Clustering Model Selection for Reduced Support Vector Machines*

Lih-Ren Jen and Yuh-Jye Lee **
lrjen@niu.edu.tw, yuh-jye@mail.ntust.edu.tw

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, 106 Taiwan

Abstract. The reduced support vector machine was proposed for the practical objective that overcomes the computational difficulties as well as reduces the model complexity by generating a nonlinear separating surface for a massive dataset. It has been successfully applied to other kernel-based learning algorithms. Also, there are experimental studies on RSVM that showed the efficiency of RSVM. In this paper we propose a robust method to build the model of RSVM via RBF (Gaussian kernel) construction. Applying clustering algorithm to each class, we can generate cluster centroids of each class and use them to form the reduced set for RSVM. We also estimate the approximate density for each cluster to get the parameter used in Gaussian kernel. Under the compatible classification performance on the test set, our method selects a smaller reduced set than the one via random selection scheme. Moreover, it determines the kernel parameter automatically and individually for each point in the reduced set while the RSVM used a common kernel parameter which is determined by a tuning procedure.

Keywords: Gaussian Kernel, k-means Clustering Algorithm, Model Selection, Radial Basis Function Network, Reduced Set, Support Vector Machine.

1 Introduction

In recent years support vector machines (SVMs) with linear or nonlinear kernels [1, 3, 10] have become one of the most promising learning algorithms. For the binary classification problems, SVMs are able to construct a nonlinear separating surface (if it is necessary), which is implicitly defined by a kernel function [10]. However, there are some major computational difficulties such as huge memory usage and long CPU time, in generating a nonlinear SVM classifier for a massive dataset. To overcome these difficulties the reduced support vector machine (RSVM) [4] was proposed.

* This is a revision for an original paper in *Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning*, Lecture Notes in Computer 3177, pages 714-719, Springer-Verlag, Exeter, UK, 2004.

** Corresponding author

In this paper, we apply the k-means clustering algorithm to each class to generate cluster centroids of each class and then use them to form the reduced set that is randomly selected in RSVM [4]. One of the most important ideas of SVM is kernel technique that uses a kernel function to represent the inner product of two data points in the feature space after a nonlinear mapping. We will use the Gaussian kernel through this paper. The value of the Gaussian kernel can be interpreted as a measure of similarity between data points. In this case, the reduced kernel matrix records the similarity between the reduced set and the entire training dataset. This observation inspires us to select the most representative points of the entire training dataset to form the reduced set. Using the cluster centroids should be an intuitive heuristic. In order to catch the characteristic of each class we run the clustering algorithm on each class separately. This idea originally comes from [6]. The Gaussian kernel function contains a tuning parameter σ, which determines the shape of the kernel function. Choosing this tuning parameter is called the model selection which is a very important issue in nonlinear support vector machine. A smaller value of this parameter will give a better discriminate ability on training examples while may cause the overfitting risk, fitting the training data too well but losing the prediction ability on unseen data. In practice, the conventional SVM as well as RSVM determines this tuning parameter which is commonly used in kernel function via a tuning procedure [2]. While, in our approach the kernel parameter is determined automatically and individually for each point in the reduced set. This can be achieved by estimating the approximate density of each resulting cluster [9]. Once we have the reduced kernel matrix, we apply smooth support vector machine [4] to generate the final classifier. We apply our method on four benchmark datasets from the UCI Machine Learning Repository [8] and the face detection dataset[1]. Under the compatible classification performance on the test set, our method selects a smaller reduced set than the one via random selection scheme. Moreover, it determines the kernel parameter automatically and individually for each point in the reduced set while the RSVM used a common kernel parameter which is determined by a tuning procedure.

We briefly outline the contents of the paper and a word about our notation is given below. Section 2 provides the main idea and formulation for RSVM. In Section 3, we explain why we could use the former research results of RBFN, and describe our algorithm. The experimental results of our method are presented in Section 4. In Section 5, we conclude this paper. All notations used in the paper are listed as follows. All vectors will be column vectors unless otherwise specified or transposed to a row vector by a prime superscript $'$. The plus function x_+ is defined as $(x)_+ = \max\{0, x\}$. The scalar (inner) product of two vectors x and z in the n-dimensional real space R^n will be denoted by $x'z$ and the p-norm of x will be denoted by $\|x\|_p$. For a matrix $A \in R^{m\times n}$, A_i is the ith row of A which is *a row vector* in R^n. A column vector of ones of arbitrary dimension will be denoted by e. For $A \in R^{m\times n}$ and $B \in R^{n\times l}$, the kernel $K(A, B)$ maps $R^{m\times n} \times R^{n\times l}$ into $R^{m\times l}$. In particular, $K(x', z)$ is a real number, $K(x', A')$ is a

[1] Available at http://www.ai.mit.edu/projects/cbcl/.

row vector in R^m, $K(A,x)$ is a *column* vector in R^m and $K(A,A')$ is an $m \times m$ matrix. The base of the natural logarithm will be denoted by ε.

2 Reduced Support Vector Machines

We now briefly describe the RSVM formulation, which is derived from the generalized support vector machine (GSVM) [7] and the smooth support vector machine (SSVM) [5]. We are given a training dataset $\{(x^i, y_i)\}_{i=1}^m$, where $x^i \in R^n$ is an input data point and $y_i \in \{-1, 1\}$ is class label, indicating one of two classes, A_- and A_+, to which the input point belongs. We represent these data points by an $m \times n$ matrix A, where the ith row of the matrix A, A_i, corresponds to the ith data point. We denote alternately A_i (a row vector) and x^i (a column vector) for the same ith data point. We use an $m \times m$ diagonal matrix D defined by $D_{ii} = y_i$ to specify the membership of each input point. The main goal of the classification problem is to find a classifier that can predict the label of new unseen data points correctly. This can be achieved by constructing a linear or nonlinear separating surface, $f(x) = 0$, which is implicitly defined by a kernel function. We classify a test point x belonging to A_+ if $f(x) \geq 0$, otherwise x belonging to A_-. We will focus on the nonlinear case that is implicitly defined by a Gaussian kernel function. The RSVM solves the following unconstrained minimization problem

$$\min_{(\bar{v},\gamma)\in R^{\bar{m}+1}} \frac{\nu}{2}\|p(e - D(K_\sigma(A,\bar{A}')\bar{v} - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(\bar{v}'\bar{v} + \gamma^2), \tag{1}$$

where the function $p(x,\alpha)$ is a very accurate smooth approximation to $(x)_+$ [5], which is applied to each component of the vector $e - D(K_\sigma(A,\bar{A}')\bar{v} - e\gamma)$ and is defined componentwise by

$$p(x,\alpha) = x + \frac{1}{\alpha}\log(1 + \varepsilon^{-\alpha x}), \alpha > 0. \tag{2}$$

The function $p(x,\alpha)$ converges to $(x)_+$ as α goes to infinity. The reduced kernel matrix $K_\sigma(A,\bar{A}') \in R^{m\times\bar{m}}$ in (1) is defined by

$$K_\sigma(A,\bar{A}')_{ij} = \varepsilon^{-\frac{||A_i - \bar{A}_j||_2^2}{2\sigma^2}}, \tag{3}$$

where $\bar{A}$ is the reduced set that is randomly selected from A in RSVM [5]. The positive tuning parameter ν here controls the tradeoff between the classification error and the suppression of $(\bar{v},\gamma)$. Since RSVM has reduced the model complexity via using a much smaller rectangular kernel matrix we will suggest using a larger tuning parameter ν here. The solution of this minimization problem (1) for $\bar{v}$ and γ leads to the nonlinear separating surface

$$f(x) = \bar{v}'K_\sigma(\bar{A},x) - \gamma = \sum_{i=1}^{\bar{m}} \bar{v}_i K_\sigma(\bar{A}_i,x) - \gamma = 0. \tag{4}$$

The minimization problem (1) can be solved via the Newton-Armijo method [5] directly and the existence and uniqueness of the optimal solution of this problem are also guaranteed. We note that this nonlinear separating surface (4) is a linear combination of a set of kernel functions $\left\{1, K_\sigma(\bar{A}_1, \cdot), K_\sigma(\bar{A}_2, \cdot), \cdots, K_\sigma(\bar{A}_{\bar{m}}, \cdot)\right\}$, where σ is the kernel parameter of each kernel function. In next section, we will apply the k-means algorithm to each class to generate cluster centroids and then use these centroids to form the reduced set. Moreover we also give a formula to determine the kernel parameter σ for each point in the reduced set automatically.

3 Clustering Reduced Support Vector Machine

We propose our new algorithm, Clustering RSVM (CRSVM), that combines the RSVM [4] and RBF networks algorithm together. The most popular RBF networks can be describe as

$$f(x) = w_0 + \sum_{h=1}^{\bar{m}} w_h \, \varepsilon^{-\frac{||x - c^h||_2^2}{2\sigma_h^2}}, \tag{5}$$

where $c^h = (c_1^h, c_2^h, ..., c_n^h)$ is also a vector in the n-dimensional vector space and $||x - c^h||_2$ is the distance between training (test) vectors x and c^h. We can use the same decision rule in previous section for binary classification. That is, we classify a test point x belonging to A_+ if $f(x) \geq 0$, otherwise x belonging to A_-. By RBFN approaches, we have to choose three parameters (c^h, σ_h ,w_h) in equation (5) based on the training dataset. For the first two parameters, many RBFN approaches were proposed that apply variant clustering algorithms such as k-means to training set to generate the cluster centroids as c^h. The parameter σ_h is estimated based upon the distribution of clusters. Based on uniform distribution assumption, [9] estimates σ_h as

$$\sigma_h = \frac{\overline{R}(c^h) \cdot \delta \cdot \sqrt{\pi}}{\sqrt[n]{(r+1)\Gamma(\frac{n}{2}+1)}} \quad \text{, where } \delta \cdot \sqrt{\pi} = 1.6210 \tag{6}$$

and $\overline{R}(c^h)$ is defined as

$$\overline{R}(c^h) = \frac{n+1}{n}(\frac{1}{r}\sum_{q=1}^{r} ||\hat{x}_q - c^h||_2), \tag{7}$$

where $\hat{x}_1, \ldots, \hat{x}_r$ are the r nearest samples to the cluster centroid c^h. If the cluster size is smaller than r, we use the all examples in this cluster to compute $\bar{R}(c^h)$.

When the first two type-variables are selected, RBFN is trained to get the w_h and the estimation of σ_h as kernel parameter to generate the reduced kernel matrix [9]. We proposed a variant RSVM method that uses clustering centroids as reduced set. The Clustering Reduced Support Vector Machine (CRSVM) algorithm is described below.

Algorithm 3.1 Clustering Reduced Support Vector Machine:

Let k be the number of cluster centroids for each class and r be a positive integer.
STEP1. For each class, runs k-means algorithm to find the cluster centroids c^h. Use the clustering results to form the reduced set $\bar{A} = [c^1 c^2 ... c^{2k}]'$.
STEP2. For each centroid c^h, computes the corresponding kernel parameter σ_h using the formula (6, 7).
STEP3. Let A_i denote the ith training point, use the resulting parameters from STEPs 1 and 2 to construct the rectangular kernel matrix $K_\sigma(A, \bar{A}')_{ih} = \varepsilon^{-\frac{||A_i - c^h||_2^2}{2\sigma_h^2}}$, where $K \in R^{m \times 2k}$, for $i = 1, 2, ..., m$ and $j = 1, 2, ..., 2k$.
STEP4. Apply the Newton-Armijo Algorithm [4] to solve the problem (1), where $K_\sigma(A, \bar{A}')$ is the reduced kernel matrix obtained in Step 3.
STEP5. The separating surface is given as formula (4), where $(\bar{v}^*, \gamma^*) \in R^{\bar{m}+1}$ is the unique solution of problem (1) that got from Step 4.
STEP6. A new unseen data point $x \in R^n$ is classified as class +1 if $\bar{v^*}' K_\sigma(\bar{A}, x) - \gamma^* \geq 0$, otherwise x is classified as class -1.

The conventional SVMs as well as RSVM determine parameter used in kernel function via a tuning procedure [2]. While, in our approach the kernel parameter is determined automatically and individually for each point in the reduced set. This can be achieved by estimating the approximate density of each resulting cluster [9]. The numerical results are shown in Section 4.

4 Numerical Results

We normalized the dataset such that each attribute has 0-mean and 1-deviation, so that we can assume that each attribute has the similar contribution to the Gaussian kernel. In our experiment, the normalization procedure is very crucial.

The numerical results on four benchmark datasets and a real one are shown in Table 1. SSVM stands for nonlinear smooth support vector machine with full kernel. RSVM1 and RSVM2 stand for RSVM [4] with different size of reduced set. CRSVM used the same size of reduced set with RSVM1 which is smaller than the one used in RSVM2. We note that CRSVM only used a smaller reduced set than random selection scheme with compatible classification performance.

Classifier Dataset	CRSVM correctness(%) $\bar{m}$, time(sec.)	RSVM1 correctness(%) $\bar{m}$, time(sec.)	RSVM2 correctness(%) $\bar{m}$, time(sec.)	full kernel SSVM correctness(%) m, time(sec.)
Ionosphere 351 × 34	95.7 14, 2.98	94.4 14, 2.52	95.19 35, 3.64	94.35 351, 40.16
BUPA 345 × 6	73.4 14, 2.51	70.4 14, 1.31	74.86 35, 4.31	73.62 345, 34.25
Pima 768 × 8	77.6 30, 12.3	77.8 30, 6.31	78.64 50, 7.47	76.59 768, 234.8
Cleveland 297 × 13	85.7 12, 1.93	84.1 12, 1.01	86.47 30, 3.47	85.92 297, 27.14
Face Detection 6977 × 361	98.2 16, 280.97	95.2 16, 125.3	96.7 24, 205.4	96.7 6977, out of memory

Table 1. Results of benchmarks (Test set correctness of ten-fold cross validation).

5 Conclusion

In this paper we propose a robust method to build the model of RSVM via RBF (Gaussian kernel) construction. Applying clustering algorithm to each class, we can generate cluster centroids of each class and use them to form the reduced set for RSVM. We also estimate the approximate density for each cluster to get the kernel parameter that is used in Gaussian kernel. Under the compatible classification performance on the test set, our method selects a smaller reduced set than the one via random selection scheme. Moreover, it determines the kernel parameter automatically and individually for each point in the reduced set while the RSVM used a common kernel parameter which is determined by a tuning procedure.

References

1. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
2. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.
3. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.
4. Y.-J. Lee and O. L. Mangasarian. RSVM: Reduced support vector machines. Technical Report 00-07, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, July 2000. Proceedings of the First SIAM International Conference on Data Mining, Chicago, April 5-7, 2001, CD-ROM Proceedings. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-07.ps.
5. Y.-J. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001. Data Mining Institute, University of Wisconsin, Technical Report 99-03. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.
6. A. Lyhyaoui, M. Martinez, I. Mora, M. Vazquez, J.-L. Sancho, and A. R. Figueiras-Vidal. Sample selection via clustering to construct support vector-like classifier. *IEEE Transactions on Neural Networks*, 10:1474–1481, 1999.
7. O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.
8. P. M. Murphy and D. W. Aha. UCI machine learning repository, 1992. www.ics.uci.edu/∼mlearn/MLRepository.html.
9. Y.-J. Oyang, S.-C. Hwang, Y.-Y. Ou, C.-Y. Chen, and Z.-W. Chen. A novel learning algorithm for data classification with radial basis function networks. In *Proceeding of 9th International Conference on Neural Information Processing*, pages 18–22, Singapore, Nov. 2001.
10. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.