Journal of the Korean Data & Information Science Society
2010, **21**(5), 959–965

Variable selection for multiclassification by LS-SVM[†]

Hyungtae Hwang[1]

[1]Department of Statistics, Dankook University

Received 9 July 2010, revised 23 August 2010, accepted 28 August 2010

Abstract

For multiclassification, it is often the case that some variables are not important, while some variables are more important than others. We propose a novel algorithm for selecting such relevant variables for multiclassification. This algorithm is based on multiclass least squares support vector machine (LS-SVM), which uses results of multiclass LS-SVM using one-vs-all method. Experimental results are then presented which indicate the performance of the proposed method.

Keywords: Generalized cross validation function, kernel function, least squares support vector machine, multiclassification, variable selection.

1. Introduction

A modified version of support vector machine (SVM) originally introduced by Vapnik (1995, 1998) in a least squares sense has been proposed for classification in Suykens and Vandewalle (1999a). In LS-SVM concerning classification problems, we have regression interpretations and direct links to work in classical statistics. The solution is given by a linear system instead of a quadratic programming problem. The fact that LS-SVM has explicit primal-dual formulations has lots of advantages. Kernel tricks are used in LS-SVM to treat the nonlinear relation between input variables and output variable. See Cho *et al.* (2010), Hwang (2010), Shim and Lee (2009), and Shim *et al.* (2009a) for the reference.

The binary classification by SVM or LS-SVM is known to be well developed. Multiclassification is typically performed using voting scheme method based on combining a set of binary classifications (Scholkopf *et al.*, 1995). Suykens and Vandewalle (1999b) proposed multiclassification method using LS-SVM in a step but its linear equation is composed of several linear equations corresponding to each of binary classifications. Weston and Watkins (1998) proposed multiclassification method using SVM which does not use a combination of binary classifications.

Variable selection is very important in microarray technology which allows us to look at many genes at once and determine which are expressed in a particular cell type. This technology has various applications such as gene discovery, disease diagnosis and drug discovery.

[†] The present research was conducted by the research fund of Dankook university in 2010.

[1] Professor, Department of Statistics, Dankook University, Gyeonggido 448-701, Korea.
E-mail: hthwang@dankook.ac.kr

In most microarray data, some genes are irrelevant and some relevant genes (marker genes) play a more important role than others in classification. The selection of marker genes for classification of different phenotypes, predominantly cancer types, using microarray gene expression data is to provide a better understanding of the underlying biological system and to improve the prediction performance of classifiers. There are lots of literatures in studies of variable selection, Guyon *et al.* (2002), Tibshirani *et al.* (2002), Zhang *et al.* (2006) and Koo *et al.* (2006). Guyon *et al.* (2002) developed SVM with a recursive features elimination (SVM-RFE) algorithm and Tibshirani *et al.* (2002) developed the prediction analysis of microarrays (PAM) method based upon an enhancement of the simple nearest prototype classifier. Recently, Koo *et al.* (2006) proposed the structured polychotomous machine (SPM) based on a functional analysis of variance decomposition using structured kernels.

In this paper we propose a variable selection method for multiclass LS-SVM, which uses results of multiclassification by LS-SVM. From the quadratic programming problem we obtain weights whose magnitudes imply the importance of variables on multiclassification.

The rest of paper is organized as follows. In Section 2 we present an overview of multiclass LS-SVM and model selection methods. In Section 3 we propose the variable selection method. In Section 4 we perform the numerical studies with real data sets. In Section 5 we give the concluding remarks.

2. Multiclass LS-SVM

2.1. LS-SVM

Let the training data set be denoted by $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, with each input $\boldsymbol{x}_i \in R^d$, the output $y_i \in R$. We consider the case of nonlinear regression. Then we take the form

$$f(\boldsymbol{x}) = \boldsymbol{w}^t \boldsymbol{\Phi}(\boldsymbol{x}) + b.$$

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

The optimization problem is defined with a regularization parameter $C > 0$ as

$$\text{Minimize } \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^n e_i^2 \tag{2.1}$$

over $\{\boldsymbol{w}, b, \boldsymbol{e}\}$ subject to equality constraints

$$y_i = \boldsymbol{w}^t\boldsymbol{\Phi}(\boldsymbol{x}_i) + b + e_i \ , \ i = 1, \cdots, n.$$

The Lagrangian function can be constructed as

$$L(\boldsymbol{w}, b, e : \alpha) = \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^n e_i^2 - \sum_{i=1}^n \alpha_i \ (\boldsymbol{w}^t\boldsymbol{\Phi}(\boldsymbol{x}_i) + b + e_i - y_i \ ), \tag{2.2}$$

where α_i's are the Lagrange multipliers. The conditions for optimality given by

$$\begin{aligned}
&\frac{\delta L}{\delta \boldsymbol{w}} = 0 \rightarrow \ \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\Phi}(\boldsymbol{x}_i) \\
&\frac{\delta L}{\delta b} = 0 \rightarrow \ \sum_{i=1}^{n} \alpha_i = 0 \\
&\frac{\delta L}{\delta e_i} = 0 \rightarrow \ e_i = \frac{1}{C}\alpha_i, \ i = 1, \cdots, n \\
&\frac{\delta L}{\delta \alpha_i} = 0 \rightarrow \ \boldsymbol{w}^t \boldsymbol{\Phi}(\boldsymbol{x}_i) + b + e_i - y_i = 0, \ i = 1, \cdots, n,
\end{aligned}$$

lead to the linear equation,

$$\begin{bmatrix} \boldsymbol{K} + \frac{1}{C}\mathbf{I}_n & \mathbf{1}_n \\ \mathbf{1}_n^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \tag{2.3}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones and $\boldsymbol{K}$ is the $n \times n$ kernel matrix with elements $K_{ij} = \boldsymbol{\Phi}(\boldsymbol{x}_i)^t \boldsymbol{\Phi}(\boldsymbol{x}_j), i, j = 1, \cdots, n$, which are obtained from the application of Mercer's conditions (1909). Solving the linear equation (2.3) the optimal bias and Lagrange multipliers, b and α_i's are obtained, then the optimal regression function for a test data point $\boldsymbol{x}_t^*$ is obtained as

$$\widehat{y}(\boldsymbol{x}_t^*) = \sum_{i=1}^{n} K(\boldsymbol{x}_t^*, \boldsymbol{x}_i)\alpha_i + b. \tag{2.4}$$

In the nonlinear case $\boldsymbol{w}$ is no longer explicitly given. However, it is uniquely defined in the weak sense by the dot products. Here the linear regression model can be regarded as the special case of the nonlinear regression model by using identity feature mapping function, that is, $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x}$ which implies the linear kernel matrix such that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^t \boldsymbol{x}_2$.

Note that it can be easily shown that Lagrange multipliers of LS-SVM for binary classification are identical to product of diagonal matrix of $\boldsymbol{y}$ and Lagrange multipliers of LS-SVM for regression obtained from equation (2.3), when $\boldsymbol{y}$ consists of class labels -1 and 1. That is, if $\boldsymbol{y}$ consists of class labels -1 and 1, $\widehat{\boldsymbol{y}}$ obtained by LS-SVMs for regression and classification are identical. Thus, for the binary classification, each observation of the test data can be classified into either class according to the sign of $\widehat{y}(\boldsymbol{x}_t^*)$ in (2.4) for $t = 1, \cdots, n_t$. See for details Shim *et al.* (2008). We use LS-SVM for regression, instead of LS-SVM for classification, to approximate the cross validation function of multiclass LS-SVM.

2.2. Multiclass LS-SVM using one-against-all method

In this section we give a simple overview on multiclass LS-SVM using one-against-all method (Shim *et al.*, 2008). Let the training data set be denoted by $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, with each input vector $\boldsymbol{x}_i \in R^d$ and the class label $y_i \in \{1, 2, \cdots, m\}$, where m is number of classes. For multiclassification using one-against-all method, we transform $\boldsymbol{y}$ into $n \times m$ matrix $\boldsymbol{Y}$ which consists of -1 and 1 such that $Y_{ij} = 1$ and $Y_{ik} = -1$ for $j \neq k$ implies that the i th observation belongs to the j th class. We have m LS-SVMs for binary classification with $\{\boldsymbol{x}_i, Y_{ij}\}_{i=1}^n$ for $j = 1, \cdots, m$. From the linear equation,

$$\begin{bmatrix} \boldsymbol{K} + \frac{1}{C}\mathbf{I}_n & \mathbf{1}_n \\ \mathbf{1}_n^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^j \\ b^j \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_{.j} \\ 0 \end{bmatrix}, \tag{2.5}$$

the optimal bias and Lagrange multipliers, b^j and α_i^j 's are obtained. Here $\boldsymbol{Y}_{.j}$ is the j th column of $\boldsymbol{Y}$.

For the test data point $\boldsymbol{x}_t^*$, we have

$$\widehat{Y}_{tj}(\boldsymbol{x}_t^*) = \sum_{i=1}^{n} K(\boldsymbol{x}_t^*, \boldsymbol{x}_i)\alpha_i^j + b^j \text{ for } t = 1, \cdots, , n_t. \tag{2.6}$$

Thus, if $\widehat{Y}_{tj}(\boldsymbol{x}_t^*) > 0$ and $\widehat{Y}_{tk}(\boldsymbol{x}_t^*) < 0$ for $k \neq j$ then the test data point $\boldsymbol{x}_t^*$ is classified into the j th class for $t = 1, \cdots, n_t$.

2.3. Model selection for multiclass LS-SVM

The functional structure of multiclass LS-SVM is characterized by hyperparameters, the regularization parameter C and the kernel parameters. To select the parameters of multiclass LS-SVM, we define a cross validation (CV) function as follows:

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(Y_{im_i} - \widehat{Y}_{im_i}^{(-i)}(\boldsymbol{\lambda}))^2, \tag{2.7}$$

where $\boldsymbol{\lambda}$ is the set of hyperparameters and $\widehat{Y}_{im_i}^{(-i)}(\boldsymbol{\lambda})$ is the predicted value of Y_{im_i} obtained from data without i th observation. Here m_i is the column number of the i th row of $\boldsymbol{Y}$ such that $Y_{im_i} = 1$, which implies that the i th observation belongs to the m_i th class. The CV function can be rewritten as

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(1 - \widehat{Y}_{im_i}^{(-i)}(\boldsymbol{\lambda}))^2. \tag{2.8}$$

Since for each candidates of hyperparameters, $\widehat{Y}_{im_i}^{(-i)}(\boldsymbol{\lambda})$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally formidable. By leaving-out-one lemma (Kimeldorf and Wahba, 1971) and the first order Taylor expansion, we have a generalized cross validation (GCV) function (Shim *et al.* 2008),

$$GCV(\boldsymbol{\lambda}) = \frac{n\sum_{i=1}^{n}(1 - \widehat{Y}_{im_i}(\boldsymbol{\lambda}))^2}{(n - trace(\boldsymbol{S}))^2}. \tag{2.9}$$

where $\boldsymbol{S}$ is the hat matrix obtained from the linear equation (2.5) such that $\widehat{\boldsymbol{Y}}_{.j} = \boldsymbol{S}\boldsymbol{Y}_{.j}$ for $j = 1, \cdots, m$.

3. Variable selection for multiclassification

We express the estimate of Y_{ij} as the weighted sum of $\widehat{Y}_{ij}^k$'s, $\widehat{Y}_{ij} = \sum_{k=1}^p c_k \widehat{Y}_{ij}^k$, where $\widehat{\boldsymbol{Y}}_{\cdot j}^k = \{\widehat{Y}_{ij}^k\}_{i=1}^n$ is obtained from the linear equation (2.5) with replacing $\boldsymbol{K}$ by $\boldsymbol{K}^k$ where $\boldsymbol{K}^k$ is the $n \times n$ kernel matrix constructed from $\{x_{ik}\}_{i=1}^n$ with x_{ik} the k th variable of the i th observation, $k = 1, \cdots, p$. The important variables can be selected according to magnitude of $c_k' s$, which are obtained by minimizing the objective function,

$$L(\boldsymbol{c}) = \sum_{i=1}^n (1 - \sum_{k=1}^p c_k \widehat{Y}_{im_i}^k)^2 \tag{3.1}$$

subject to $\sum_{k=1}^p c_k = 1$ and $c_k \geq 0$ for $k = 1, \cdots, p$. Here m_i is the column number of the i th row of $\boldsymbol{Y}$ such that $Y_{im_i} = 1$, which implies that the i th observation belongs to the m_i th class. The equation (3.1) can be rewritten as a quadratic programming problem,

$$\min L(\boldsymbol{c}) = \frac{1}{2}\boldsymbol{c}'\widehat{\boldsymbol{Y}}^{*\prime}\widehat{\boldsymbol{Y}}^{*}\boldsymbol{c} - \mathbf{1}_N'\widehat{\boldsymbol{Y}}^{*}\boldsymbol{c} \text{ subject to } \mathbf{1}'\boldsymbol{c} = 1 \text{ and } \boldsymbol{c} \geq \mathbf{0}, \tag{3.2}$$

where $\hat{\boldsymbol{Y}}^*$ is a $n \times p$ matrix with $\widehat{Y}_{ik}^* = \widehat{Y}_{im_i}^k$ for $i = 1, \cdots, n, k = 1, \cdots, p$.

To determine the optimal values of $\boldsymbol{c}$ which represent the importance of variables, we use the two stepwise procedure as follows.

1) Find $\widehat{Y}_{ik}^*$'s with the specified values of hyperparameters obtained from GCV function in (2.9).

2) Find $\widehat{\boldsymbol{c}}$ which minimizes the objective function $L(\boldsymbol{c})$ in (3.2).

4. Numerical studies

In this section we illustrates how well the proposed variable selection method works for selection of marker genes through real microarray data sets. To evaluate the performance of our proposed method in practice, we analyzed four publicly available microarray data sets: (i) Leukemia data set (Golub *et al.*, 1999). (ii) Lymphoma data set (Alizadeh *et al.*, 2000). (iii) Small Round Blue Cell Tumor (SRBCT) data set (Khan *et al.*, 1999). (iv) Brain tumor data set (Pomeroy *et al.*, 2002).

All data were transformed to the base 10 log scale, and the arrays were standardized for analysis. For each given data set, there is no applicable test set, so we performed 3-fold cross validation and examined classification error rates. This procedure was repeated 50 times to obtain necessary performance measures to compare with other methods. The radial basis function kernel was applied to SRBCT data set,

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\frac{1}{\sigma^2}||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2)$$

and the linear kernel was applied to the rest of data sets. The optimal values of hyperparameters for each data set are obtained by GCV function (2.9).

Table 4.1 Number of variables selected (number of classes in parenthesis)

	Proposed	SVM-RFE	PAM	SPM
Leukemia (3)	6.88	3.62	22.42	3.80
Lymphoma (3)	12.22	11.92	24.68	3.88
SRBCT (4)	8.40	14.24	18.12	4.96
Brain (5)	10.52	14.12	23.56	2.04

Table 4.2 Misclassification rates (standard error in parenthesis)

	Proposed	SVM-RFE	PAM	SPM
Leukemia	0.0442 (0.0092)	0.0833 (0.0057)	0.0633 (0.0062)	0.0708 (0.0066)
Lymphoma	0.0238 (0.0076)	0.0447 (0.0061)	0.1800 (0.0130)	0.0152 (0.0032)
SRBCT	0.0350 (0.0083)	0.0507 (0.0073)	0.0235 (0.0042)	0.0614 (0.0069)
Brain	0.3871 (0.0364)	0.3742 (0.0171)	0.4257 (0.0229)	0.3785 (0.0102)

Error rates and the average number of genes selected were compared between our method and three other methods: PAM, SPM and SVM-RFE. Results by three other methods are from Shim *et al.* (2009b). Tables 4.1 and 4.2 display the average numbers of the genes selected, mean error rates and standard errors, respectively. As shown in Table 4.1, the proposed method gives relatively smaller average numbers of the genes selected compared with SVM-RFE and PAM but larger compared with SPM. However, as shown in Table 4.2, the proposed method gives generally lower or almost same mean error rates for all four data sets. In particular, for Leukemia data set the proposed method gives remarkably lower mean error rates than other methods.

5. Concluding remarks

In this paper, we proposed a variable selection method to identify the important variables in multiclassification. To show the performance of the proposed variable selection method, we used four real data sets (Leukemia, Lymphoma, SRBCT, Brain), and we compared the proposed method with three other existing methods (SVM-RFE, PAM, SPM). The experimental results show that the proposed variable selection method has better performance in some data sets than existing methods. In addition, our variable selection method has the advantage that the computing time is much shorter in comparison to other existing methods.

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X. *et al.* (2000). Distinct types of diffuse large celllymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Cho, D. H., Shim, J. and and Seok, K. H. (2010). Doubly penalized kernel method for heteroscedastic autoregressive data. *Journal of Korean Data & Information Science Society*, **21**, 155-162.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M. and Downing, J. *et al.* (1999). Molecular classification of cancer: Classdiscovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 89-422.

Hwang, H. (2010). Fixed size LS-SVM for multiclassification problems of large data sets. *Journal of Korean Data & Information Science Society*, **21**, 1561-567.

Khan, J., Bittner, M. L., Saal, L. H., Teichmann, U., Azorsa, D. O., Gooden, G. C., Pavan, W. J., Trent, J. M. and Meltzer, P.S. (1999). cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. *Proceedings of the National Academy of Sciences*, **96**, 132 64-13269.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Koo, J. Y., Sohn, I., Kim, S. and Lee, J. W. (2006). Structured polychotomous machine diagnosis of multiple cancer types using gene expression. *Bioinformatics*, **22**, 950-990.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, **A**, 415-446.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P. and Lau, C., *et al.* (2002). Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature*, **415**, 436-442.

Scholkopf, B., Burges, C. and Vapnik, V. (1995). Extracting support data for a given task. *Proceedings of First Conference on Knowledge Discovery and Data Mining*, 252-257.

Shim, J., Bae, J. S. and Hwang, C. (2008). Multiclass classification via LS-SVR. *Communications of the Korean Statistical Society*, **15**, 441-450.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of Korean Data & Information Science Society*, **20**, 467-4720.

Shim, J., Park, H. and Hwang, C. (2009a). A kernel machine for estimation of mean and volatility functions. *Journal of Korean Data & Information Science Society*, **20**, 905-912.

Shim, J., Sohn, I., Kim, S., Lee, J.W., Green, P. E. and Hwang, C. (2009b). Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. *Computational Statistics and Data Analysis*, **53**, 1736-1742.

Suykens, J. A. K. and Vandewalle, J. (1999a). Least square support vector machine classifier. *Neural Processing Letters*, **9**, 293-300.

Suykens, J. A. K. and Vandewalle, J. (1999b). Multiclass least squares support vector machines. *Proceeding of the International Joint Conference on Neural Networks*, 900-903.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, **99**, 6567-6572.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, New York.

Vapnik, V. N. (1998). *Statistical learning theory*, Springer, New York.

Weston, J. and Watkins, C. (1998). *Multi-class SVM*, Technical Report 98-04, Royal Holloway University of London.

Zhang, H.H., Ahn, J., Lin, X. and Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, **22**, 88-95.