2011 International Conference on Information and Electronics Engineering
IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore

Training Data Selection for Support Vector Machines Model

Dang Huu Nghi[1]and Luong Chi Mai[2]

[1]Hanoi University of Mining and Geology, Tu Liem, Ha Noi. E-mail: dhnghi2005@yahoo.com

[2]Vietnamse Academy of Science and Technology, Institute of Information Technology, 18 Hoang Quoc Viet Road, Ha Noi. E-mail: lcmai@ioit.ac.vn

Abstract. In recent years, Support Vector Machines (SVM) have become a popular tool for pattern recognition and machine learning. In training a support vector machine we need to select the parameters of the kernel function as well as soft margin parameter C of SVM. Thus to develop the optimal classifier we need to determine the optimal kernel parameter and the optimal value of C. Determining the optimal classifier is called model selection. When applied to a large data set, however, it requires a long time for training so the model selection task and its performance can be degraded a long time. To speed up training thereby shortening the time for model selection, several methods have been proposed, one of which is to reduce the training set size called training data selection. Recently, there has been considerable research on data selection for SVM training. The main idea is to select only the patterns that are likely to be located near the decision boundary. In this paper we propose a methods that select a subset of data for SVM training. Our experimental results show that a significant amount of training data can be removed by our proposed method without degrading the performance of the resulting SVM classifiers.

Keywords: Support vector machines, model selection, training data selection.

1. Introduction

Support vector machines are very powerful classifiers in theory but their effiency in practice rely on an optimal selection of hyper-parameters. This hyper-parameter estimation with respect to the aforementioned performance measures is often called the model selection problem. One method often used to select the parameters is grid search on the log ratio of the parameters associated with cross-validation. When applied to a large data set, however, it requires a long time for training so the model selection task and its performance can be degraded a long time. To speed up training thereby shortening the time for model selection, several methods have been proposed, one of which is to reduce the training set size called data selection. Recently, there has been considerable research on data selection for SVM training. For example, Shin and Cho proposed a method that selects patterns near the decision boundary based on the neighborhood properties [7]. In [2], k-means clustering is employed to select patterns from the training set. In [9], Zhang and King proposed a β_skeleton algorithm to identify support vectors. In [1], Abe and Inoue used Mahalanobis distance to estimate boundary points. In [8] Y. Koshiba proposed a method that for each datum belonging to class 1, find the datum belonging to class 2 with the minimum distance in input space then add these data into training set. In this page we propose a method that is variants of the method propsed in [8] on data selection. After that we perform model selection for SVM on the training data selected. Our experimental results show that a significant amount of training data can be removed by our proposed method without degrading the performance of the resulting SVM classifiers.

2. Support Vector Machine

Support vector machines are extensively used as a classification tool in a variety of areas. They map the input (x) into a high-dimensional feature space ($z = \phi(x)$) and construct an optimal hyperplane defined by w.z + b = 0 to separate examples from the two classes. For SVMs with L1 soft-margin formulation, this is done by solving the primal problem [5][6].

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i \qquad (1)$$

With constraints:

$$y_i\left(\langle w.x_i\rangle - b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1....N \tag{2}$$

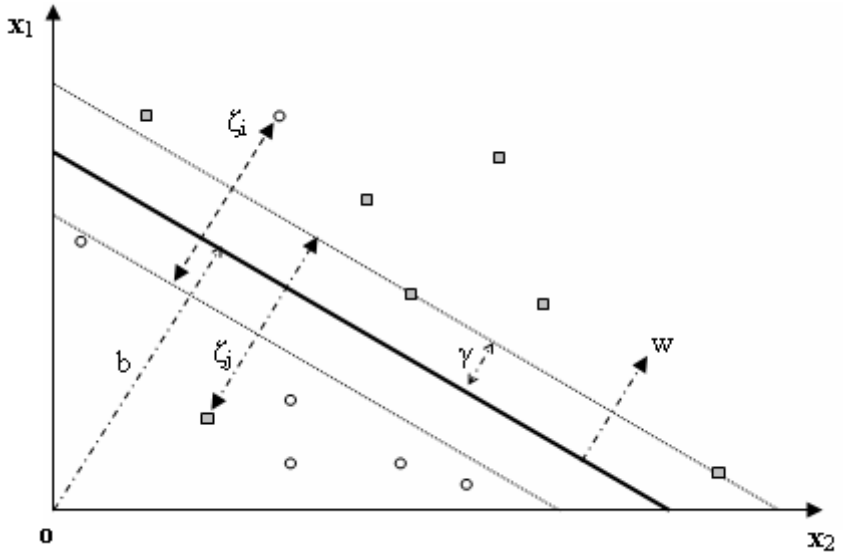

Fig. 1. SVM soft margin and classification problem kernel.

The problem of the Lagrange dual form is:

$$\max \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k\left(x_i, x_j\right) \tag{3}$$

With constraints

$$0 \leq \alpha_i \leq C, \forall i, \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \tag{4}$$

In which $k(x_i, x_j) = \Phi(x_i).\ \Phi(x_j)$ is a kernel function of the nonlinear mapping implementation. A number of commonly used kernel functions are:

Gaussian kernel: $$k\left(x_i, x_j\right) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)$$

Polynomial kernel: $$k\left(x_i, x_j\right) = \left(1 + x_i.x_j\right)^d$$

RBF kernel: $$k\left(x_i, x_j\right) = \exp\left(-\gamma \| x_i - x_j \|^2\right)$$

To obtain a good performance, some parameters in SVMs have to be chosen carefully. These parameters include the regularization parameter C, which determines the tradeof between minimizing the training error and minimizing model complexity and parameter (d, σ or γ) of the kernel function.

In this page we test the classification using the kernel function RBF so two parameters need to be choice is the γ width of the RBF function and soft margin parameter C of SVM. Choice of parameters is usually done by minimizing generalization error was estimated as k-fold cross-validation error or leave-one-out error (LOO). With k-fold cross-validation error first the training set is divided into k subsets of the same size then turn evaluated a subset using hyperplane was drawn from the training k-1 subset left.

One method often used to select the parameters is grid search on the log ratio of the parameters associated with cross-validation. Value pairs (C, γ), respectively was assessed using cross-validation and then choose the pair with highest precision. The value of C and γ are increasing exponentially (eg C = 2^{-6}, 2^{-3},……, 2^{14}; γ = 2^{-15}, 2^{-13},……, 2^{10}) [3][4].

3. Training data selection

According to the architecture of the support vector machine, only the training data near the boundaries are necessary. In addition, because the training time becomes longer as the number of training data increases, the training time is shortened if the data far from the boundary are deleted. Therefore, if we can delete unnecessary data from the training data effciently prior to training, we can speed up training.

In [8] Y. Koshiba proposed a method for training data selection is as follows:

1. For each datum belonging to Class 1, find the datum belonging to Class 2 with the minimum distance in the input space which are included in the reduced training set (Fig. 2).
2. Iterate Step 1 exchanging Class 1 and Class 2.

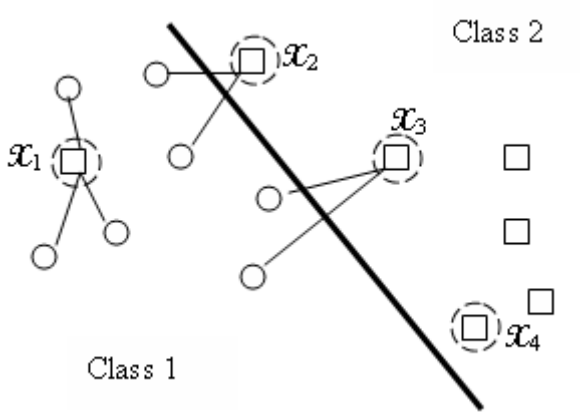

Fig. 2. Find the datum belonging to Class 2 with the minimum distance

As we can see in figure 2 three patterns x_1, x_2 and x_3 are selected and x_4 is not selected by step 1. However x_1 is noisy pattern which should be identified and removed. The x_4 pattern need to be selected. To overcome this problem we propose a method that is variants of the method propsed by Y. Koshiba as follows:

The first, for each training example x_i we compute the number of training patterns that are contained in the largest sphere centered at a training patterns without covering an pattern of a same class. We denote this number by $N(x_i)$. To remove the noisy patterns, we delete k% (of the size of the training sets) patterns with largest numbers $N(x_i)$ from the original training set.

The second, we select only the patterns that are likely to be located near the decision boundary. Like the method propsed by Y. Koshiba, for each datum belonging to class 1 in training set, find the datum belonging to class 2 with the minimum distance in the input space but among the selected pairs with the same class 2 datum, we select the pair with the minimum distance which are included in the reduced training set.

Description of the algorithm:

1. Compute the number of training patterns that are contained in the largest sphere centered at a training patterns without covering an pattern of a same class $N(x_i)$.
2. Delete k% (of the size of the training set) patterns with largest numbers $N(x_i)$ from the original training set.
3. For each datum belonging to class 1 in training set received form step 2, find the datum belonging to class 2 with the minimum distance in the input space (Fig 3.a)
4. Among the selected pairs with the same class 2 datum, select the pair with the minimum distance which are included in the reduced training set (Fig 3.b).
5. Iterate Steps 3 and 4 exchanging class 1 and class 2.

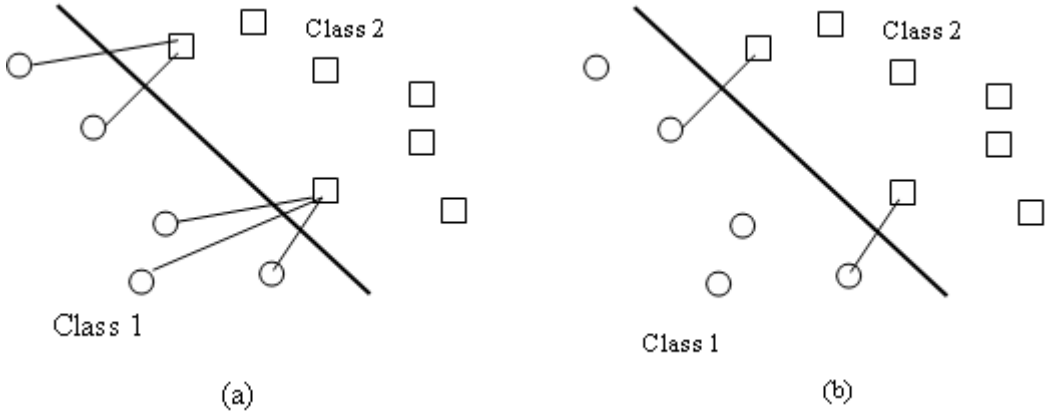

Fig. 3: Select the patterns that are likely to be located near the decision boundary

4. Experiments

SVM is a new technique used for regression and classification data. SVM usually depends on some parameters, problem posed is how to choose the parameters of SVM so as to obtain the best results. Here we focus on the selection parameter γ of the RBF function and soft margin parameter C of SVM for classification problems. We perform grid search on C and γ using 5 fold cross-validation with $C \in [2^{-6}, 2^{14}]$ and $\gamma \in [2^{-15}, 2^{10}]$. We ran the software LIBSVM [3] on Acer Aspire 5573ZWXMI NoteBook.

To speed up training thereby shortening the time for parameter selection we implemented training data selection and then applied SVM model selection over the reduced training set. In order to test the effectiveness of the proposed method, a series of experiments were carried out. We test methods on the 5 benchmark datasets was downloaded from http://www.csie.ntu.edu.tw/ ~ cjlin /libsvmtools/datasets.

Table 1: Test datasets

Dataset	Training	Test	Features	Classes
Astroparticle	3089	4000	4	2
Vehicle	1243	41	21	2
Satimage	3104	2000	36	6
Splice	1000	2175	60	2
Letter	15000	5000	61	26

Table 2: Compares the accuracy rates and size between reduced training set by Y. Koshiba's proposed method, reduced training set by our proposed method with k =1 and original training set.

Dataset	Reduced training set by Y. Koshiba's proposed method		Reduced training set by our proposed method		Original training set	
	Size	Accuracy	Size	Accuracy	Size	Accuracy
Astroparticle	298	96.02%	358	96.55%	3089	96.87%
Vehicle	421	82.93%	485	87.80%	1243	87.80%
Satimage	974	81.55%	1207	89.50%	3104	90.35%
Splice	373	52%	523	52%	1000	52%
Letter	7350	96.26%	8743	96.78%	15000	97.04%

Table 3: Computional time spent for training data selection and model selection on reduced training data set. Results over original training data sets are also reported.

Dataset	Reduced training data set (h:m:s)	Original training data set (h:m:s)
Astroparticle	0:3:01	0:8:40
Vehicle	0:6:20	0:38:03
Satimage	0:16:50	0:58:00
Splice	0:4:30	0:13:46
Letter	7:50:15	13:51:00

5. Conclusion

The selection optimum values of the parameters for SVM is an important step in SVM design. When applied to a large data set, however, it requires a long time for training so the model selection task and its performance can be degraded a long time. To reduce the time for model selection, in this page we propose a training data selection method then aplly the model selection on reduced training set.

Our experimental results show that a significant amount of time for model selection can be saved by our proposed method without degrading the performance of the resulting SVM classifiers**.**

6. References

[1] Abe, S., Inoue, T. *Fast training of support vector machines by extracting boundary data*. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN) (2001) 308–313.

[2] Almeida, M. B., Braga, A. P., Braga, J. P. *SVM-KM: speeding SVMs learning with a priori cluster selection and k-means*. In Proceedings of the 6th Brazilian Symposium on Neural Networks (2000) 162–167.

[3] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, *A practical guide to Support Vector Classification*, Libsvm: a library for support vector machines, 2005.

[4] Marcelo N. Kapp, Robert Sabourin, Patrick Maupin, *A PSO-Based Framework for Dynamic SVM Model Selection*, GECCO'09, Montréal Québec, Canada, 2009.

[5] Nello Cristianini, John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[6] Shigeo Abe, Support *Vector Machines for Pattern Classification*, Springer, 2005.

[7] Shin, H. J., Cho, S. Z. *Fast pattern selection for support vector classifiers*. In Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence (LNAI 2637) (2003) 376–387

[8] Y. Koshiba. Acceleration *of training of support vector machines*. Master's thesis, Graduate School of Science and Technology, Kobe University, Japan, 2004.

[9] Zhang, W., King, I. *Locating support vectors via β_skeleton technique*. In Proceedings of the International Conference on Neural Information Processing (ICONIP) (2002) 1423–1427