Support Vector Machines Based on Set Covering

Jiaqi Wang, and Chengqi Zhang

Abstract—**Support Vector Machines (SVMs) have been the promising method in the field of machine learning. But for the real applications there are still some drawbacks in SVMs, e.g. the high training cost and too many support vectors. This paper presents a novel method based on set covering to overcome these drawbacks, called SC-SVMs. Some experiments on real data show the effectiveness of this new method.**

Index Terms—**SVMs, Set Covering, Kernel**

I. Introduction

SUPPORT Vector Machines (SVMs) have been one of promising methods in many fields of information technology such as pattern recognition, regression analysis, clustering, data compression and so on [1]-[3]. Their success depends on two excellent ideas. First, SVMs maximize the margin such that the good generalization performance can be guaranteed and second, the kernel functions are used in SVMs to overcome the curse of dimension.

While SVMs have the advantage of the generalization performance over the traditional learning methods such as Neural Networks, Decision Trees, etc., they still suffer from some drawbacks compared to other methods. On one hand, the training cost can be high since the quadratic programming is required in SVMs. On the other hand, testing unseen samples can be time consuming because there may be very many support vectors in the learning model obtained by SVMs.

A lot of work has been done to speed up the training of SVMs. There are two most important ideas, which speed up the training of SVMs, chunking [4] and shrinking [5]. They use the heuristics to reduce the size of the training set. In addition, the reduced set [6], [7] and the sequential approaches [8], [9] have been presented to reduce the number of support vectors.

This paper presents a new method, SVMs based on Set Covering (SC-SVMs), to improve the efficiency of training and testing of the traditional SVMs at the same time. Set covering is another important machine learning method [10], [11] and its training efficiency is much higher than SVMs. In addition, we find that the number of support vectors obtained by SC-SVMs is low compared to the traditional SVMs. Some experiments on the real data verify the effectiveness of SC-SVMs.

The remainder of the paper is organized as follows. SVMs are introduced in Section II. SC-SVMs are presented in Section III. Some experiments on real data are provided to support SC-SVMs in Section IV. The conclusion is given in Section V.

II. SVMs Overview

The theory and method a bout SVMs are originally established by Vapnik et al [1] and have been very popular in 1990's. SVMs benefit from two good ideas: maximizing the margin and the use of the kernel function. The former guarantees the good generalization performance of the leaming model and the latter can overcome the curse of dimension.

For the classification problem, the traditional SVMs solve the following quadratic optimization problem.

$$\min \ \frac{1}{2}\|\omega\|^2 + C\sum_i \xi_i \qquad (1)$$
$$s.t. \ (\langle \omega, x_i \rangle + b) y_i \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\ldots, l$$

The example of solving the linear classification problem using the linear SVMs is shown as Fig. 1.

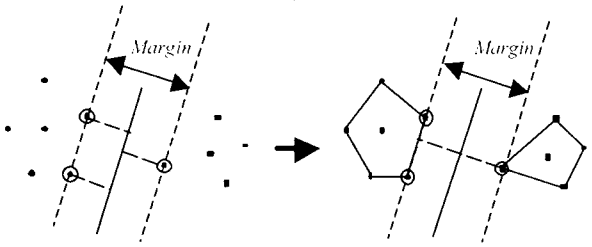


Fig. 1. (from [12]) SVMs maximize the margin between two linear separable sample sets. Maximizing the margin is equivalent to finding the shortest distance between two disjoint convex hulls spanned by the two linear separable sample sets.

For the regression analysis, the traditional SVMs solve the following quadratic optimization problem.

$$\min \ \frac{1}{2}\|\omega\|^2 + C\left(\sum_i \xi_i + \sum_i \xi_i^*\right) \qquad (2)$$
$$s.t. \ \begin{array}{l} y_i - (\langle \omega, x_i \rangle + b) \leq \varepsilon + \xi_i^*, \\ (\langle \omega, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i, \\ \xi_i \geq 0, \xi_i^* \geq 0, i = 1,2,\ldots, l \end{array}$$

In addition, SVMs use the kernel functions to solve the non-linear learning problems. Now the popular kernel functions include the polynomial function, the Gaussian radius basis function (RBF), and the sigmoid function. The example of solving the non-linear classification problem using the kernel function is shown as Fig. 2.

Jiaqi Wang is a PhD student in the Faculty of Information Technology, University of Technology, Sydney, Australia, (telphone: 61-2-9514-4534, e-mail: jqwang@it.uts.edu.au).

Chengqi Zhang is a professor in the Faculty of Information Technology, University of Technology, Sydney, Australia, (telphone: 61-2-9514-7941, e-mail: chengqi@it.uts.edu.au).

ICITA2004 ISBN 0-646-42313-4

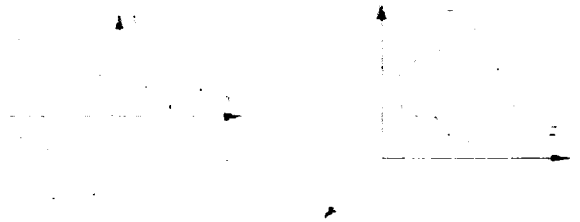


Fig. 2. (from [13]) The samples are mapped from the 2-dimensional original space to the 3-dimensional feature space. The non-linear classification problem is changed to the linear classification problem by using the kernel mapping.

Many researchers tried their best to improve the efficiency of solving the quadratic optimization problems in (1) and (2). The ideas on chunking and shrinking are presented just for reducing the training time. Chunking solves the sub-problems by iteratively building a set of samples that includes those violating the optimization conditions the most. By contrast, shrinking temporarily removes the samples from the training set that are not likely to become support vectors and then train the model. The efficiency of SVMs using chunking and shrinking is much higher than that of the traditional methods solving the quadratic optimization problem.

In addition, one always wishes to reduce the number of support vectors since this affects the efficiency of testing and the rate of data compression. The two important ways to solve this problem have been presented. One is to compress support vectors into a reduced set by solving a complex optimization problem. The other is a sequential approach, that is, the learning model stops computing the output of the unseen sample as soon as the output meets some statistical criteria.

III. SVMs Based on Set Covering

This section introduces SC-SVMs, which can reduce the training time and the number of support vectors for the classification problem. The idea on set covering has been presented in the field of machine learning [10], [11]. Set covering, e.g. ball covering, is a classic geometric problem. It can be run efficiently compared to solving a quadratic optimization problem. So in this paper the idea about set covering is applied to improve the traditional SVMs and this method is called SC-SVMs.

SC-SVMs includes the following steps:

Step1, the radius r of set covering is set.

Step2, a sample x is randomly selected.

Step3, the samples with the same class label as x within the ball, whose center is x and radius is r, are removed from the training set.

Step4, the samples in the training set are removed based on the above principle until no samples can be removed.

Step5, the compressed training set, whose size is much smaller than that of the original training set, is generated.

Step6, the compressed training set is trained using the traditional SVMs.

Step7, the learning model based on SC-SVMs is built.

An example of set covering is shown as Fig. 3. There are twenty positive and negative samples respectively in this example. After set covering is run, forty original samples are reduced to only six samples *C1*, *C2*, ..., *C6*. From Fig. 3, the statistical distribution of the original sample set is not drastically changed. So while the efficiency of training is improved and the number of support vectors is reduced, the generalization performance is guaranteed to some extent.

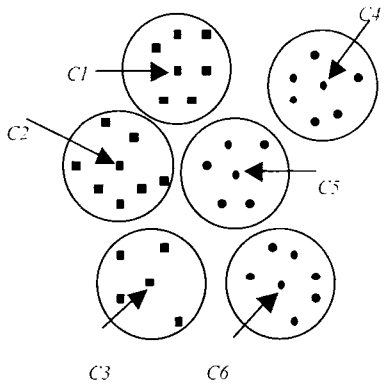


Fig. 3. An example of set covering.

SC-SVMs use the Gaussian RBF

$$K(x, y) = e^{-\frac{\|x - y\|^2}{2\sigma^2}}$$

as the kernel function because of its good properties [14].

Property1, for any sample x in the sample set, $K(x, x) = 1$.

Property2, for any two different samples x, y in the sample set,

$$0 < K(x, y) < 1.$$

IV. Experiments

This section provides two experiments on real data to verify the effectiveness of SC-SVMs. "*Compression Rate*" in Tables 1 and 2 means

No. of compressed samples / No. of original samples,

where compressed samples are those obtained by set covering from the original samples. For example, "*Compression Rate = 100*" means that the whole original training set is optimized by SVMs. In addition, Gaussian RBF kernel is used in all the following experiments. All the following experiments are performed on a computer with Pentium4 CPU and 256M Memory.

Experiment1. This experiment is performed on U.S. Postal Service (USPS) data [15], which is often used to evaluate the performance of the classification model. It includes a lot of real handwritten digits (1-10) similar to those shown in Fig. 1. There are 7291 training samples and 2007 testing samples in USPS data.

The values of all parameters in this experiment are set as follows:

1. Gaussian RBF kernel: $sigma = 8$.
2. The penalty factor in (1): $C = 5$.

The results obtained by SC-SVMs are shown in Table 1. "*Compression Rate = 100*" means that the traditional SVMs are used to build the model. From Table 1, we can find that SC-SVMs improve the efficiency of training and reduce the number of support vectors compared to the traditional SVMs. Moreover,

the generalization performance (test correct rate) of SC-SVMs is comparable to that of the traditional SVMs.

Fig. 1. Some handwritten digit samples are shown and some of them are atypical.

Table 1. The results of running SC-SVMs on USPS data.

Compression Rate (%)	Compression Time (s)	Training Time (s)	No. of Support Vector	No. of Test Errors
100	0	32.366	1449	91
55.48	11.276	18.777	1275	95
37.9	6.92	12.428	1138	100
16.36	2.433	4.256	763	125

Experiment2. This experiment is conducted on UCI Adult benchmark data set [16]. This data includes 22698 training samples and 9866 testing samples with 123 binary attributes. The task is to predict whether the household has an income greater than $50,000 using the census form of a household.

The values of all parameters in this experiment are set as follows:

1. Gaussian RBF kernel: $sigma = 10$.
2. The penalty factor in (1): $C = 10$.

The results obtained by SC-SVMs are shown in Table 2. From Table 2, we find that when a lot of samples are removed from the original training set by using set covering, the efficiency of training is obviously improved and the support vectors are reduced. For example, when 4.49% of the original training samples are trained, both the learning time and the number of support vectors are more than one hundred times less than those by the original samples. Moreover, the generalization performance is comparable to the traditional SVMs and even when *Compression Rate* = 24.49%, the generalization performance (*Test Correct Rate* = 85.2%) is better than that (Test Correct Rate = 84.96%) of the original samples.

Table 2. The results of training using SC-SVMs on UCI Adult data.

Compression Rate (%)	Compression Time (s)	Training Time (s)	No. of Support Vector	Test Correct Rate (%)
100	0	586.373	8213	84.96
24.49	38.615	46.677	2586	85.20
4.49	2.894	1.982	636	83.39
4.29	2.634	1.823	615	82.94
1.81	0.881	0.43	301	79.76

V. CONLUSION

This paper presents a new method to train SVMs, called SC-SVMs. Based on the good property of set covering and the Gaussian RBF kernel, SC-SVMs reduce the training cost and the number of support vectors compared to the traditional SVMs. So this new method will be helpful for many applications such as, classification, regression, clustering, data compression, and so on. In further work, we will verify SC-SVMs on more real data set.

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory, 2nd ed.*, Springer-Verlag, New York, 1999.

[2] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, " Support Vector Cluster," *Journal of Machine Learning Research*, vol. 2, pp. 125-137, 2001.

[3] J. Robinson and V. Kecman, " The Use of Support Vector Machines in Image Compression" , *Proceedings of the EIS' 2000, Second International ICSC Symposium on Engineering of Intelligent Systems*, June, 2000.

[4] B. E. Boser, I. M. Guyon, and V. Vapnik, " A Training Algorithm for Optimal Margin Classifiers," *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 144-152, ACM Press, 1992.

[5] T. Joachims, " Making Large-scale SVM Learning Practical," *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1998.

[6] C. Burges, " Simplified Support Vector Decision Rules," *International Conference on Machine Learning*, 1996.

[7] C. Burges amd B. Schö lkopf, " Improving the Accuracy and Speed of Support Vector Machines," *Advances in Neural Information Processing Systems*, 1997.

[8] S. Romdhani, P. Torr, B. Schö lkopf, and A. Blake, " Computationally Efficient Face Detection," *International Conference on Computer Vision*, 2001.

[9] D. DeCoste and D. Mazzoni, " Fast Query-Optimized Kernel Machine Classification via Incremental Approximate Nearest Support Vectors," *International Conference on Machine Learning*, 2003.

[10] L. Zhang and B. Zhang, " A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications," *IEEE Transactions on Neural Networks*, vol. 10(4), pp. 291-295, 1999.

[11] M. Marchand and J. Shawe-Taylor, " The Set Covering Machine," *Journal of Machine Learning Research*, vol. 3, pp. 723-746, 2002

[12] J. Q. Wang, Q. Tao, and J. Wang, " Kernel Projection Algorithm for Large-scale SVM Problems," *Journal of Computer Science and Technology*, vol. 17(5), pp. 556-564, 2002.

[13] K. R. Muller , S. Mika, G. Ratsch, K. Tsuda, and B. Schö lkopf, " An Introduction to Kernel-based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12(2), pp. 181-201, 2001.

[14] J. Q. Wang, C. Q. Zhang, X. D. Wu, H. W. Qi, and J. Wang, " SVM-OD: a New SVM Algorithm for Outlier Detection," *Foundations and New Directions of Data Mining Workshop in IEEE International Conference of Data Mining*, 2003.

[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel, " Handwritten Digit Recognition with Back-propagation Network," *Advances in Neural Information Processing Systems*, 1990.

[16] J. Platt, " Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999.

Second International Conference on Information Technology & Applications

Conference Organising Committee

Conference Chair Prof Shi Guangfan Heilongjiang University, China

Conference Vice-Chair Dr. Fu Yuzuo Shanghai Jiao-tong University, China

Technical Chair Dr. Sean He University of Tech Sydney, Australia

International Advisor Dr Dapeng Tien Charles Sturt University, Australia

ICITA 2004 is organised by:

Heilongjiang University, Harbin, China

http://www.hlju.edu.cn

And supported by:

Shanghai Jiao Tong University, Shanghai, China

http://www.sjtu.edu.cn

IEEE, NSW Section, Australia

http://ewh.ieee.org/r10/nsw/

IEEE, CS Chapter, Beijing, China

http://www.cie-china.org/ieee-beijing/

Saora Inc., Japan

http://www.saora.com/

ICITA 2004 Harbin, China http://www.icita.org

Second International Conference on Information Technology & Applications

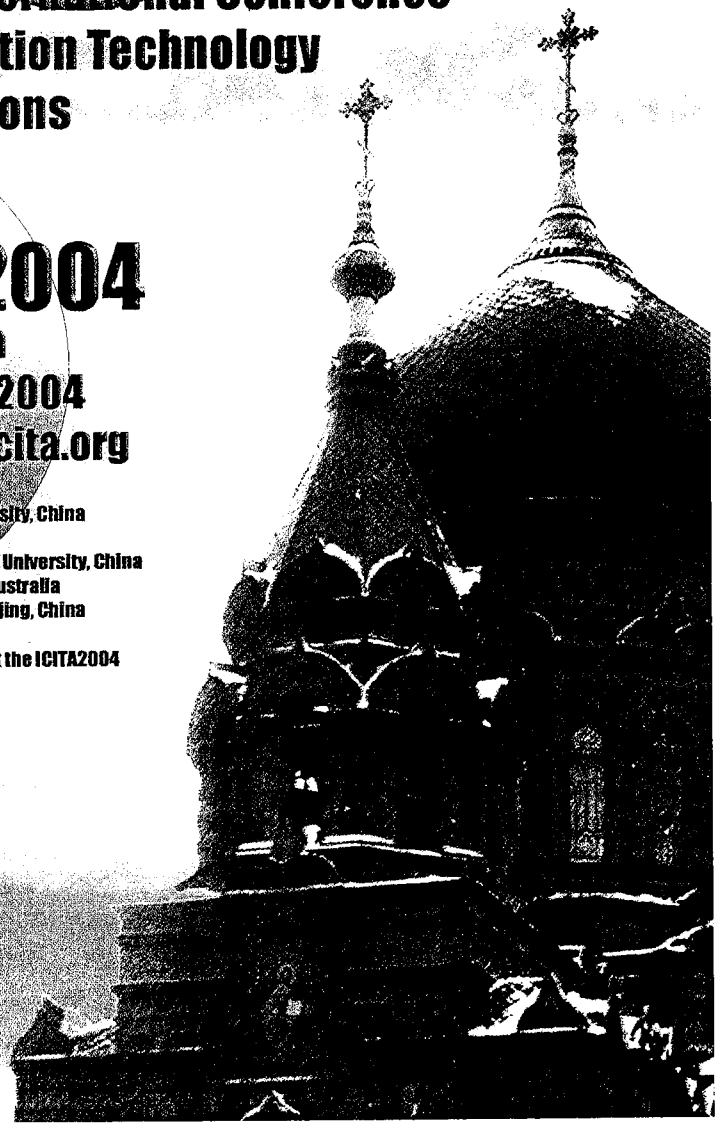
ICITA 2004

Harbin, China

8-11 January 2004

http://www.icita.org

Organised by
Heilongjiang University, China
And supported by:
Shanghai Jiao Tong University, China
IEEE NSW Section, Australia
IEEE, CS Chapter, Beijing, China
Saora Inc., Japan
papers published at the ICITA2004 will be EI indexed.



IEEE