International Conference on Information and Automation
June 20 -23, 2008, Zhangjiajie, China

A Fast Two-Stage Classification Method of Support Vector Machines

Jin Chen, Cheng Wang, *Member, IEEE,* and Runsheng Wang
ATR Laboratory, School of Electronic Science and Engineering
National University of Defense Technology
47 Yanwachi, Changsha 410073, China
chenjin_wonder@hotmail.com

***Abstract*-Classification of high-dimensional data generally requires enormous processing time. In this paper, we present a fast two-stage method of support vector machines, which includes a feature reduction algorithm and a fast multiclass method. First, principal component analysis is applied to the data for feature reduction and decorrelation, and then a feature selection method is used to further reduce feature dimensionality. The criterion based on Bhattacharyya distance is revised to get rid of influence of some binary problems with large distance. Moreover, a simple method is proposed to reduce the processing time of multiclass problems, where one binary SVM with the fewest support vectors (SVs) will be selected iteratively to exclude the less similar class until the final result is obtained. Experimented with the hyperspectral data 92AV3C, the results demonstrate that the proposed method can achieve a much faster classification and preserve the high classification accuracy of SVMs.**

I. Introduction

Pattern classification is important due to emerging applications such as hyperspectral classification, protein classification, speech recognition, and so on. Compared to traditional classification approaches, support vector machines (SVMs) have been found to be particularly promising because of its lower sensitivity to the curse of dimensionality [1]. The high generalization ability of SVMs is ensured by special properties of the optimal hyperplane that maximizes the distance to training examples in a high dimensional feature space [2]. Another important property is their good generalization capability supported by their sparse representation of the decision function.

However, in many applications, data are represented by high dimensional feature vectors and a large number of classes. Both situations increase the computational complexity of test phase of SVMs. As a result, it seems that, for such classification problems, SVMs may not be comparable to traditional classifiers, such as maximal likelihood classification (MLC) method, in terms of test time. In the literature, dimensionality reduction is motivated mainly by the consideration of classification speed [3].

Dimensionality reduction mainly consists of feature selection and feature extraction approaches. Feature selection methods can be further classified into two categories: filter and wrapper methods [4]. The filter method employs intrinsic properties of data such as Mahalanobis class separability measure as the criterion, while the wrapper method evaluates feature subsets based on the performance of the classifier such as classification error rate. Feature extraction methods mainly including principal component analysis (PCA), independent component analysis (ICA), and kernel principal component analysis (KPCA), a comparison of these methods for dimensionality reduction in SVMs can be see in [5].

SVMs were originally designed for binary classification. One-against-all (OAA) [6] and one-against-one (OAO) [7] [8] are the two most common methods to address the multiclass classification problem. The discrimination of OAA between an information class and all others often leads to the estimation of complex discriminant functions [9]. OAO needs $C(C\text{-}1)/2$ binary SVMs for one classification, which may result in slow classification.

To obtain a faster classification, direct acyclic graph SVM (DAGSVM) [10] and binary tree of SVM (BTS) [11] were proposed to reduce the number of binary SVMs of OAO. DAGSVM only needs C-1 binary SVMs, and BTS needs $\log_{4/3}((C+3)/4)$ binary SVMs on average for one classification. There are also other multiclass SVM methods, which try to achieve higher classification accuracy, such as pairwise decision tree of SVM (PDTSVM) [12] and error correcting output codes (ECOC) methods [13]-[15]. PDTSVM selects binary SVMs with larger geometric margin and reduces the layers to decrease the accumulated errors, while ECOC methods use the error correcting coding theory to improve the decision accuracy.

Besides, reduced set methods [16], which try to approximate the original solution by a much smaller number of newly constructed support vectors (SVs), were also proposed to obtain a fast classification of SVMs.

In this paper, we propose a fast two-stage method for classification with SVMs, depicted in Fig. 1. First, it is carried out by a feature selection algorithm after decorrelation with principal component analysis (PCA). For the feature selection, we revised the criterion based on Bhattacharyya distance to get rid of influence of some binary problems with large distance. In order to further reduce the computation complexity, a simple method called fast OAO (FOAO) is proposed to combine C-1 binary SVMs with the fewest support vectors. Experimented on an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data set, the results demonstrate that the proposed method can be much faster than different multiclass SVM

978-1-4244-2184-8/08/$25.00 © 2008 IEEE.

methods with original feature set, while the classification accuracies almost remain the same. The results also show that it can even be faster than maximum likelihood classifier (MLC) with the same feature set, while the classification accuracy of proposed method is much better.

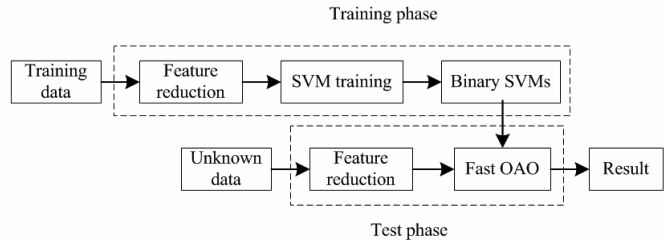

Fig. 1. Block diagram of the proposed fast two-stage method.

II. Feature Reduction

In this section, we present a feature reduction method. First, principal component analysis is applied to the data for feature reduction and decorrelation, and then a feature selection method is used to further reduce feature dimensionality. The criterion for feature selection is based on Bhattacharyya distance. We also revise it to get rid of influence of some binary problems with large distance.

Bhattacharyya distance has been used for feature selection and is known to give lower and upper bounds of the classification error [17]. For normal distributions, the Bhattacharyya distance is given by [18]:

$$B = (1/8)(\mu_i - \mu_j)^T((\Sigma_i + \Sigma_j)/2)^{-1}(\mu_i - \mu_j) + (1/2)\ln\left(\left|(\Sigma_i + \Sigma_j)/2\right| \Big/ \sqrt{|\Sigma_i||\Sigma_j|}\right) \quad (1)$$

where μ_i and Σ_i denote the mean vector and covariance matrix of class *i*.

Assuming equal a priori probabilities, the bounds are given by [19]:

$$\frac{1}{2}\left(1-\sqrt{1-e^{-2B}}\right) \le \varepsilon \le \frac{1}{2}e^{-B} \quad (2)$$

where ε is the classification error, and *B* is the Bhattacharyya distance. To reduce the classification error, we should increase the Bhattacharyya distance between the two classes. In other words, if some features can effectively increase the Bhattacharyya distance, they will have more contribution to classify the two classes. Thus, these feature sets can be selected to classify the two classes instead of all the features.

If the data of classes *i* and *j* are uncorrelated, equation (1) becomes [20]:

$$B = \sum_{n=1}^{N} B_{ij}^n = \sum_{n=1}^{N}[(1/4)(\mu_i(n)-\mu_j(n))^2(\sigma_i^2(n)+\sigma_j^2(n))^{-1} + (1/2)\ln(\sigma_i^2(n)/2+\sigma_j^2(n)/2)(\sigma_i^2(n)\sigma_j^2(n)^{-1/2})] \quad (3)$$

where $\mu_i(n)$, $\sigma_i^2(n)$ represent, respectively, the mean and variance of the *n*th band of class *i* and *N* is the number of features .

This suggests that when the data have low correlation (close to zero), like those after principal component analysis (PCA), the class separability of a set of features is mainly determined by the individual feature separability and can be estimated by summing these single feature separability. The contribution of feature *n* to class *i* and class *j* is denoted by B_{ij}^n.

For a multiclass problem with *C* classes, we can use the following equation B^n to denote the contribution of feature *n*:

$$B^n = \sum_{i=1}^{C}\sum_{j=i+1}^{C} B_{ij}^n \Big/ \max_{m=1,2,\ldots,N}(B_{ij}^m) \quad (4)$$

where $B_{ij}^n / \max_{m=1,2,\ldots,N}(B_{ij}^m)$ can be seen as the relative Bhattacharyya distance of feature *n* for binary problem of class *i* and class *j*. It can get rid of influence of some binary problems with large distance and can select features important for all binary problems, which can result in a good accuracy.

III. Fast One-Against-One SVM

OAA [6] and OAO [7] [8] are the two most common methods to address the multiclass classification problem. OAA trains each class against the remaining *C*-1 classes that have been collected together. The "winner-takes-all" rule is used for the final decision, where the winning class is the one corresponding to the SVM with the highest output (discriminant function value).

OAO needs to train $C(C\text{-}1)/2$ binary SVMs, where each one is trained on data from two information classes. When testing, for each information class ω_i, score will be computed by a score function:

$$S_i(\mathbf{x}) = \sum_{j=1,\, j\neq i}^{C} \operatorname{sgn}\{f_{ij}(\mathbf{x})\} \quad (5)$$

where $f_{ij}(\mathbf{x})$ is the discriminant function trained for classes ω_i and ω_j. Then, the unlabeled sample $\mathbf{x}$ will be associated with the class with the largest score.

To obtain a fast multiclass classification, DAGSVM and BTS were proposed to reduce the number of binary SVMs for one classification. DAGSVM only needs *C*-1 binary SVMs, and BTS needs $\log_{4/3}((C+3)/4)$ binary SVMs on average for one classification.

The number of binary SVMs can affect the test time. However, reducing the number of binary SVMs for one classification can not always assure a faster classification, since the computational complexity of a binary test is $O(n_{SV})$ where n_{SV} is the number of SVs.

In practical problems, the number of SVs may vary a lot from each other. Take 92AV3C problem for example (an eleven-class problem of hyerspectral data used in the experiments), from Fig. 2, we can see the largest number of SVs is much larger (about 32 times) than the smallest number of SVs among the binary SVMs of OAO. As a result, binary SVMs with the fewest number of SVs should be selected to achieve a fast classification.

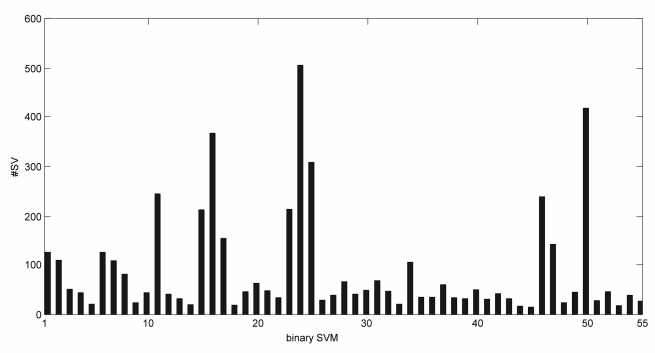

Fig. 2. The numbers of SVs of OAO binary SVMs for problem 92AV3C.

Based on such analysis, we propose a simple method called fast one-against-one (FOAO) to iteratively select one binary SVM with the fewest number of SVs. Each selected binary SVM will be used to exclude one less similar class. After excluding C-1 classes, we can assign the unknown sample to the remaining class. Accordingly, FOAO reduces the number of binary SVMs of OAO to C-1, and selects binary SVMs with fewest SVs.

After training $C(C\text{-}1)/2$ binary SVMs for each two-class problems, the classification process includes the following five steps.

1). Add all information classes to a class list.
2). Select the binary SVM with the fewest number of SVs, where both two classes trained for the selected binary SVM should be in the class list.
3). Use the selected binary SVM to exclude the less similar class and update the class list.
4). If the class list contains more than one class, then go to step 2.

Otherwise, output the remaining class as the classification result.

IV. Experiments and Results

Experiments use an AVIRIS data set covering an area of mixed agriculture and forestry landscape in the Indian Pine Test Site in northwestern Indiana (see Fig. 3). The data set is well-known in the literature [1] [9] [12]. It was recorded in June 1992 with 220 bands. Water absorption bands, 104-108 and 150-162, were removed, leaving 202 bands for analysis. Eleven classes including "corn", "corn-min", "corn-notill", "grass/pasture", "grass/trees", "hay-windrowed", "soybean-clean", "soybean-min", "soybean-notill", "woods", and "wheat" were selected. There are 9791 data points.

All the data were scaled into [-1, 1] and were randomly partitioned into two parts. 75% of the original data were used for training and 25% of the original data were used for testing. To assess the influence of the number of training data, we further varied the number of training samples drawn from the training set such that 30%, 45%, 60%, 75% of the original data were used for training while maintaining a constant testing set.

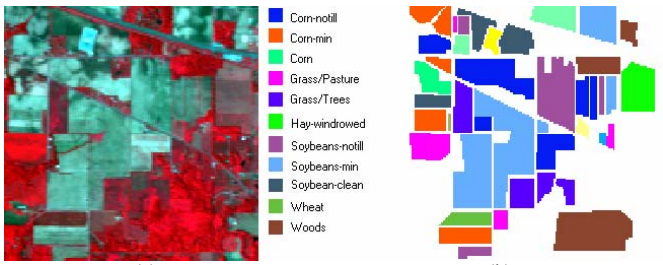

Fig. 3. The AVIRIS data used in the experiment: (a) The data set displayed in simulated color (bands 50, 27, 17 for RGB channels); (b) The ground truth data.

LIBSVM [21] with radial basis function (RBF) kernel was used to solve the binary problem. It is worth noting that the nonlinear SVM is more robust to the parameter settings than linear SVM. This is explained by the fact that a linear separation between classes involves a large number of error samples, which lie on the wrong side of the separating hyperplane [9].

To assess the effectiveness of the proposed method, denoted as FOAO with feature reduction (FOAO-FR), four well-known multiclass SVM methods with the original feature set, including OAO, OAA, BTS and DAGSVM, are used for comparison. Moreover, MLC with the same feature reduction method (MLC-FR) is used for comparison.

Both parameters of SVMs were set according the cross-validation. Parameters C and γ were set to be 32 and 1 for FOAO-FR. They were set to be 32 and 0.0625 for OAO and BTS, while they were set to be 32 and 0.125 for OAA. The parameter δ of BTS was set to be 1.5%. All the experiments were done on Pentium D CPU 2.80 GHZ with 1 GB RAM.

Considering the number of classes is eleven, we adopt ten features for classification in the MLC-FR and FOAO-FR methods. The comparison of test time of different methods is shown in Table I, while the comparison of overall classification accuracy of different methods is shown in Table II.

From Table I, we can see that the proposed method is much faster than OAO, OAA, BTS, and DAGSVM with original feature set. It is worth noting that it can even be a little faster than MLC with the same reduced feature set. The property that FOAO-FR can achieve a much faster classification is due to its greatly reduced feature dimensionality and the selection of binary SVMs with fewest SVs.

When the percentage of training data increases, all the classification process of methods of SVMs slows down. That is because the number of SVs grows as the number of training samples increases and the number of SVs is proportional to the computational complexity of the binary SVM. The test time of MLC-FR is not influenced by the percentage of training data.

From Table II, we can see that the classification accuracy of FOAO-FR is a little lower than OAO, OAA, BTS, and DAGSVM. As reported in previous references [22] [23], all methods of SVMs achieved much better results than MLC.

TABLE I
TEST TIME (IN SECONDS) OF DIFFERENT METHODS

Train	OAO	OAA	BTS	DAGSVM	MLC-FR	FOAO-FR
30%	13.28	13.19	5.47	5.09	1.20	0.75
45%	17.06	17.94	7.64	6.74	1.20	0.86
60%	20.11	21.84	9.02	8.03	1.21	1.00
75%	23.38	25.94	10.64	9.08	1.20	1.11

When the percentage of training data increases, the classification accuracies of the methods of SVMs increase. However, the classification accuracy of MLC-FR does not increase so apparently when the percentage of training data increases. That is mainly because MLC-FR is influenced by the accuracy of estimation of the multivariate normal model.

TABLE II
OVERALL CLASSIFICATION ACCURACY OF DIFFERENT METHODS

Train	OAO	OAA	BTS	DAGSVM	MLC-FR	FOAO-FR
30%	91.76	91.39	91.68	91.52	81.77	89.68
45%	92.82	91.93	92.70	92.74	81.81	91.19
60%	93.47	92.86	93.39	93.31	81.48	91.35
75%	93.84	93.96	93.92	93.84	81.36	91.52

V. CONCLUSIONS

To reduce computational complexity of classification process of SVMs, a fast two-stage method is proposed. First, it adopts a feature reduction algorithm. Then, the FOAO is proposed to further reduce the computational complexity. Experimental results show that the proposed method can achieve a much faster classification than previous SVM methods. Moreover, the proposed method can even be faster than MLC method with the same reduced feature set.

In the future, some research can be done to further reduce the computational complexity. For example, reduced set methods (or SV simplification methods) can be adopted after the training of binary SVMs. However, it is worth noting that the classification accuracy might decrease due to the approximation.

ACKNOWLEDGMENT

We thank Prof. D. Landgrebe of Purdue University, West Lafayette, IN, USA, for providing the AVIRIS data set.

REFERENCES

[1] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci.Remote Sens.*, vol. 44, no. 11, pp. 3374-3385 Nov. 2006.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Ann. Workshop on Computat.Learn. Theory*, 1992, pp. 144–152.

[3] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* vol. 34, no. 1, pp. 60-67, Feb. 2004.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[5] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, pp. 321-336, 2003.

[6] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," in *Proc. Int. Conf. Pattern Recognition*, 1994, pp. 77-87.

[7] J. Friedman. (1996) Another Approach to Polychotomous Classification. Dept. Statist., Stanford Univ., Stanford, CA. [Online]. Available: http://www-stat.stanford.edu/reports/friedman/poly.ps.Z.

[8] U. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods-Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 255-268.

[9] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci.Remote Sens.*, vol. 42, no. 8, pp. 1778-1790, Aug. 2004

[10] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*, S.A.Solla, T.K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 547-553.

[11] B. Fei and J. Liu, "Binary tree of SVM: A new fast multiclass training and classification algorithm," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp.696-704, May 2006.

[12] J. Chen, C. Wang, and R. Wang, "Combining support vector machines with pairwise decision tree," *IEEE Geosci. Remote Sens. Lett.*, to be published.

[13] T. G. Dietterich and G. Bakiri. "Solving multiclass learning problems via error-correcting output codes", *J. Artif. Intell. Res.*, vol. 2, pp. 263-286, 1995.

[14] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers", *J. Mach. Learn. Res*, vol. 1, pp. 113-141, 2000.

[15] O. Pujol, P. Radeva, and J. Vitià, "Discriminant ECOC: A heuristic method for appliction dependent design of error correctiong output codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1001-1006, June 2006.

[16] D. D. Nguyen and T. B. Ho, "A bottom-up method for simplifying support vector solution," *IEEE Trans. Neural netw.*, vol. 17, no.3, pp. 792-796, May 2006.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press), 1990.

[18] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. on Communication Technology*, vol. 15, pp. 52-60, 1967.

[19] C. Lee and E. Choi, "Bayes error evaluation of the Gaussian ML classifier," *IEEE Trans. Geosci.Remote Sens.*, vol. 38, no. 3, pp. 1471-1475, May 2000.

[20] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Trans. Geosci.Remote Sens.*, vol. 37, no. 1, pp. 538-542, Jan. 1999.

[21] LIBSVM: A Library for Support Vector Machines, C. -C. Chang and C. -J. Lin. (2001). [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, Feb. 2002.

[23] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 1007-1011.