\} \tag{2}$$

$$\text{subject to:}\quad \begin{array}{l} y_i - f(x_i) - \xi_i^* \le \varepsilon \\ f(x_i) - y_i - \xi_i \le \varepsilon \\ \xi_i^*, \xi_i \ge 0 \end{array} \tag{3}$$

where ξ_i^*, ξ_i are slack variables that define the "soft margin" to measure the deviation of training samples outside the ε insensitive zone; and λ is the regularization parameter that determines the trade-off between model complexity (flatness) and the degree to which deviations larger than ε are tolerated in the optimization function. By tuning λ in eq. (2) we can achieve maximal prediction accuracy.

One important property of using L_1-norm penalty in eq. (2) rather than L_2-norm is its greater tendency of producing some large fitted coefficients and leaving others at 0. This allows us to automatically identify relatively small proportion of SNPs with nonzero values that are predictive for the disease outcomes [40]. Furthermore, it also benefits the reduction in the estimated coefficients' variance and produces more accurate estimates given high dimensional data challenges.

2.2 Sequential Support Vector with Minimum Entropy Selection Algorithm

Due to high correlations and redundancy among SNPs, selection instability is a concern [41]. To deal with the redundancy and correlations among SNPs, we propose the entropy information measure together with a sequential procedure embedded in the ε-SV model for the selection of a minimal optimum subset of SNPs that will have the highest predictive accuracy while retaining most of the information. The sequential procedure starts from a SV model with randomly selected SNPs; if the candidates SNPs have equal prediction accuracy, then the entropy measure is used as “tie-breaker” to refine the selection and to reduce the redundancy.

To further improve the computational efficiency and to speed up the selection given the high dimensionality with the affordable sample size, we adopt the sliding window strategy instead of one SNP at a time. This strategy helps to incorporate the LD patterns, SNP interactions and may combine information both locally and in long range over the chromosome [42,43]. Here, each sliding window includes a set of consecutive SNPs across a chromosome, and the windows are adjacent, but not overlapped. The size of the window is pre-determined; for example, we may use 5, 10 or 50 SNPs once per sliding window. The sliding window with the best prediction accuracy will be chosen first. If there is a tie between two windows (equal prediction accuracy), then the Shannon entropy [44,19] is used as "tie-breaker". With the same principle, the entropy measure is used to assess the information gain and loss within each window considering individual SNPs. After that, other sliding windows are sequentially tested with the SV model, one at a time. The one that improves the prediction most will be subsequently chosen (with tie breaker if needed). The procedure stops when including an additional window cannot improve the prediction. The prediction accuracy is used as the stopping criteria.

The computation of the entropy for each SNP within each window is as follows: assuming a discrete random SNP_j having three possible genotypes coded as 0, 1, or 2 with 0 and 2 representing homozygote and 1 representing heterozygote; the genotype frequencies for SNP_j are denoted as p_{0j}, p_{1j}, $p_{2j} = 1 - p_{0j} - p_{1j}$. The entropy for SNP_j is defined as:

$$H(SNP_j) = -[p_{0j}\log p_{0j} + p_{1j}\log p_{1j} + (1 - p_{0j} - p_{1j})\log(1 - p_{0j} - p_{1j})] \quad (4)$$

We define the entropy of a sliding window, e.g. the i’th sliding window with K_i SNPs as follows:

$$\sum_j H(SNP_j) = -\sum_{j=1}^{K_i}[p_{0j}\log p_{0j} + p_{1j}\log p_{1j} + (1 - p_{0j} - p_{1j})\log(1 - p_{0j} - p_{1j})] \quad (5)$$

For improving the computational efficiency, entropies are calculated only for those candidate SNPs, which are selected based on best prediction performance in the ε-SV model with regularization. The calculated entropies from different sliding windows are used to compare the information gain or loss if they have ties in prediction accuracy.

The details of the sequential support vector with minimum entropy selection algorithm are given below:

Step 1. Train a Support Vector regression model with the given parameters (λ, ε, window size, etc.).

Step 2. Choose the candidate window with the highest prediction accuracy as the next set of predictors. If there is a tie between two or more sliding windows, then calculate the Shannon entropy based on Eq. (5) to rank the sliding windows. Choose the lowest entropy window.

Step 3. To further extract the maximal amount of information within C_i, calculate the entropy of each SNP_j within C_i based on Eq. (4), and their summations based on Eq. (5). Keep only the SNPs that provide entropy decrease in C_i (denoted as s_i). Discard the rest.

Step 4. Repeat *Steps 1–3* to produce $S_{N+1} = (s_1, s_2, \ldots, s_N, s_{N+1})$ by adding S_{N+1} to $S_N = (s_1, s_2, \ldots, s_N)$ until the prediction accuracy increase is less than ε.

The choice of variable (s_{N+1}) to be added and the total number of variables (N) to be included in the model are not predetermined, but based on the best prediction performance and entropy gains between and within sliding windows. Also note that the above algorithm is a forward selection method. Instead of adding variables at step 5, we can modify the algorithm by starting from all variables and deleting variables, which is similar to backward elimination.

3. Applications

3.1 A working example

3.1.1 Myocardial Infarction (MI) data set—To illustrate our proposed model and algorithm described in section 2, we implemented and applied them to cardiovascular disease (CVD) with Myocardial Infarction (MI) case and control data set that was collected in Western New York. CVD including MI is one of the leading causes of death and disability in the western world. Numerous clinical and epidemiological studies have shown that CVD may be the result of common environmental exposures and potential genes that may regulate the individual response to these exposures. The identification of SNPs and environmental variables that influence the risk of diseases and the prediction to the disease outcomes based on the selected variables with the proposed model and algorithm are main purposes of this application.

The study sample was collected from residents of Erie and Niagara counties in New York state and all were within age range 35 to 69 years. There were 614 white male patients with MI matched with 614 control males by age (+/− 5 years) and smoking habits; 206 white pre and postmenopausal females with MI were matched with 412 control females by age (+/− 5 years), menopausal status, years since menopause (+/− 2 years), and smoking habits. The variables in the data set include 29 environmental variables, such as smoking status, menopausal status, blood pressure, blood cholesterol, body mass index, drinking status, etc. and 2 protein variables (ACHMN and CALMEA) that are known to be related to the disease.

To limit this example to a low number of variables, we preselected SNPs from the well known Seattle web site (http://pga.mbt.washington.edu/) using candidate gene approach. We included 31 SNPs in 9 genes as follows, plus the above 31 environmental variables to test our proposed approach: IL-1 beta gene: rs1143634, rs16944, rs3917354, rs3917356; IL-6

gene: rs2069825, rs1818879, rs1548216, rs1800795; MMP3: rs522616, rs595840, rs602128, rs680753; TF: rs1324214, rs1361600, rs3354, rs391763…, and so forth.

Figure 1 displays a subset of our MI data for one of the selected gene (IL-1 beta), which is known to be related to metabolic pathways and involved in MI. The figure shows 62 SNPs (columns) for this gene and 47 subjects (rows). For each SNP, there are three genotype categories as we have discussed in section 2.2: homozygous common marked with blue color (coded as 0), homozygous rare marked with yellow (coded as 2), heterozygous marked with red (coded as 1) and missing or unknown are marked with gray. In the figure we can see that many SNPs are strongly correlated, which indicate that the genetic information is highly redundant. One of the important questions is: can we reduce the number of SNPs needed to be genotyped to save on the cost, while preserving the good predictive performance for disease outcomes?

To evaluate the performance of our proposed model and algorithm without bias and to examine if the sequential procedure is asymptotically consistent for both selections and predictions, Cross-Validation (CV) is used for optimizing the number of selected variables in the context of building prediction model, which also helps to avoid over-fitting. Standard 10-fold Cross-Validation was performed. In our MI data set, there are 820 cases and 1026 controls. The sample was randomly divided into 10 equal subsets for both case and control. One subset including both cases and controls was randomly picked as first CV set; the rest was used as training data. The training data was presented to the sequential algorithm to build the model and selection of the optimal set of SNPs was performed. CV sets were applied to the best prediction model, which was associated with the highest training accuracy. SNPs were selected and the average CV accuracies from 10 CV sets were calculated to evaluate the performance, which is reported in our result section below.

3.1.2 Results—Table 1 lists the average 10-fold CV accuracies and the standard deviations for our proposed Sequential Support Vector Regression with Minimum Entropy Selection method (SSVRMES). The results of other popular approaches discussed earlier: SVMRFE, LRLR, CART, and LOGICFS are also presented. The penalty parameter λ values of our SSVRMES and SVMRFE were tested in the range of 0.0–1.0 by increments of 0.01, and the final value was chosen based on the lowest 10-fold CV error rate. The parameters for the stopping rule were 0.001 for the prediction accuracy and 0.01 for entropy gain or loss. The window sizes evaluated for SSVRMES were varied from 5 to 15. The numbers of nodes for CART were tested from the smallest tree containing only 10 seed nodes and allowed to grow until the classification or prediction performance would not change in any profound way. All the parameters from the compared approaches were tested first and selected based on best 10-fold CV performance for each, and then they were further applied to the data sets for comparing and evaluating their performances.

According to the results, our proposed SSVRMES is superior to these four methods. Note that LOGICFS is designed only for categorical SNPs data and the 10-fold CV accuracies of LOGICFS presented in Table 1 were obtained only based on SNPs data. This might be a reason it performed worst since MI is a complex disease, which involves both genetic and environmental factors, and their interactions. To better visualize the performance of our proposed SSVRMES, we recorded and examined the prediction accuracy changes of the constructed SV model as the number of variables selected from SSVRMES. We also implemented and compared SSVRMES to the simpler Naive Bayesian classifier (denoted as NBCPT) and a more complex classifier CART and applied both to our MI data. Figure 2 displays the 10-fold CV accuracies of NBCPT, CART and our SSVRMES for MI data as the number of dimensions increases from 1 to 30 for two typical runs. The results of SVMRFE, LOGICFS and LRLR are also displayed in the same figure.

Results show that the performance of SVMRFE is relatively close to that of our approach, but on the average, our method outperformed SVMRFE. These results demonstrate our rational reasons for our proposed model and algorithm: 1) Although SVMRFE was designed for high dimensional data with embedded learning and optimization strategy and has shown good performance on problems of gene selection for microarray data, the information redundancy and highly correlated variables are not considered; 2) Besides embedded optimization strategy, our algorithm embedded additional entropy measure that helped to remove the redundant variables that may hurt the prediction performance while SVMRFE did not. Our results confirm that better generalization performance can be achieved if embedding the minimum redundancy versus maximum relevance in supervised learning framework to achieve scarcity.

The 10-fold CV accuracy may not be independent of one another. Instead, there are positive correlations among them, and the standard deviation may be under-estimated. Therefore, we also run a 3-fold CV, where there is less correlation present. Results show that the CV accuracy of standard deviations is consistent with 10-fold CV. Moreover, although the first set of SNPs selected was varied; the variations or standard deviations in prediction error of our proposed SSVRMES are consistently better than those of other approaches, even with different window sizes (e.g., 5 or 10). However, the selected SNPs in the final data set were slightly different with different window sizes. To further evaluate our proposed model and algorithm, we performed a simulation study and applied the proposed approach to a larger SNP-disease data set: North American Rheumatoid Arthritis Consortium (NARAC) data, which are discussed next.

3.2. Simulation study

To further evaluate our proposed model and algorithm and compare with existing approaches, we conducted a simulation study using simulated SNP data sets that mimic some of the genetic and epidemiological features of the North American Rheumatoid Arthritis Consortium Data [45]. Detailed descriptions of the simulated data, data generation scheme, and the parameters are given by Witte, et al. [45] and the Genetic Analysis Workshop 15 - problem 3. A large population of nuclear families was generated that contains 2 million sibling pairs, and the RA affection status was determined for everyone from a complex model [48]. For our simulation purpose, a random sample of 1500 families was selected from families with an affected sib pair (case group) and another random sample of 2000 families was selected from families where no member was affected (control group). 2000 simulated SNPs on nine unobserved trait loci were selected from 17,820 SNP markers on chromosome 6. These generated SNPs had similar properties, such as LD, to those observed in the real RA data [47]. Since GAW 15 - problem 3 provided 100 replicates of simulated SNP and covariate data sets which were modeled after real RA data, these data sets were ideal for evaluating our proposed model and algorithm for SNP selection and disease classification purposes.

We extracted tuning data from 100 replicates generated from the same simulated population. The penalty parameter λ values were tested in the range of 0.0–1.0 by increments of 0.2, and the final value was chosen based on the highest classification accuracy the tuning set. The parameters for the stopping rule were 0.001 for the prediction accuracy and 0.01 for entropy gain or loss. The window sizes for SSVRMES were varied from 5 to 15. To reduce the selection bias and for unbiased evaluation of the performance, we conducted 10-fold CV to estimate the classification accuracy of the resulting model. Training was performed 10 times with the same parameter settings, but each time using a different 9/10 of the data for training and 1/10 of the data for cross validation. Table 2 reports the classification accuracies from our proposed SSVRMES for the simulated data sets with different λ values and window sizes of 5 and 15.

The numerical summary from Table 2 for the simulation study show that different λ values lead to different numbers of selected SNPs and as λ increases the number of selected SNPs decreases. This indicates that the penalty parameter did affect the SNP selection and also the corresponding classification accuracy. This suggests that with a proper penalty value, the selected SNPs and the constructed classifier will lead to less over-fitting and increased prediction accuracy. According to our simulation study, with a window size of 5 (SNPs), an average of 87 SNPs were found with an average of 78% classification accuracy with our proposed SSVRMES. As the window sizes were varied (e.g., from 5 to 15), the variation was low in the classification error rates based on Table 2. However, increasing the window size tended to reduce the number of selected (relevant) SNPs. These may be due to the combination smoothing effect on the high/low LD in the studied region.

The simulation data sets were also analyzed with publicly available software SVMRFE, CART and LRLR for quantitative performance comparison with our SSVRMES. The penalty parameter λ values of SVMRFE were tested similarly to SSVRMES. However, CART and LRLR do not have penalty parameter λ and window sizes. Therefore, we only compared the classification accuracy and the selected SNP sets from each approach. All model parameters from all compared approaches were tested first, similarly to our working example discussed in section 3.1. The final model of each approach was chosen based on the best 10-fold CV accuracy. The performance comparison in SNP selections obtained by different approaches shows that there is a relative consistent consensus among the various methods as to the set of predictable SNPs (e.g., all indicated SNP0688 and SNP1240 as being important) relevant to the disease outcomes, which is encouraging despite the difference of error rates. For all compared approaches, about 4–8% of the SNPs (80–160 SNPs) were selected from the initial 2000 SNPs. Overall, SSVRMES performed better than other approaches, with 76–84% accuracy and 2–4% standard deviations using 10-fold CV. However, the methods did not agree on all predictive SNPs. For instance, SSVRMES, SVMRFE and CART have found the three simulated loci (C, D, DR) on chromosome 6 to be predictive for RA status, but LRLR missed D locus. This lack of agreement may be due to the fact that some methods detected additive effects, such as LRLR, while others detected both additive and multiplicative effects, such as SSVRMES.

3.3 Application to real Rheumatoid Arthritis data

3.3.1 North American Rheumatoid Arthritis Consortium Data—Rheumatoid Arthritis (RA) is an autoimmune disease that causes chronic inflammation of the joints, the tissue around the joints, or other organs in the body. RA affects more than two million people in the United States; 70 percent of people with RA are women. RA may be triggered by an infection in people with an inherited susceptibility. Although the disease itself is not inherited, certain genes that create an increased susceptibility are. People who have inherited these genes will not necessarily develop RA, but they may have higher risk. The severity of the disease may also depend on the inherited genes [46]. The North American Rheumatoid Arthritis Consortium (NARAC) lead by Peter Gregersen has provided microsatellite and SNP scans, quantitative phenotypes, and clinical measures and was provided as part of the Genetic Analysis Workshop 15. We evaluated our approach using a subset of the RA case-control data file “CHR18SNP.dat” offered by NARAC. In the data file, a dense panel of 2300 SNPs was genotyped by Illumina [47] for an approximately 10 kb region of chromosome 18q. This region has shown evidence of association with RA in previous studies. These SNPs were individually genotyped on 460 cases and 460 controls recruited from a New York City population. All 2300 SNPs with associated case control outcomes were included to evaluate our approach.

3.3.2 Results—Prior to our approach, we examined the association between RA case control status and each of the SNPs using non-model based Pearson Chi-Square test. Since the SNP allele expressions are not ordinal, a general test of association was used. 10.43% (240) of the 2300 SNPs were found to be significantly associated with RA with significance level of alpha taking the value of 0.10; no consideration was given to the effects of multiple testing. Although this testing based approach provides the list of SNPs that are significantly related to the RA statuses, it did not consider the correlations among SNPs and ignored the redundancy. Therefore, these 240 selected SNPs could be still redundant and highly correlated, and can be further eliminated to reduce the cost for genotyping and disease prediction.

Table 3 lists the average 10-fold CV accuracies and standard deviations for our proposed SSVRMES. The CV results of popular approaches SVMRFE, NBCPT, CART, LRLR, and LOGICFS are also presented in the same table. Since the NARAC CHR18SNP case control data only included SNP variables, the performance of SVMRFE, NBCPT, CART, and LOGICFS were relatively close. LRLR performed worst and our proposed SSVRMES performed best. Note that the number of SNPs that were actually selected by individual methods varied.

To better visualize the performance of the SSVRMES we recorded the 10-fold CV accuracies as the number of selected SNPs increased from 1 to 200. Note that the CV accuracies stopped to increase after 200 SNPs for most of the runs, which is consistent with the association analysis based on Pearson Chi-square test (240 SNPs). Figure 3 displays two typical runs, and the average CV accuracies on the NARAC RA data for variable dimensions 1 to 200 with NBCPT, LOGICFS, SVMRFE, CART, LRLR, and SSVRMES approaches. It indicates that the CV accuracies of our SSVRMES are better than those of the others. This is due to that their selections ignored the correlations among SNPs, while our SSVRMES approach included embedded entropy measure, which reduced the redundancy and helped to improve the generalization performance.

Furthermore, we implemented the popular scan statistic and sliding window methods for the RA data, and present the results in Figure 4 left and right with window sizes 5 and 15, respectively. For a window size of 5 SNPs, the scan statistic identified 76 SNPs as members of 19 chromosomal regions with significance level alpha < 0.01 associated with Rheumatoid Arthritis. Significant regions contained 4 SNPs on the average, ranging from 1 to 15 SNPs. The two largest regions, each with 15 SNPs were at SNP number 136–150 and 1190–1204.

Compared to the scan statistic method, the sliding window was more likely to identify a region as significant. Occasionally, the sliding window splits a scan statistic region into sub-regions. The sliding window regions tend to be wide. For example, in the vicinity of SNP 183, the sliding window identified a significant region of 21 SNPs. The corresponding scan statistic region included only 10 SNPs, roughly centered on the sliding window region. Similar circumstances arose around SNPs 374, 600, and 1089. The sliding window identified several regions containing one SNP with a significant disease association, with the rest being marginally insignificant; while the scan statistic method did not have this issue. This finding corroborates the notion that the scan statistic tends to be more conservative in identifying regions.

Increasing the window size tended to combine significant regions together, resulting in a smoothing effect. This can be seen by comparing Figure 4-left and Figure 4-right. The number of significant regions decreased from 12 to 6, and the number of SNPs included in the significant regions increased. The scan statistic is more selective than the sliding window in identifying significant regions. This may be considered as strength of the scan statistic

approach. The more specific results of the scan statistic may increase efficiency of the overall search process.

The drawbacks of sliding window and scan statistic methods compared to our proposed methods are the following: the analysis conducted based on sliding window and scan statistic methods cannot incorporate or control for environmental factors, which are believed to be associated with RA. Specifically, factors such as age, body weight and smoking, which potentially interact with genotypes, may change the results in a meaningful way. Moreover, scan statistic and sliding window methods do not incorporate gene-gene interactions. In a complex disease, it is not likely that the outcome is attributable to a single gene. Our proposed sequential support vector regression method can incorporate both genomic and environmental factors and their interactions and therefore may better account for these interactions. It is obvious that the association tests in either scan statistic or sliding window methods may not have been independent. The violated assumption of independence would result in exaggerated p-values based on either scan statistic or sliding window methods, or identification of regions that are more significant because of correlation rather than association with the outcome. Last, scan statistic accounted for the spacing and ordering of SNPs on the chromosome and quickly identified regions of the chromosome that are associated with RA, but they are ranking methods. Therefore, they are less robust and may not generalize well to unseen data for better prediction purpose.

4. Discussion

The type of model, learning algorithm and the number of variables are main factors for good predictive and generalization performances given thousands of SNPs that are available in the genome wide data for complex diseases. In this paper, our goal is to develop a joint supervised learning model (classifier) and sequential embedded variable (SNP) selection algorithm in order to achieve minimum redundancy and maximum relevance in the Support Vector framework. The practical motivation is to develop prediction model that maximizes the disease prediction performance and achieves the lowest possible genotyping effort for diagnosis by simultaneous selections of non-redundant predictive SNPs. To do so, we proposed a new statistical embedded learning algorithm, “Sequential Support Vector Regression with Embedded Entropy” to deal with the redundancy in the highly correlated high dimensional SNP data and to find an optimal set of informative SNPs enabling and maximizing the disease prediction and classification of individuals in disease risk.

We compared our proposed methods to some popular methods for SNP-disease study. Compared to the well known variable selection methods SVMRFE, LOGICFS, and LRLR, our method gained higher CV accuracy on the average on two independent real complex disease data sets. Results have demonstrated that our proposed method provides good prediction, but with a much reduced SNP/variable set compared to methods in the literature. Moreover, we have found that as the number of variables increases, the performances of the complex models, such as our SSVRMES and SVMRFE increase while simpler models, such as NBCPT stay at the same level of performance. The reason behind this may be due to the fact that these complex models may detect those epistatic effects (gene-gene interactions) which do not exhibit statistically significant marginal effects. The detection of epistatic effects in higher dimensions requires even more complex models. In contrast, when lower levels of LD are observed at given loci, a larger number of SNPs are required to predict disease status, such as in the NARAC RA data set.

Our results also indicate that simpler models, such as logistic regression or naive Bayesian classifier perform better when the number of variables is low. As the number of variables increases, the prediction performance of complex models, such as SSVRMES or SVMRFE

increase. This indicates that complex models are more suitable to identify multiple susceptibility SNPs simultaneously. On the other hand, the requirement for more training data is larger for complex models in order to catch the relationship of the variables in the data. If the sample size is not sufficient, then the relation and model mined from the training data is not suitable for the CV data and, as a result, poor generalization performance occurs. Our study demonstrates that SSVRMES is more appropriate and is able to adjust model architecture for large, rapidly evolving data sets.

Our study indicates that if high level LD occurs in the population that can be captured by the prediction models and prior knowledge, then one to five SNPs are sufficient to obtain good predictive performance. In our MI example, to illustrate our proposed method on a small scale rather than genome wide, a small set of SNPs was selected based on the available knowledge of high level LD and environmental factors. By applying our method, we have achieved greater than 72% prediction accuracy with as few as 3 to 5 variables (Figure 2). This demonstrates that the prediction accuracy can be improved if prior biological knowledge, such as high LD regions or low minor SNP allele frequency is utilized during the variable selections. In contrast, when lower levels of LD are observed at given loci, a larger number of SNPs are required to predict disease status, such as in the case of the NARAC CHR18SNP data set (Figure 3).

Although our study shows that there are low variations regarding the CV accuracy as the window size varied from 5 to 15, the selected SNPs were slightly different due to the smoothing effect. Optimal window size is dependent on the biological properties across genome such as LD in the studied region and SNP density, and is not determined by the algorithm itself. Our proposed algorithm is a greedy search method and it detects significant SNPs in a sequential manner. Consequently, search space becomes smaller and smaller. Therefore, for computational efficiency and to speed up the search, larger window size is recommended for the context of dense SNP maps. For better determination of the optimal window size, one may utilize biological knowledge on the studied SNPs, such as high or low LD and the length of homozygous segments. Our future work may consider extending the model and taking population parameters, such as allele frequency and recombination rate into account.

Since the methods proposed in this paper are supervised learning methods, there are two possible limitations for disease predictions with SNP data: 1) Although our method can identify informative SNPs associated with clinical outcomes, if a researcher only recruits participants from the healthy population, then these supervised methods cannot handle, such "control only" data. In this case, unsupervised approaches and haplotype block approaches are more suitable. 2) Currently, the approach proposed in this paper cannot infer/capture the hidden SNPs that are not available on the genetic map. Future investigations of these areas and investigations of whether there is power reduction compared to the selected SNPs with direct assays of all common SNPs will be conducted. Since this work focuses on designing embedded optimization and regularization in the SV framework, comparison with other equivalent and comparable hierarchical Bayesian shrinkage methods for optimizing the process of selecting predictive variables that are related to complex diseases will be considered in our future work.

Acknowledgments

This work was supported in part by the NIH *P30*-NR011396-01.

REFERENCES

1. The International HapMap Consortium. The International HapMap Project. Nature. 2003; 426:789–796. [PubMed: 14685227]
2. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. Nat Rev Genet. 2004; 5:467–475. [PubMed: 15153999]
3. The International HapMap Consortium. Haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]
4. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet. 2004; 74:106–120. [PubMed: 14681826]
5. Zhang K, Jin L. HaploBlockFinder: Haplotype block analysis. Bioinformatics. 2003; 19:1300–1301. [PubMed: 12835279]
6. Li J, Jiang T. Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics. 2005; 21:4384–4393. [PubMed: 16249262]
7. Ao S, Yip K, Ng M, Cheung D, Fong PY, Melhado I, Sham PC. CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. Bioinformatics. 2005; 21(8):1735–1736. [PubMed: 15585525]
8. Lin Z, Altman RB. Finding haplotype tagging SNPs by use of principal components analysis. Am. J. Hum. Genet. 2004; 75:850–861. [PubMed: 15389393]
9. Benjamin DH, Nicola JC. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genetic Epidemiology. 2004; 26(1):11–21. [PubMed: 14691953]
10. Toivonen HT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, Kere J. Data mining applied to linkage disequilibrium mapping. Am. J. Hum. Genet. 2000; 67(1):133–145. [PubMed: 10848493]
11. Beckmann L, Thomas DC, Fischer C, Chang-Claude J. Haplotype sharing analysis using Mantel statistics. Human Heredity. 2005; 59:67–78. [PubMed: 15838176]
12. Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. Genetic Epidemiology. 1996; 13:423–449. [PubMed: 8905391]
13. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score test for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002; 70:425–443. [PubMed: 11791212]
14. Levin AM, Ghosh D, et al. A model-based scan statistics for identifying extreme chromosomal regions of gene expression in human tumors. Bioinformatics. 2005; 21:2867–2874. [PubMed: 15814559]
15. Neale B, Sham P. The future of association studies: Gene-based analysis and replication. American Journal of Human Genetics. 2004; 75:353–362. [PubMed: 15272419]
16. Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med. 2006; 25:105–126. [PubMed: 16220513]
17. Anderson EC, Novembre J. Finding haplotype block boundaries by using the minimum-description-length principle. American Journal of Human Genetics. 2003; 73:336–354. [PubMed: 12858289]
18. Mannila H, Koivisto M, Perola M, Varilo T, Hennah W, Ekelund J, Lukk M, Peltonen L, Ukkonen E. Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. Am. J. Hum. Genet. 2003; 73:86–94. [PubMed: 12761696]
19. Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. Hum Genet. 2003; 114:36–43. [PubMed: 14505034]
20. Zhao J, Boerwinkle E, Xiong M. An entropy-based statistic for genomewide association studies. American Journal of Human Genetics. 2005; 77:27–40. [PubMed: 15931594]

21. Sun YV, Levin AM, Boerwinkle E, Robertson H, Kardia SL. A scan statistic for identifying chromosomal patterns of SNP association. Genet Epidemiol. 2006; 30(7):627–635. [PubMed: 16858698]
22. Schwender H, Ickstadt K. Identification of SNP Interactions Using Logic Regression. Biostatistics. 2008; 9(1):187–198. [PubMed: 17578898]
23. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and Regression Tress. Belmont: Wadsworth; 1984.
24. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.
25. Liang Y, Kelemen A. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. Statistics Surveys. 2008; 2:43–60. (electronic). DOI: 10.1214/07-SS026.
26. Wang L, Zhu J, Zou H. Doubly regularized support vector machine. Statistica Sinica. 2006; 16:589–615.
27. Sun W, Cai T. Oracle and adaptive compound decision rules for false discovery rate control. J. American Statistical Association. 2007; 102:901–912.
28. Lal, TN.; Chapelle, O.; Weston, J.; Elisseeff, A. Embedded methods. Feature Extraction: Foundations and Applications. In: Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, LA., editors. Berlin, Germany: Springer; 2006.
29. MacKay, DJC. London: Cambridge University Press; 2003. Information Theory, Inference, and Learning Algorithms, Chap. 4; p. 73-74.
30. Care M, Needham C, Bulpitt A, Westhead D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics. 23(6):664–672. [PubMed: 17234639]
31. Vapnik, VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.
32. Vapnik, VN. Statistical Learning Theory. New York: Wiley; 1998.
33. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2(2):121–167.
34. Smola AJ, Lkopf BS. A tutorial on support vector regression. Statistics and Computing. 2004; 14:199–222.
35. Mukherjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time series using a support vector machine. Proceedings of the IEEE Workshop on Neural Networks for Signal Processing. 1997; 7:511–519.
36. Stitson, M.; Gammerman, A.; Vapnik, V.; Vovk, V.; Watkins, C.; Weston, J. Support vector regression with ANOVA decomposition kernels. In: Scholkopf, B.; Burges, CJC.; Smola, AJ., editors. Advances in Kernel Methods—Support Vector Learning. Cambridge, MA: MIT Press; 1999. p. 285-292.
37. Girosi F. An equivalence between sparse approximation and support vector machines. Neural Computation. 1998; 10(6):1455–1480. [PubMed: 9698353]
38. Liang Y, Kelemen A. Temporal Gene Expression Classification with Regularised Neural Network. International Journal of Bioinformatics Research and Applications. 2005; 1(4):399–413. [PubMed: 18048144]
39. Guyon I, Weston J, Barnhill S, Vapnik VN. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002; 46(1–3):389–422.
40. Cao JF, Braak T. Regression by L1 regularization of smart contrasts and sums (ROSCAS) beats PLS and elastic net in latent variable model. Journal of Chemometrics. 2009; 23(5):217–228.
41. Leng, C.; Lin, Y.; Whaba, G. A note on the Lasso and related procedures in model selection. 2004. [Online]. Available http://www.stat.wisc.edu/~wahba/ftp1/tr1091rxx.pdf
42. Cheng R, Ma JZ, Wright FA, Lin S, Gao X, Wang D, Elston RC, Li MD. Nonparametric disequilibrium mapping of functional sites using haplotypes of multiple tightly linked single-nucleotide polymorphism markers. Genetics. 2003; 164:1175–1187. [PubMed: 12871923]
43. Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet. 2003; 73:115–130. [PubMed: 12796855]

44. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948; vol. 27:379–423. 623–656.

45. Witte, JH.; Schnell, AH.; Cordell, HJ.; Almasy, L.; MacCCluer, JW., editors. Genetic Epidemiology. Vol. 31. 2007. Genetic Analysis Workshop 15: Summaries of the Design and Analysis of Genomic Data; p. S1-S148.

46. Seldin M, Amos C, Ward R, Gregerson P. The genetics revolution and the assault on rheumatoid arthritis. Arthritis and Rheumatism. 1999; 42(6):1071–1079. [PubMed: 10366098]

47. Jawaheer D, Seldin M, Amos C, Chen W, Shigeta R, Monteiro J, Kern M, Criswell L, Albani S, Nelson J, Clegg D, Pope R, Schroeder H Jr, Bridges S Jr, Pisetsky D, Ward R, Kastner D, Wilder R, Pincus T, Callahan L, Flemming D, Wener M, Gregersen P. A Genomewide Screen in Multiplex Rheumatoid Arthritis Families Suggests Genetic Overlap with Other Autoimmune Diseases. Am. J. Hum. Genet. 2001; 68:927–936. [PubMed: 11254450]

48. Miller MB, Lind GR, Li N, Jang SY. Genetic analysis workshop 15: simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense SNP map with linkage disequilibrium between marker loci and trait loci. BMC Proceedings. 2007; 1 Suppl 1:S4. [PubMed: 18466538]

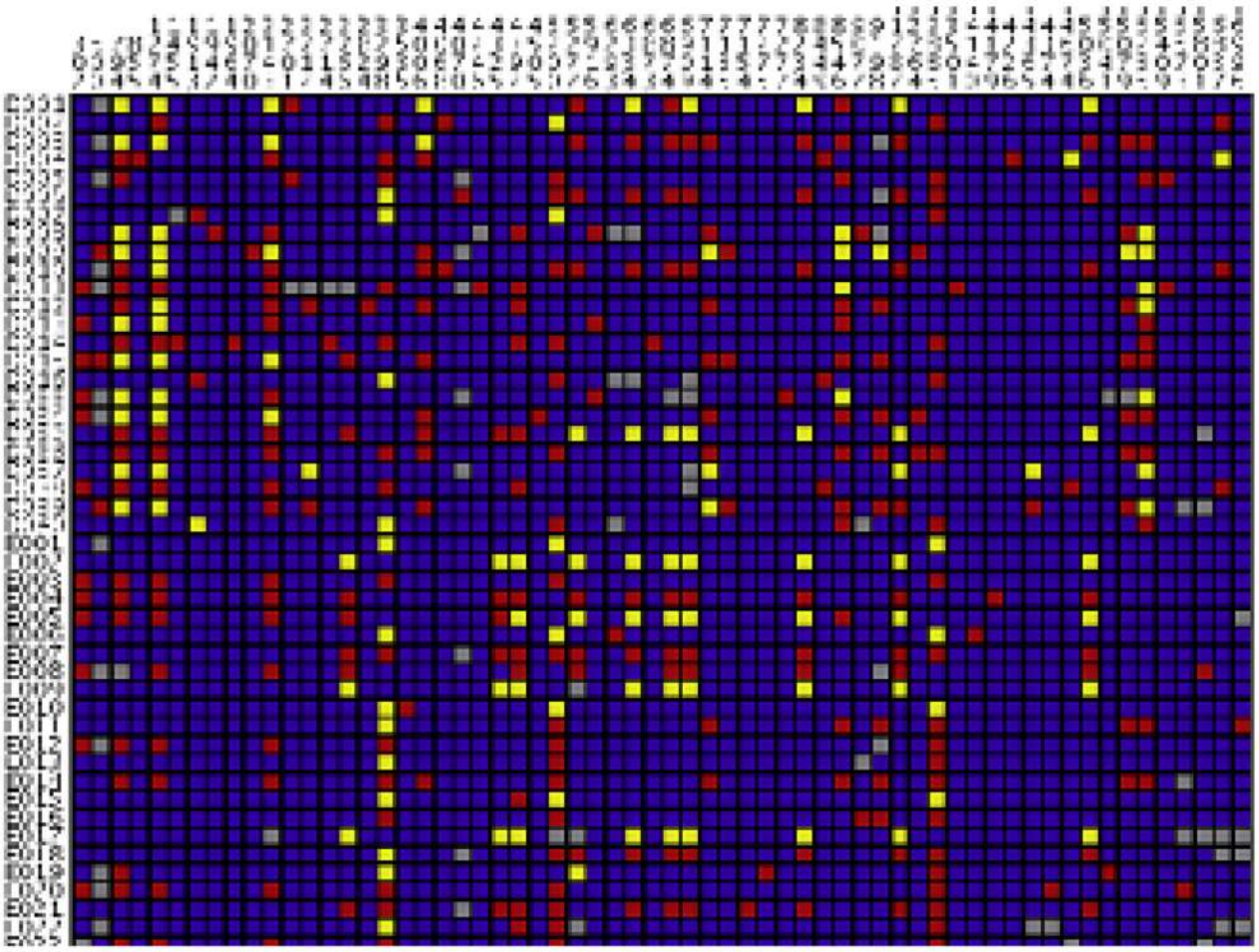


Figure 1.
Graphical representation of individual genotypes for Gene IL-1 beta related to metabolic pathways of Myocardial Infarction. 62 SNPs (columns) for 47 subjects (rows) are shown. Blue: homozygous common allele; yellow: homozygous rare allele; red: heterozygous; gray: missing or unknown. Many SNPs are highly correlated with redundancy.

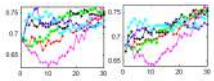


Figure 2.
10-fold CV accuracies from two typical runs for the MI data set under different number of variables selected; variable dimensions 1 to 30. X: the number of variables selected; Y: prediction accuracy. The color marks represent different methods: *black*: SSVRMES; *dark blue*: SVMRFE; *light blue*: LOGICFS; *red*: CART; *green*: NBCPT; *pink*: LRLR.

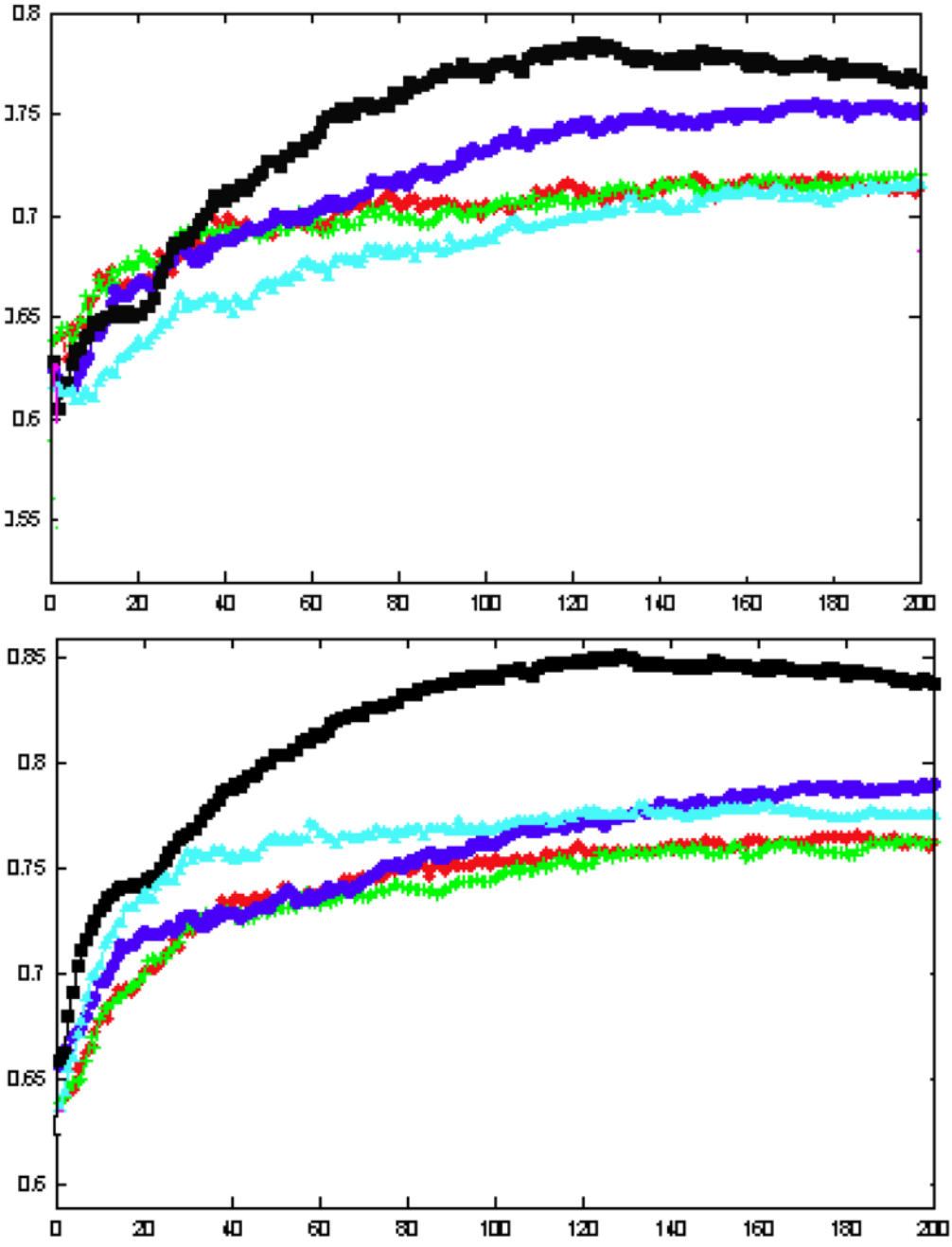
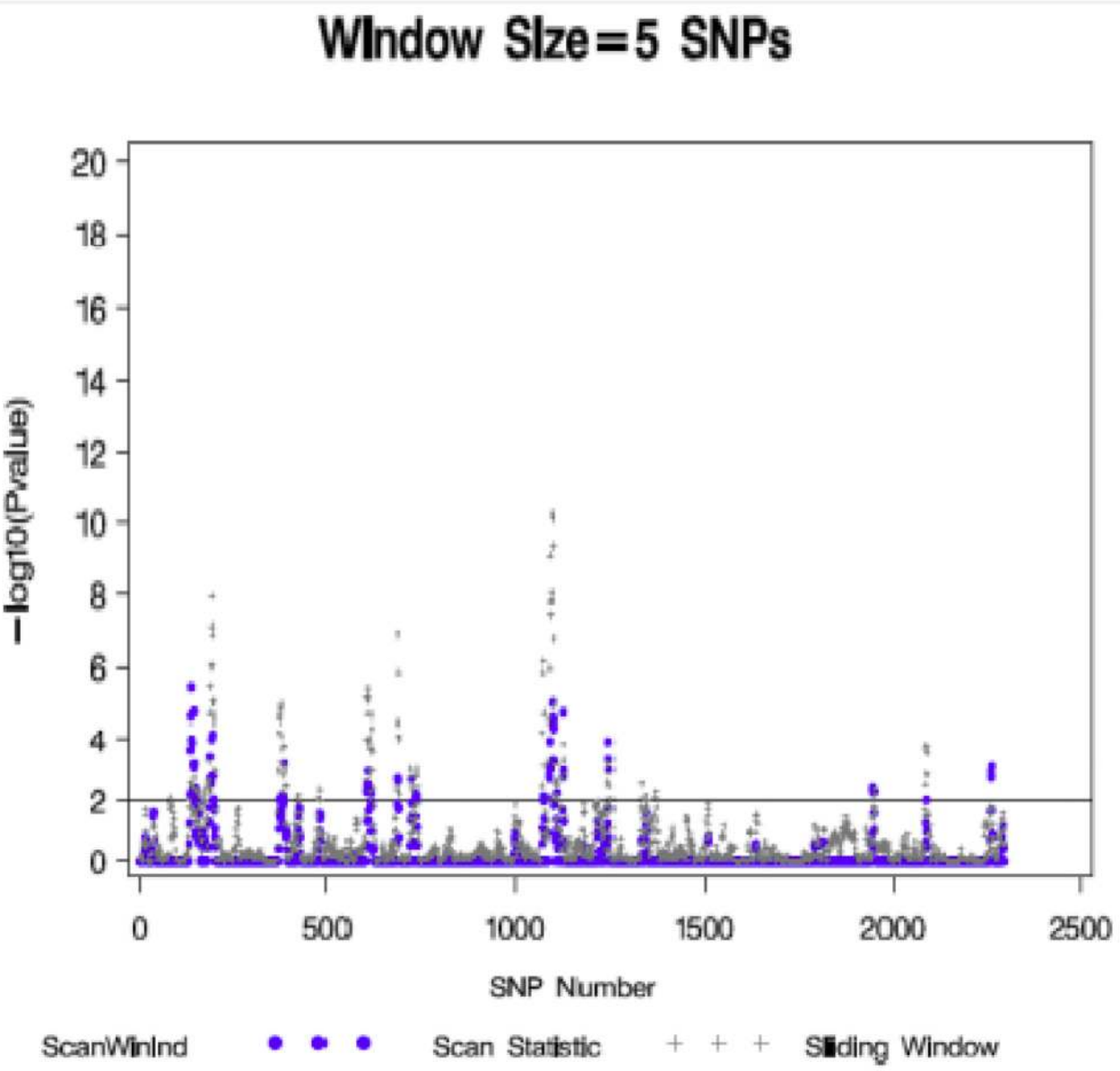


Figure 3.
10-fold CV accuracies from two typical runs for the Rheumatoid Arthritis data set under different number of variables selected; variable dimensions increase from 1 to 200 given 2300 SNPs. X: number of variables selected; Y: prediction accuracy. The color marks represent different methods: *black*: SSVRMES; d*ark blue*: SVMRFE; *light blue*: LOGICFS; *red*: CART; *green*: NBCPT. LRLR is not included due to its consistent worst performance compared with the above listed approaches.



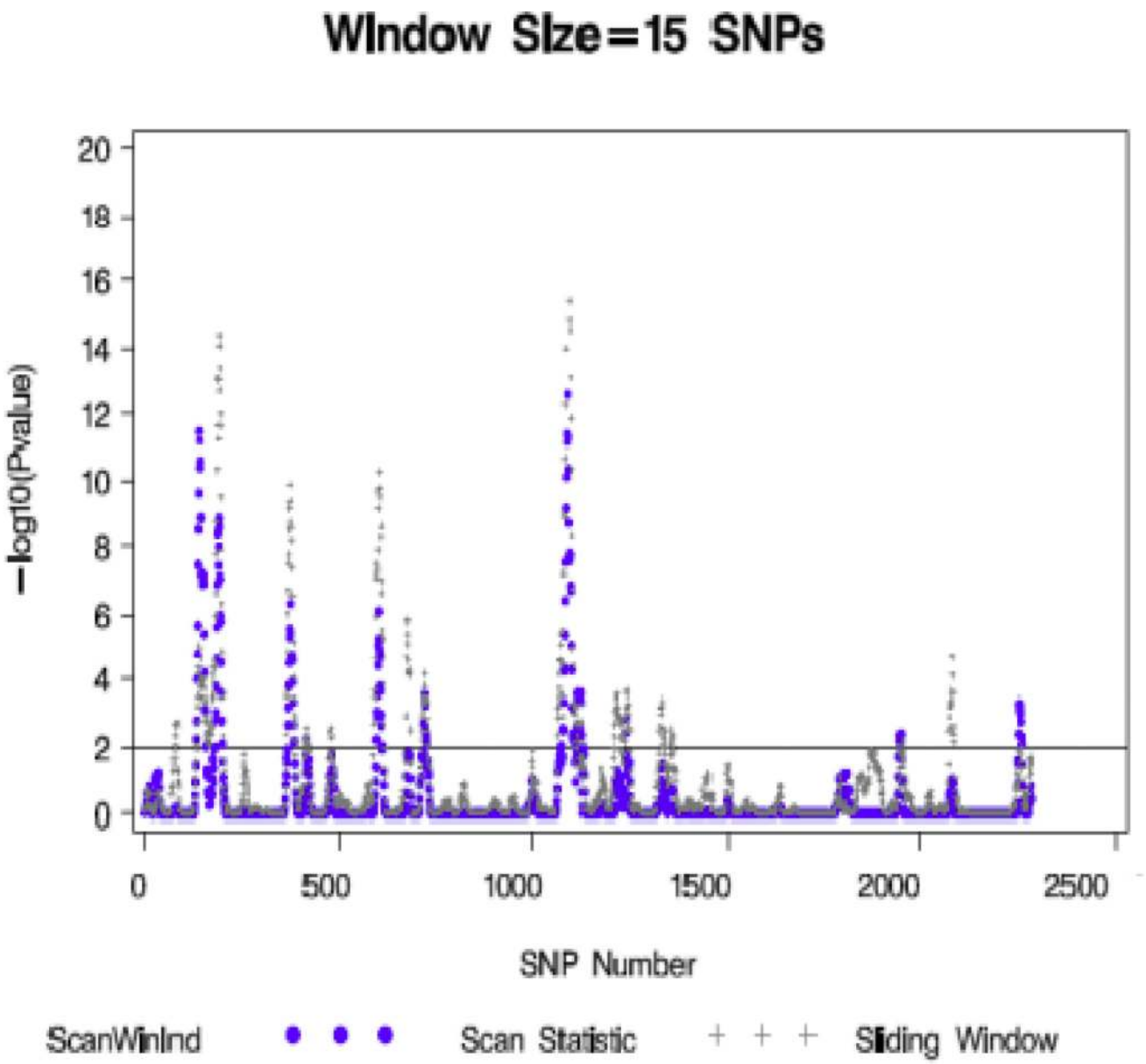


Figure 4.
Results of scan statistic and sliding window methods for Rheumatoid Arthritis data with window sizes 5 and 15. For window size 5 (5 SNPs) the scan statistic identified 76 SNPs as members of 19 chromosomal regions with significant alpha < 0.01 RA association. Significant regions contained 4 SNPs on the average, ranging from 1 to 15 SNPs. The two largest regions, each with 15 SNPs were at SNP numbers 136–150 and 1190–1204.

Table 1

10-fold CV accuracies for Myocardial Infarction data set.

Method	10-fold CV accuracy (mean (%) ± standard deviation (%))
SSVRMES	77.4 ± 1.9
SVMRFE	74.5 ± 2.1
CART	73.3 ± 2.7
LRLR	73.1 ± 4.0
LOGICFS	54.4 ± 1.5

Table 2

10-fold CV accuracies and the number of selected SNPs for the simulated RA data sets with different λ values and window sizes for SSVRMES.

SSVRMES		**λ values**			
		0.2	0.4	0.6	0.8
Window size 5	# of SNP	112	91	82	63
	CV accuracy (mean±SD(%))	76±3.4	79±2.5	81±2.4	75±3.1
window size 15	# of SNP	132	116	112	87
	CV accuracies (mean±SD(%))	75±4.1	77±3.2	80±2.6	76±3.5

Table 3

10-fold CV accuracies for NARAC Rheumatoid Arthritis data set.

Method	10-fold CV accuracy (mean (%) ± standard deviation (%))
SSVRMES	81.4 ± 2.8
SVMRFE	75.8 ± 3.1
CART	72.3 ± 2.7
LRLR	67.4 ± 2.9
LOGICFS	74.5 ± 3.5