The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 237–244

A Novel SVM-RFE for Gene Selection*

Jun-Yan Tan[1] Zhi-Xia Yang[2,3] Naiyang Deng[1,†]

[1] College of Science, China Agricultural University, 100083, Beijing, P.R.China
[2] College of Mathematics and Systems Science, Xinjiang University, 830046, Urumuqi, P.R.China
[3] Academy of Mathematics and Systems Science, CAS, 100190, Beijing, P.R.China

Abstract Selecting a subset of informative genes from microarray expression data is a critical data preparation step in cancer classification and other biological function analysis. The support vector machine recursive feature elimination (SVM-RFE) is one of the most effective feature selection method which has been successfully used in selecting informative genes for cancer classification. While, the SVM-RFE selects genes only using the gene expression data without using any other biological information of the genes. Based on the biology information of the genes, it may be beneficial to identify the genes that are relevant to the cancer. We propose a novel SVM-RFE method for gene selection by incorporating the Kyoto Encyclopedia of genes and genomes (KEGG) pathway information into feature selection process. Numerical results indicate that the novel SVM-RFE tends to provide better variable selection results than the SVM-RFE.

Keywords Support vector machine; microarray data; gene selection

1 Introduction

The microarray datasets usually contain only a small number of samples. This characteristic raises new challenges for data analysis. In the classification, data overfitting arises when the number of features is much larger than the number of the samples. In order to overcome the risk of over-fitting, it is necessary to reduce the data dimensionality by selecting a subset of features(genes) that are relevant for classification. In microarray analysis, researchers are more interested in identifying the genes that are relevant to the cancer. The current gene selection methods aims at doing nothing but the enhance of the classification accuracy, this may lead to the selection result violate the biology fact, it is desirable to have a tool that can consider both the classification accuracy and biology fact.

Guyon *et al.* (2002)[1] proposed the support vector machine-recursive feature elimination (SVM-RFE). The SVM-RFE method ranks all the genes according to some score function and eliminates one or more genes with the lowest scores. This process is repeated until the highest classification accuracy. Magasarian (1998) [2] and Magasarian (2007) [3] proposed the feature selection via concave minimization (FSV), which can automatically select features by the l_0- norm penalty of the number of features. But their classification accuracy was not very good due to the loss of the maximum margin . Neumann

*This work is supported by the Key Project of the National Natural Science Foundation of China(No. 10631070), the National Natural Science Foundation of China (No.10801112) and the China Postdoctoral Science Foundation funded project(No.20080430573).

†corresponding author:dengnaiyang@vip.163.com

(2005) [4] proposed the $l_2 - l_0$ norm SVM to improve the generalization performance of the classifiers. It combined the l_2- norm with the l_0- norm and performs better in the classification accuracy than the FSV due to the l_2-norm of w in the objective function. Wang (2008) [5] proposed a hybrid huberized support vector machine (HHSVM) which replaced the loss function in the SVM by the huberized hinge loss function.

One limitation of all the above approaches is that the methods are developed purely from computational or algorithmic points without utilizing any prior biological knowledge or information. The large body of information is now available through databases on different aspects of biological systems. Some well known databases include the Gene Ontology (GO)([6]), Kyoto Encyclopedia of genes and genomes(KEGG) pathways ([7]). This kind of prior information is a useful supplements to the standard numerical data coming from an experiment. Furthermore, in microarray analysis, many genes are known to have the same function or involved in the same pathways as some of known/putative cancer-related genes, and the genes in the same functional group or pathway are more likely to work together. So, the ideal gene selection methods eliminate the trivial genes and automatically include the whole group genes into the model once one gene among them is selected. Therefore, we have to deal with one important problem in gene selection: how to take into account the pathway information between genes. In this paper, we propose a gene selection method by involving in the information of KEGG pathways to achieve the object that the genes in the same pathway can be selected or removed together.

The rest of the paper is organized as follows. In section 2, we introduce the dataset used in this paper. In section 3, we introduce the linear SVM and the SVM-RFE. In section 4, we first prove the grouping effect of SVM, then propose our algorithm for gene selection. In section 5, we apply our method to simulation and real microarray datasets . We conclude the article in section 6.

2 Datasets

To evaluate the performance of our proposed method, we apply it to two simulation datasets and two real microarray gene expression datasets respectively. Now, let us address to the description of the datasets used in the experiments.

Simulation data 1

We construct 20 samples($l = 20$) with 300-dimensionality($n = 300$), including 10 positive samples and 10 negative samples. The dimensionality n is much larger than the number of samples, which fits the fatual gene microarray data. The positive samples are generated from a normal distribution with the mean $\mu_+ = (\underbrace{1,1,\cdots,1}_{10},\underbrace{0,0,\cdots,0}_{290})^{\mathrm{T}}$ and the covariance$\Sigma_{i,j} = \begin{pmatrix} \Sigma^*_{10\times 10} & 0_{10\times 290}; \\ 0_{290\times 10}, & I_{290\times 290}, \end{pmatrix}$, where the diagnoal elements of Σ^* are 1 and the off-diagnoal elements are all equal to $\rho = 0.9$. The negative samples have similar distribution to positive samples except that $\mu_- = (\underbrace{-1,-1,\cdots,-1}_{10},\underbrace{0,0,\cdots,0}_{290})^{\mathrm{T}}$. Obviously, the optimal classification rule depends on the first ten variables (genes).

Simulation data 2

This simulation data is a modified version of simulation data 1. The only difference between them is that some of the genes are in the same pathway. We assume that, the 1th, 2th and 300th gene are in the same pathway, they should be selected together. The first ten genes and the 300th gene are informative.

Colon cancer dataset

The colon cancer datatset provided by Alon and Barkai (1999), consists of expression levels of 62 samples of which 40 samples are colon cancer samples and the remaining are normal samples. Each sample consists of 2000 genes.

Prostate dataset

This dataset provides the expression levels of 12600 genes for 50 normal samples and 52 prostate cancer samples. The details can be found in [8].

3 Methods

Given the training set

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (\mathscr{X} \times \mathscr{Y})^l, \tag{1}$$

where $x_i = ([x_i]_1, [x_i]_2, \cdots, [x_i]_n) \in \mathscr{X} \subseteq R^n$ is the input and its n components are called 'features'. For the microarray data, the n features are n gene expression levels. $y_i \in \mathscr{Y} = \{-1, 1\}$ is the output, it means 'normal' or 'cancerous' for microarray data.

3.1 Background

Support vector machine(SVM) We briefly introduce linear SVM and refer interested readers to ([11],[12]) for detail. The training set T is given by (1), the SVM is to find a hyperplane that separates the two classes of data points by the maximizing margin:

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \ , \tag{2}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i \ , \ i = 1, \cdots, l \ , \tag{3}$$

$$\xi_i \geq 0 \ , \ i = 1, \cdots, l \ , \tag{4}$$

where the constant $C(> 0)$ is parameter. The dual problem of the problem $(2) \sim (4)$ is

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^{l} \alpha_i \ , \tag{5}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \ , \tag{6}$$

$$0 \leq \alpha_i \leq C \ , \ i = 1, \cdots, l \ , \tag{7}$$

Suppose $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$ is the solution of the dual problem $(5) \sim (7)$, if there exists some j such that $0 < \alpha_j^* < C$, the solution about (w, b) of the primal problem $(2) \sim (4)$ can be calculated by the following:

$$w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i, x_j), b^* = y_i - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j). \tag{8}$$

A new point x is to be assigned with the label $f(x) = sign((w^* \cdot x) + b^*)$. And, for microarray analysis, the $i-$th element of w is the weight of the $i-$th gene.

SVM-RFE In SVM-RFE, the parameter q, here named "filter-out" factor, is used to decide how many genes are removed at one step. Notice if $0 < q < 1$, a fraction of q bottom-ranked genes are removed at each step; if $q = -1$, only one gene is removed; if $q = -2$, two genes are removed, and so on. Now, we describe the algorithm in detail.

Algorithm 1. SVM-RFE

```
1. Given the training set $T$ (1), the subscripts set of all input features (genes) $F$ , the filter out factor $q$, the size of final informative genes $s$ and the parameter $C > 0$ ;
2. Solve the problem (5) $\sim$ (7) and get its solution $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$ , compute the weight vector $w^*$ according to the equation (8);
3. Rank the features of $F$ by $w_i^2$ in the descending order.
4. If $q < 0$, $F2 = F - \{q$ bottom ranked features in F$\}$;
   if $0 < q < 1$, $F2 = F - \{q * 100\%$features with the largest rank in F$\}$;
5. If the size of $F2 = s$ or the size of $F2 < s$, adjust $F2$ to be composed of $s$ top ranked features in $F$ and stop, otherwise, F=F2 goto step 2.
```

3.2 New methods

We fist prove the ‘grouping effect’ of the linear SVM which means that highly correlated genes can have similar weights, then propose the novel algorithm for gene selection.

The grouping effect of SVM The training set T is given by (1), we pay particular attention to the gene vector $g_i = ([x_1]_i, [x_2]_i, \cdots, [x_l]_i)^{\mathrm{T}}$, which comprises the i-th feature of all inputs to denote the expression levels of i-th gene in all inputs, where $i = 1, 2, \cdots, n$. Now, we illustrate the ‘grouping effect’ of the linear SVM with the following theorem:

Theorem 1. Let (w^*, b^*) is the solution of problem (2) $\sim$ (4), for any (i, j), we have $|w_i^* - w_j^*| \leq \sqrt{lM}||g_i - g_j||$, if the input vector g_i and g_j are normalized with mean 0 and norm 1, then$|w_i^* - w_j^*| \leq 2\sqrt{lM}\sqrt{1-\rho}$ where ρ is the correlation between g_i and g_j.
Proof: From (8), we have $w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i$, using the Cauchy inequality, we get

$$|w_i - w_j|^2 = (w_i - w_j)^2 = (\sum_{k=1}^{l} \alpha_k y_k (x_{ki} - x_{kj}))^2 \leq \sum_{k=1}^{l} \alpha_k^2 y_k^2 (x_{ki} - x_{kj})^2 \tag{9}$$

Let $M = \max\{\alpha_1^2, \alpha_2^2, \cdots, \alpha_l^2\}$, then $\sum_{k=1}^{l} \alpha_k^2 y_k^2 (x_{ki} - x_{kj})^2 \leq lM||g_i - g_j||^2$. Moreover, we have the result that $|w_i^* - w_j^*| \leq 2\sqrt{l C}\sqrt{1-\rho}$, where C is the parameter in the SVM.

Theorem 1 suggests that highly correlated variables (genes) tend to have similar estimated weights, this can be seen clearly from Table 1. Table 1 shows the ranking results of ten informative genes when C gets different values for simulation data 1. From table 1, we can see that C can control the similarity of the weights between g_i and g_j except for the determination of the trade-off between the margin and the classification accuracy.

Novel algorithm Now, we refine the SVM-RFE algorithm for gene selection by incorporating the pathway information from KEGG. In KEGG, a gene may be annotated in multiple pathways. To deal with this problem, we keep a gene to the pathway with the smallest ID (the pathway serial number in KEGG). We introduce a vector P whose

Table 1: The grouping effect of SVM on simulation data 1

		$corr = 0.9$		$corr = 1$
C	Error	Relevant gene location	Error	Relevant gene location
0.09	0	$41 \sim 50$	4	$39 \sim 48$
0.1	1	$38, 39$ $42 \sim 48$ $49, 50$	0	$41 \sim 50$
1	0	$17, 18, 21, 22, 24, 29, 30, 34, 36, 39$	0	$39 \sim 48$
10	0	$19, 25, 27, 30, 31, 32, 37, 40, 41, 42$	0	$29 \sim 38$

dimensions are the same with the number of genes by the following method: the ith element of P is i if the $i-$th gene is not in the same pathway with any other genes; if the $i-$th gene is in the same pathway with the k-th gene and $(i < k)$, the i-th and k-th element of P is i. Then the KEGG pathway information of all the genes is given by this vector P. According to the vector P, we can divide all the genes into some groups, a group is a set of genes that are in the same pathway. We use group$_i$ to express the genes that are in the same KEGG pathway with gene i, $|$group$_i|$ is the numbers of genes in group$_i$. We still incorporate the 'filter- out' factor q, which decides how many genes are removed at one step. Notice, if $0 < q < 1$, a fraction of q bottom-ranked groups ($q * 100\%$ bottom-ranked genes and the genes that are in the same pathway with them) are removed at each step; if $q = -1$, only one bottom ranked group is removed; if $q = -2$, two bottom-ranked groups are removed, and so on. Next, we present the algorithm which aims at selecting genes.

Algorithm 2. Novel SVM-RFE

1. Given the training set T (1), the subscripts set of all input features (genes) F , the gene pathway information vector P,the filter out factor q, the size of final informative genes s and the parameter $C > 0$;
2. Solve the problem (5) $\sim$ (7) and get its solution $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$, compute the weight vector w^* according to the equation (8);
3. We recompute the weight of gene i by $|w1_i| = \max_{i \in group_i}\{|w_i^*|\}$ and rank genes by $|w1_1|, |w1_2|, \cdots |w1_n|$ and rank the features in F by $w1_i^2$ in descending order;
4. if $q < 0$, $F2 = F - \{q\ bottom\ ranked\ groups\ in\ P\}$;
   if $0 < q < 1$, $F2 = F - \{q * 100\%\ groups\ in\ P\}$;
5. if the size of $F2 \leq s$, adjust $F2$ to be composed of s top ranked groups in P and stop, otherwise, F=F2 goto step 2.

Because of the 'grouping effect ' of SVM and the incorporating of gene pathway information, the algorithm can select or remove those genes whose expression level are highly correlated and the genes that are in the same pathway with them. This will be clearly seen in the numerical experiments.

4 Results

4.1 Comparison on simulation data 2

We compare the SVM-RFE and the Novel SVM-RFE on simulation data 2. The 'filter-out' factor $q = -1$, the parameter C is selected by ten-fold cross validation.

Table 2: Comparison of average number of selected relevant genes during 50 experiments on simulation data 2

	Error (%)	Number of relevant gene
SVM-RFE	8.2(1.14)	5.64
Novel SVM-RFE	5.8(0.98)	7.72

Table 3: Comparison of selected features in one experiment on simulation data 2

	Relevant feature number
SVM-RFE	$(1,2,3,4,5,8)$
Novel SVM-RFE	$(1,2,3,4,5,8,10,300)$

We random split the simulation data 2 into training and test set 50 times, for each split, 12 samples (6 '+' samples and 6 '-' samples) are used for training and the rest for testing. For simulation data 2, the relevant genes refer to the first ten and the 300-th gene. Table 2 shows the mean of test errors and the selected relevant genes by different methods. We can see that the test error of our Novel SVM-RFE is lower due to the fact that our method can select more relevant genes than SVM-RFE. From Table 3, we can see that our method can select gene 1, gene 2 and gene 300 simultaneously (they are in the same pathway and they should be selected together), while, the SVM-RFE cannot achieve this goal.

4.2 Comparison on real data

For colon cancer dataset, we randomly split the dataset into training and test sets 100 times; for each split, the training set consists of 42samples(27 cancer samples and 15 normal samples), the rest samples form the test set. We apply t-test+SVM, SVM-RFE and the Novel SVM-RFE for each split. The t-test+SVM means that the genes are selected by t-test, the SVM is used for classification. For the SVM-RFE and the novel SVM-RFE, the filter-out factor $q = 0.1$. The parameter C is selected by ten-fold cross validation.

Table 4: The most frequently selected genes by SVM-REF and Novel SVM-RFE for Colon dataset

SVM-RFE			Novel SVM-RFE		
Gene number			Gene number		
493	377	765	493	377	765
792	1423	1772	1772	1423	14
353	70	1570	70	116	346
1346	1976	1740	432	526	632
1873	249	419	792	982	986
1482	1641	1924	1033	1173	1474
-	-	-	1830	1831	1968

Table 5: The most frequently selected genes by Novel SVM-RFE for Colon dataset

Gene number	Description
493	Myosin heavy chain
765	Human CRP gene
1423	Smooth muscle isoform (human)
hsa04514	Cell adhesion molecules
377	mRNA for GCAP-II/UGN precursor(human)
1772	Collagen alpha 2(XI) chain
14	Smooth muscle isoform (human)

Table 6: The frequently selected genes by SVM-REF and Novel SVM-RFE for Proatate dataset

SVM-RFE			Novel SVM-RFE		
Gene number			Gene number		
205	6185	7623	6185	11570	12146
11942	4525	6390	205	6390	4525
9044	10234	12495	9044	10234	11942
6220	7139	9093	7623	7139	9850
9172	10537	12153	9172	55	12153
7298	8123	8965	5890	8123	8416
10956	470	-	9255	-	-

Table 4 summarize the genes that are selected more than 50 times out of 100 experiments by the SVM-RFE and the Novel SVM-RFE. As we can see, some of the most frequently selected genes by the SVM-RFE and the Novel SVM-RFE are the same, such as the $493, 765, 1423, 377, 1772, 70, 792$, these genes are proved to be more relevant to colon cancer in [5]. The genes whose gene number are 70,792,116,346,432,526,632,792, 982,986,1033,1173,1474,1830, 1831,1968 respectively are in the same KEGG pathway ‘has04514’, they should be selected together. The SVM-RFE selects two of them ‘70,792’, the Novel SVM-RFE selects all of them together. The selected genes and the description is listed in Table 5.

For prostate dataset, we also randomly split the dataset into training and test sets with the sample size 68(33 normal samples and 35 prostate cancer samples) and 34 respectively. We repeat it 100 times. The right part of Table 7 summarizes the results. From Table 6, we can see that the gene selection behavior of the SVM-RFE and the Novel SVM-RFE for the prostate dataset are similar to those for the colon dataset.

The average test errors of the three methods and the number of selected genes are summarized in the upper part of Table 7. We can see that the novel SVM-RFE seemed to have a slightly better classification accuracy than other methods.

Table 7: Results on 100 randomsplits of the original datasets

		Error(%)	Number of genes
Colon dataset	T-test	15.6(1.28)	70
	SVM-RFE	15.8(1.35)	64
	Novel SVN-RFE	13.9(1.48)	71.85
Prostate dataset	T-test	10.69(1.41)	100
	SVM-RFE	7.9(1.65)	60
	Novel SVN-RFE	7.8(1.53)	99.02

5 Conclusions and the future work

In this paper, we have proposed the novel SVM-RFE for gene selection. This algorithm incorporates biology prior knowledge of the genes into the process of gene selection. We have presented some evidence that the gene selection result of this algorithm tends to accord more with the biology facts. Furthermore, how to handle genes with multiple annotations warrants more research.

References

[1] Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46:** , 389-422,2002.

[2] Bradley P.S., Mangasarian O.L. Feature selection via concave minimization and support vector machines. *In Proc. 13th ICML*,82-90,1998.

[3] Mangasarian O.L., Wild E.W. Feature selection for nonlinear kernel support vector machines.*IEEE Seventh International Conference on Data Mining (ICDM'07)*,2007.

[4] Neumann J., Schnörr C., Steidl G. Combined SVM-based feature selection and classification. *Mach. Learn.*. **61:** 129-150,2005.

[5] Wang L., Zhu J., Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, **23:** 2507-2517,2008.

[6] www.geneontology.com.

[7] www.KEGG.com

[8] Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., Lander E., Loda M., Kantoff P., Golub T.,Sellers W. Gene expression correlates of clinical prostate cancer behavior. *cancer cell*,**1:**,203-209,2002.

[9] Lai Y.L.,Genome-wide co-expression based prediction of differential expressions.*Bioinformatics*,**24:**666-673,2008.

[10] http://www.ucsf.edu/pgdb

[11] Vapnik, V.1995. *The Nature of Statistical Learning Theory* . Springer-Verlag, New York.

[12] Burges, C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Dsicov*. **2**, 121-167.