Journal of the Korean Data & Information Science Society
2009, **20**(4), 749–755

한국데이터정보과학회지

Sparse kernel classification using IRWLS procedure

Daehak Kim[1]

School of Computer & Information Communication Engineering,
Catholic University of Daegu

Received 10 June 2009, revised 8 July 2009, accepted 15 July 2009

Abstract

Support vector classification (SVC) provides more complete description of the linear and nonlinear relationships between input vectors and classifiers. In this paper. we propose the sparse kernel classifier to solve the optimization problem of classification with a modified hinge loss function and absolute loss function, which provides the efficient computation and the sparsity. We also introduce the generalized cross validation function to select the hyper-parameters which affects the classification performance of the proposed method. Experimental results are then presented which illustrate the performance of the proposed procedure for classification.

Keywords: Cross-validation, hinge loss function, iterative reweighted least squares, kernel function, support vector classification.

1. Introduction

Support vector machine (SVM), firstly developed by Vapnik (1995, 1998), is being used as a popular technique for classification and regression problems. SVM is based on the structural risk minimization (SRM) principle, which has been shown to be superior to traditional empirical risk minimization (ERM) principle. SRM minimizes an upper bound on the expected risk unlike ERM minimizing the error on the training data. By minimizing this bound, high generalization performance can be achieved. In particular for the SVC, SRM results in the regularized ERM with the hinge loss function. The introductions and overviews of recent developments of SVM can be found in Vapnik (1995, 1998), Gunn (1988), Smola and Schölkopf (1998). Oh *et al.* (2003) studied classification problem based on least squares SVM. Training an SVC requires the solution to a quadratic programming (QP) optimization problem. But QP problem presents some inherent limitations which results in computational difficulty especially for the large data sets. Platt (1998) developed the sequential minimal optimization (SMO) algorithm which divides the QP problem into a series of small QP problems to avoid such computational difficulty. Perez-Cruz et al. (2000) proposed iterative reweighted least squares (IRWLS) algorithm for SVM regression by transforming the Lagrangian function into sum of quadratic terms by defining associated weights

[1] Professor, School of Computer & Information Communication Engineering, Catholic University of Daegu, Gyeongsan 712-702, Korea. E-mail: dhkim@cu.ac.kr

of predicted errors. The IRWLS procedure can be applied to solve the QP problem of SVC with a modified hinge loss function of which original version is used by Vapnik (1995, 1998). The modified hinge loss function is attained by providing the differentiability at 1, which enables to solve QP problem by IRWLS procedure. Seok (2007) had studied semi-supervised learning by using kernel function.

Sparsity is known as an important feature of kernel machine models, which provides the efficiency on predicting the function. SVM does not provide extreme sparsity and the number of support vectors depends on the number of training data. Tipping (2001) proposed a Bayesian approach referred to as the relevance vector machine (RVM) providing more sparsity. However RVM has computational problems since there is no closed-form solutions for maximizing the marginal likelihood.

With a modified hinge loss function and absolute loss function, we propose sparse kernel classification (SKC) which provides both efficient computation including model selection and the sparsity. The rest of this paper is organized as follows. In Section 2 we give a simple review of SVC. In Section 3 we propose SKC using IRWLS procedure. We perform the numerical studies through examples in Section 4. In Section 5 we give the conclusions.

2. Support vector classification

Let the training data set D be denoted by $(\mathbf{x}_i, y_i)_{i=1}^n$, with each input vector $\mathbf{x}_i \in R^d$ including a constant 1 and the output $y_i \in \{-1, +1\}$ which is linearly or nonlinearly related to the input vector $\mathbf{x}_i$. Here the feature mapping function $\phi(\cdot) : R^d \to R_f^d$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. An inner product in feature space has an equivalent kernel in input space, $\phi(\mathbf{x}_i)'\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$ (Mercer, 1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible. We consider the nonlinear case, in which the classifier given $\mathbf{x}$ can be regarded as a nonlinear function of input vector $\mathbf{x}$.

With a hinge loss function $h(\cdot)$, the classifier can be defined as a function of any solution to the optimization problem,

$$\min \frac{1}{2}\mathbf{w}'\mathbf{w} + \lambda \sum_{i=1}^{n} h(y_i f(\mathbf{x}_i)), \tag{2.1}$$

where $h(r) = 0$ if $r \geq 1$ and $h(r) = 1 - r$ if $r < 1$ and $\mathbf{w}$ is an appropriate weight vector. We can express the classification problem by formulation for SVC as follows.

$$\min \frac{1}{2}\mathbf{w}'\mathbf{w} + C \sum_{i=1}^{n} \xi_i \tag{2.2}$$

subject to

$$y_i \mathbf{w}'\phi(\mathbf{x}_i) \geq 1 - \xi_i \quad \xi_i \geq 0$$

where λ is a regularization parameter penalizing the training errors. Also, we construct a Lagrange function as follows:

$$L = \frac{1}{2}\mathbf{w}'\mathbf{w} + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i \mathbf{w}'\phi(\mathbf{x}_i) - 1 + \xi_i) - \sum_{i=1}^{n} \eta_i \xi_i. \tag{2.3}$$

We notice that the positivity constraints $\alpha_i, \eta_i \geq 0$ should be satisfied. After taking partial derivatives of equation (2.3) with regard to the primal variables $(\mathbf{w}, \xi_i, b)$ and plugging them into equation (2.3), we have the optimization problem below.

$$\max -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \alpha_i y_i \tag{2.4}$$

with constraints

$$\alpha_i \in [0, \lambda].$$

Solving the above equation with the constraints determines the optimal Lagrange multipliers α_i. Thus, the classifier given the input vector $\mathbf{x}$ is obtained as

$$\tilde{y}(\mathbf{x}) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x})). \tag{2.5}$$

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

3. Sparse kernel classification using IRWLS procedure

In this section we propose a sparse kernel classification using IRWLS procedure to solve the QP problem of SVC and ℓ_1 norm penalty term with a modified hinge loss function and a modified absolute loss function which are differentiable at 1 and 0, respectively. The modified hinge loss function $h_\delta(\cdot)$ is attained by providing the differentiability at 1 by differing from the original hinge loss function $h(\cdot)$ in the small interval $(1-\delta,\, 1+\delta)$,

$$h_\delta(r) = \frac{1}{4\delta}(1 - r - \delta)^2 I(r \geq 1 - \delta) + (1 - r) I(r < 1 - \delta), \tag{3.1}$$

where $\delta > 0$ and $I(\cdot)$ is an indicative function.

To have the sparsity on estimation of α, we use a Laplacian prior (Williams,1995),

$$p(\alpha) \propto \exp(-\lambda ||\alpha||_1), \tag{3.2}$$

where $||\alpha||_1 = \sum_i |\alpha_i|$ denotes ℓ_1 norm and λ is a nonnegative constant. Since $||\alpha||_1$ is not differentiable with respect to α, we need a modification of $||\alpha||_1$ for IRWLS procedure.

Now the problem (2.1) becomes obtaining α to minimize,

$$L(\alpha) = \sum_{i=1}^{n} h_\delta(y_i(K_i Y \alpha)) + \lambda \sum_{i=1}^{n} g_\delta(\alpha_i), \tag{3.3}$$

where K_i is the i-th row of K and $g_\delta(\cdot)$ is attained by providing the differentiability at 0 by differing from the original absolute loss function in the small interval $(-\delta,\, \delta)$,

$$g_\delta(\alpha) = \frac{1}{\delta}\alpha^2 I(-\delta < \alpha \leq \delta) - \alpha I(\alpha \leq -\delta) + \alpha I(\alpha > \delta), \tag{3.4}$$

where $\delta > 0$ and $I(\cdot)$ is an indicative function.

Taking partial derivatives of (3.3) with regard to α leads to the optimal values of α which is the solution to

$$\mathbf{0} = -HW\mathbf{1} + HWH\alpha + U\alpha. \tag{3.5}$$

Here W is a diagonal matrix with the i-th diagonal element w_{ii} obtained from the derivative of the modified hinge loss function as

$$w_{ii} = \frac{1}{2\delta(1-r_i)}(1 - r_i - \delta)I(r_i \geq 1-\delta) + \frac{1}{1-r_i}I(r_i < 1-\delta), \tag{3.6}$$

where $r_i = y_i\hat{y}_i = y_i(K_iY\alpha)$. U is the diagonal matrix consisted of $\partial g_\delta(\alpha_i)/\partial\alpha_i$, $i = 1, \cdots, n$.

The solution to (3.5) cannot be obtained in a single step since W and U contain α. Thus we need to apply IRWLS procedure.

At t-th iteration of IRWLS procedure, $\alpha^{(t+1)}$ is obtained as

$$\alpha^{(t+1)} = (HW(\alpha^{(t)})H + \lambda U(\alpha^{(t)}))^{-1}YKW(\alpha^{(t)})\mathbf{y}. \tag{3.7}$$

During iteration, we find that some α_i's tend to zero keeping the value of objective function $L(\alpha)$ decreasing. This motivates that we can find sparse estimates of α which decrease the value of the objective function $L(\alpha)$ at the same time.

Algorithm of SKC using IRWLS procedure is given as follows:

(a) Set $v = (1:n)'$ and $\alpha(v)^{(0)}$.

(b) Find $\alpha(v)^{(t+1)}$ from (3.7).

(c) Set $\alpha_i^{(t+1)} = 0$ which is very close to zero.
Find $v = \{i : \alpha_i^{(t+1)} \neq 0\}$.

(d) Iterate (b)-(c) until $L(\alpha)$ converges.

To select the hyper-parameters of SKC using IRWLS we define the leave-one out cross validation (LOO-CV) function as follows:

$$CV(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_\theta^{(-i)}(\mathbf{x}_i))^2, \tag{3.8}$$

where θ is the set of hyper-parameters and $\hat{y}_\theta^{(-i)}(\mathbf{x}_i)$ is the classifier of $\mathbf{x}_i$ estimated without i-th observation. Since for each candidate of hyper-parameters, $\hat{y}_\theta^{(-i)}(\mathbf{x}_i)$ for $i = 1, \cdots, n$, should be evaluated. So the selection of parameters using CV function is computationally burdensome. By applying the leave-out-one lemma (Craven and Wahba, 1979) and the first order Taylor expansion, we have GCV function as follows,

$$GCV(\theta) = \frac{n\sum_{i=1}^{n}(y_i - \hat{y}_\theta(\mathbf{x}_i))^2}{(n - tr(S))^2}, \tag{3.9}$$

where $S = KY(H'WH + \lambda U)^{-1}YKW$, is the hat matrix such that $\hat{y}_\theta(\mathbf{x}) = S\mathbf{y}$. For SVC, $\hat{y}_\theta(\mathbf{x})$ cannot be expressed in the linear combination of y_i's, we cannot use GCV function but LOO-CV function or k-fold CV function whose computation time is much longer than GCV function.

4. Numerical studies

We illustrate the performance of SKC using IRWLS of Section 3 by comparing with that of SVC, and the relevance vector classification (RVC) through the Pima Indian data set and the Iris data set available at UCI Machine Learning Repository (www.ics.uci.edu/∼mlearn/MLRepository.html or http://archive.ics.uci.edu/ml/). We repeated 100 times dividing randomly the data set into the training data set and the test data set. δ of the modified hinge loss function was set to 0.0001, the linear kernel and the gaussian kernel are used for Pima Indian data set and Iris data set, respectively. For the model selection of SVC and RVC we use 10-fold cross-validation function.

Pima Indian data set case

The sizes of the training and test data sets are 384 and 384. The averages of 100 misclassification error rates and their standard errors of SVC, SKC, RVC are obtained as (0.2345, 0.2342, 0.6491) and (0.0017, 0.0016, 0.0015), respectively. The averages of numbers of retained kernel functions and their standard errors are obtained as (201.49, 10.53, 5.94) and (1.10, 0.167, 0.09), respectively. Figure 4.1 (right) shows the boxplot of numbers of retained kernel functions. From the figure 4.1, we can see that SKC provides slightly better result than other methods in this example. We can see that SKC provides more sparsity than SVC. Also we can see RVC provides the least number of retained kernel functions but SKC provides the less number of retained kernel functions than SVC.

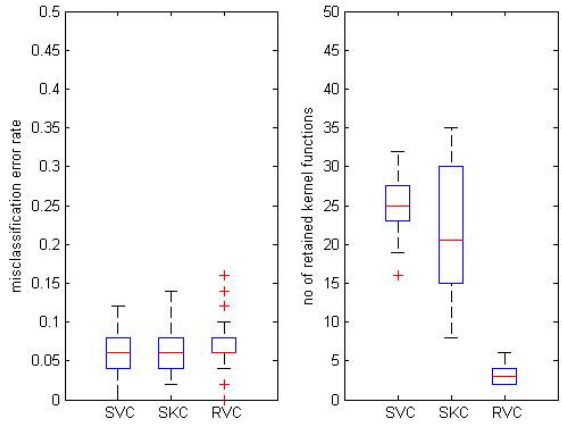

Figure 4.1 Misclassification error rates (left) and numbers of retained kernel functions (right) by SVC, SKC, and RVC, respectively, for Pima Indian data set.

Iris data set case

We use 100 data from two classes, Virginia and Setosa, which are known to be classified uncorrectly with the linear kernel. The sizes of the training and test data sets are 50 and

50. The averages of 100 misclassification error rates and their standard errors of SVC, SKC, RVC are obtained as (0.0612, 0.0616, 0.0702) and (0.0023, 0.0029, 0.0032), respectively. The averages of numbers of retained kernel functions and their standard errors are obtained as (25.02, 21.65, 3.77) and (0.31, 0.83, 0.10), respectively. Figure 4.2 (left) show the boxplot of 100 misclassification error rates obtained by SVC, SKC, and RVC, respectively. As a pima indian data set, we can also find that SKC provides slightly better result than other methods in this example. From the figure 4.2, we can see that SKC provides more sparsity than SVC and that RVC provides the least number of retained kernel functions but SKC provides the less number of retained kernel functions than SVC.

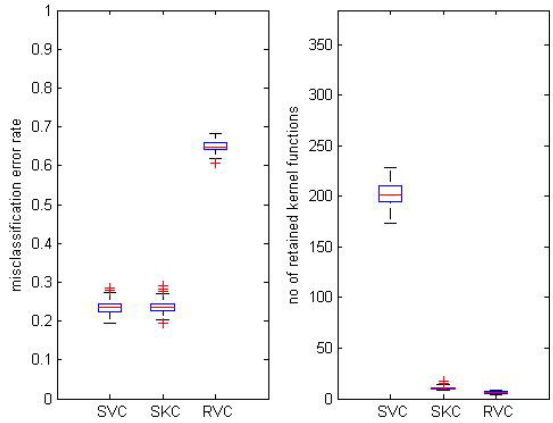

Figure 4.2 Misclassification error rates (left) and numbers of retained kernel functions (right) by SVC, SKC, and RVC, respectively, for Iris data set.

5. Conclusions

In this paper, we dealt with obtaining the classifier by SKC using IRWLS procedure. Through the example we showed that the proposed procedure derives the satisfying results. We found that SKC provides the faster computation in training and model selection than SVC. Also we found that the classification accuracy is almost same as SVC and the sparsity is smaller than SVC.

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik*, **31**, 377-390.

Gunn, S. (1998). Support vector machines for classification and regression. *ISIS Technical Report*, University of Southhampton.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society,* Series A, **209**, 415-446.

Oh, K., Shim, J. and Kim, D. (2003). Incremental multi-classification by least squares support vector machine. *Journal of Korean Data & Information Science Society*, **14**, 965-974.

Perez-Cruz, F., Navia-Vazquez, A., Alarcon-Diana, P. L. and Artes-Rodriguez, A. (2000). An IRWLS procedure for SVR. In *Proceedings of European Association for Signal Processing*, EUSIPO 2000, Tampere, Finland.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Microsoft Research Technical Report,* MSR-TR-98-14.

Seok, K. H. (2007). Semi-supervised learning using kernel estimation. *Journal of Korean Data & Information Science Society*, **18**, 629-636.

Smola, A. and Scholkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica,* **22**, 211-231.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research,* **1**, 211-244.

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

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

Williams, P. M. (1995). Bayesian regularization and pruning using a laplace prior. *Neural Computation,* **7**, 117-143.