Integrating Support Vector Machines in a Hierarchical Output Space Decomposition Framework

Yangchi Chen and Melba M. Crawford
Center for Space Research
3925 W. Braker Lane, Austin, TX 78759
The University of Texas at Austin
Email: yanji@csr.utexas.edu and crawford@csr.utexas.edu

Joydeep Ghosh
Department of Electrical and Computer Engineering
The University of Texas at Austin
Email: ghosh@ece.utexas.edu

***Abstract*— This paper presents a new approach called Hierarchical Support Vector Machines (HSVM), to address multi-class problems. The method solves a series of max-cut problems to hierarchically and recursively partition the set of classes into two-subsets, till pure leaf nodes that have only one class label, are obtained. The SVM is applied at each internal node to construct the discriminant function for a binary meta-class classifier. Because max-cut unsupervised decomposition uses distance measures to investigate the natural class groupings, HSVM has a fast and intuitive SVM training process that requires little tuning and yields both high accuracy levels and good generalization. The HSVM method was applied to Hyperion hyperspectral data collected over the Okavango Delta of Botswana. Classification accuracies and generalization capability are compared to those achieved by the Best Basis Binary Hierarchical Classifier, a Random Forest CART binary decision tree classifier and Binary Hierarchical Support Vector Machines.**

I. Introduction

Achieving both high classification accuracy and good generalization when sample sizes are small relative to the dimension of the input space continues to be a challenging problem, specially when the size of the output space (number of classes) is large. Previous studies of supervised methods show that a complex classifier tends to overtrain in such situations while a weak classifier is often inadequate for large output space problems [1].

According to Occam's razor, as classifiers become more and more complex, the generalization error will eventually increase because of over-training [2]. Ensemble methods can alleviate this problem, particularly by reducing the model variance [3]. In particular, random forest classification, a combination of bagging and random subspace method, achieves both high classification accuracies and good generalization but is computationally costly due to the large (50-100) number of classifiers required in the ensemble [4].

A new group of classifiers called Support Vector Machines (SVM) seek to maximize margin between training samples and the decision boundary [5]. Typically implemented as binary classifiers, SVMs utilize nonlinear optimization and kernel projection to find the optimal distance between two classes in a new projection. Because they search for the best hyperplane instead of the highest training accuracy, they tend not to overtrain on a sample data set.

Although SVMs were originally designed for binary classification, several class decomposition approaches, including pairwise, one-vs-all, and error correcting output codes (ECOC) have been investigated for extending the SVM approach to handle multi-class problems [6]. Even though one-vs-all and ECOC decomposition methods can achieve high quality classification results, using the associated class groups often requires a complex SVM kernel, such as the RBF kernel, to construct the decision boundary. This results in a time consuming, tedious parameter tuning process. To mitigate this problem, a new class decomposition framework is investigated here. The main idea is to group the classes into two meta-classes based on the natural affinities among the classes so that the binary problem of separating these two meta-classes is relatively easy. By recursively applying this approach to the two subgroups, a binary hierarchical output space decomposition is achieved. This approach was used in the Generalized Associative Modular Learning System (GAMLS) [7], a simulated annealing-based class decomposition algorithm utilized by the Binary Hierarchical Classifier (BHC). Previous studies have demonstrated that this framework has several advantages for classification of remotely sensed data with large output spaces: including 1) The order of the number of binary classification problems reduces from $O(C^2)$ to $O(C)$; 2) the impact of the small sample problem is mitigated; 3) the framework provides a natural, intuitive structure [8]. When an SVM is used to solve the binary classification problem at each internal node of the BHC, classification accuracies increased somewhat and generalization improved [9], [10]. However, the two ingredients of obtaining the hierarchical class decomposition and using SVMs as binary classifiers, were not integrated in a common "group distance" framework, and tuning, which is critical to good performance of the SVM, was time consuming.

The new method proposed in this paper provides an alternative to GAMLS for obtaining the hierarchical class decomposition, and is based on a max-cut formulation to search the

0-7803-8742-2/04/$20.00 (C) 2004 IEEE

maximum total distance between two (meta)-class partitions.

II. Methodology

The Hierarchical Support Vector Machines (HSVM) method is based on a max-cut hierarchical output space decomposition algorithm and uses SVM as the based classifier at each internal node to construct the decision boundary. SVM and the max-cut problem, two main algorithms of the HSVM framework, are well matched. The background for each method is first presented, then their integration into the HSVM method is presented in the reminder of this section.

A. Support Vector Machines

The Support Vector Machine projects the input vectors into a high dimensional feature space, then searches for the linear decision boundary that maximizes the minimum distance between two class groups [11]. For a binary classification problem with input space $\mathbf{X}$ and binary class labels $Y : Y \in \{-1, 1\}$.
Giving training samples

$$(y_1, \mathbf{x}_1), ..., (y_l, \mathbf{x}_l), \quad y_i \in \{-1, 1\} \tag{1}$$

the goal of SVM is to search the optimal hyperplane

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2}$$

with variables $\mathbf{w}$ and b that satisfy the following inequality.

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, ..., l. \tag{3}$$

Defining the minimum distance between two class groups in the new projection.

$$\rho(\mathbf{w}, b) = \min_{\{\mathbf{x}: y=1\}} \frac{\mathbf{x} \cdot \mathbf{w}}{|\mathbf{w}|} - \max_{\{\mathbf{x}: y=-1\}} \frac{\mathbf{x} \cdot \mathbf{w}}{|\mathbf{w}|} \tag{4}$$

From e.q. (3), $\min_{\{\mathbf{x}: y=1\}} \mathbf{x} \cdot \mathbf{w} = 1$ and $\max_{\{\mathbf{x}: y=-1\}} \mathbf{x} \cdot \mathbf{w} = -1$. Substituting back into e.q. (4), yields

$$\rho(\mathbf{w}_0, b_0) = \frac{2}{|\mathbf{w}_0|} = \frac{2}{\sqrt{\mathbf{w}_0 \cdot \mathbf{w}_0}}$$

For a given training set $\mathbf{w}, b$ that maximizes $\rho(\mathbf{w}_0, b_0)$ solves the following quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w} \cdot \mathbf{w} \\ s.t. \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, ..., l. \end{aligned} \tag{5}$$

If the given training sample set is linear separable, the optimization problem (5) has feasible solutions. The optimal solution $\mathbf{w}$, and b forms the best hyperplane that maximizes the margin between two different classes in the new projection. Because SVM search for the best separation hyperplane instead of the highest training sample accuracy, they never over-train on a sample data set. If the parameters are properly selected, SVM typically produce both excellent classification results and good generalization if parameters are properly selected. Not every problem is guaranteed to be linear separable, so a soft margin hyperplane SVM was developed to separate the training set with a minimal number of errors [5]. The associated optimization problem introduces some non-negative variables ξ_i and becomes

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + CF\left(\sum_{i=1}^{l} \xi_i\right) \\ s.t. \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, ..., l. \end{aligned} \tag{6}$$

where $F(u)$ is a monotonic convex function, and C is a user-defined penalty constant variable, optimization problem 6 allows class samples to move beyond the decision boundary, while incurring a penalty cost $CF(u)$. It has been show that when a training sample is small, it is important to select an appropriate C to mitigates the effect of outliers of the training sample set.

The primary weaknesses of SVM are that they only solve binary classification problems and are computationally intensive due to the training process. An output space decomposition algorithm is required to extend SVM to solving multi-class problems. Later experiments show that a natural decomposition framework can speed up the SVM training process and reduces time spending on parameters tuning. GAMLS [9] provides one such approach, but could not be naturally integrated with the SVM in the hierarchy building process.

B. Max-Cut Problem

In a max-cut problem, an undirected graph with nonnegative edge weights is partitioned into two groups. The cuts between these two groups have the maximum weight [12]. The max-cut problem is an NP-hard nonlinear integer programming problem. Define an undirected graph $G = (N, E)$ where N represents nodes, and E represents edges of the graph. $w_{ij} \geq 0$ represents the weight of an edge linking nodes i and j. The objective is to find the best binary partition that has the cut $\delta(K^*)$ that $K^* \subseteq N$ and $\{ij \in E : \ i \in K^*, j \notin K^*\}$ that has the maximum weight:

$$w(\delta(K^*)) = \sum_{ij \in \delta(K^*)} w_{ij}. \tag{7}$$

The graph is assumed to be complete by setting $w_{ij} = 0$ for all non-edges ij.

The max-cut problem can be represented using an integer quadratic programming formulation with decision variables X : $x_i \in \{1, -1\}$. $\forall i \in N$. To represent a cut $\delta(K)$; $x_i = 1$ iff $i \in K$. If $ij \in \delta(K)$, $x_i x_j = -1$. Thus:

$$w(\delta(K)) = \frac{1}{2} \sum_{i<j} w_{ij} (1 - x_i x_j) \tag{8}$$

and the resulting max-cut integer quadratic problem is:

$$\begin{aligned} \max \quad & w(\delta(K)) \\ s.t. \quad & x_i \in \{+1, -1\}, i \in N \end{aligned} \tag{9}$$

Because this integer quadratic problem is NP-hard, the combination of the feasible solutions grows exponentially as the number of N increases. Goemans and Williamson proposed that the original max-cut problem can be relaxed into

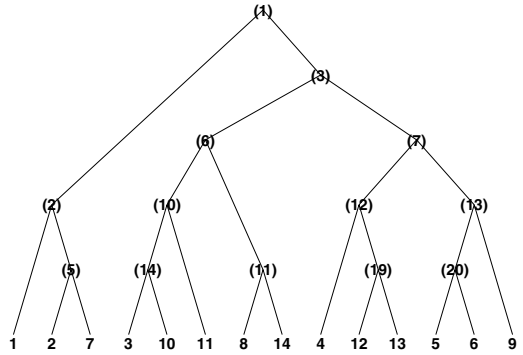

Fig. 1. Typical HSVM hierarchical structure

a constrained quadratic problem and be solved using a semi-definite programming [13]. The interior point method extended by Nesterov and Nemirovskii [14] provides a computationally efficient method for solving the semi-definite problem. The relaxed max-cut problem solved using SDP achieves optimal or near optimal results and has an expected value of 87.7% of the optimal max-cut [13].

C. Hierarchical Support Vector Machines

GAMLS treats each meta-class partition as a multivariate normal distribution with group mean vector μ and covariance matrix Σ, and seeks the best binary partitioning in terms of the maximum separability obtained through a Fisher projection. Thus the statistical distance used is $d_f = (\mu_1 - \mu_2) \cdot \Sigma_{pooled}^{-1}$ in the Fisher projection. Previous studies showed that if SVMs are used instead for the binary classification problems, classification accuracies and generalization improved but the training process slows down dramatically [9], [10]. In order to speed up the process by investigating the impact of natural class grouping in combination with the SVM base classifier, the proposed max-cut hierarchical output space decomposition method searches the maximum total distance between the two class partitions instead of a single projected distance d_f. The original class samples are treated as an undirected graph G where node n_i represents class i and the non-negative weight:

$$w_{ij} = \frac{1}{2} \sum_{\forall x} \left(f_i(x) \log \frac{f_i(x)}{f_j(x)} + f_j(x) \log \frac{f_j(x)}{f_i(x)} \right) \qquad (10)$$

is the average Kullback-Leibler distance [15] between the density function of class i and class j. The new HSVM approach solves this max-cut problem to achieve the required unsupervised class decomposition at each node of the binary hierarchical structure. The original output space is hierarchically decomposed into pure leaf nodes that have only one class label at each node (see Fig. 1). Since this max-cut unsupervised decomposition uses total pairwise distance measures to investigate the natural class grouping, the hierarchical structure results in a fast and intuitive SVM training process that requires little tuning and yields both high accuracy levels and good generalization.

TABLE I
BOTSWANA TEST DATA: ACCURACY (STD. DEV.)

Training %	BB-BHC	RF-CART	BH-SVM	HSVM
15%	89.9(1.36)	86.6(1.34)	92.3(1.15)	90.7(2.49)
30%	91.8(1.75)	89.5(1.29)	93.8(2.23)	93.2(1.07)
50%	92.9(0.73)	91.1(1.32)	96.2(0.75)	94.1(0.97)
75%	94.0(0.69)	92.0(1.34)	96.6(0.95)	95.1(0.63)

The new algorithm is applied to Hyperion hyperspectral data collected over the Okavango Delta of Botswana. Classification accuracies and generalization capability are compared to those achieved by the Best Basis Hierarchical Classifier [8], the Random Forest CART binary decision tree classifier [16] and Binary Hierarchical Support Vector Machines (BH-SVM) [9].

III. RESULTS

The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001-2003. The Hyperion sensor on EO-1 acquires data at $30m^2$ pixel resolution over a 7.7 km strip in 242 bands covering the 400-2500 nm portion of the spectrum in 10 nm windows. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220]. The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta.

Ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, these training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data. Because the training and test data are spatially collocated, an extended test set was also acquired and used to evaluate the generalization of these classifiers to another area. Note that this extended data may have substantially different characteristics as it is taken from a geographically separate location. Its purpose here is to investigate the capability of the various methods for extending results obtained from one area to other areas that are not so spatially correlated. Hereafter, these data are referred to as the test and extended test data, respectively.

Experiments were performed using Best Basis BHC (BB-BHC), Random Forest CART (RF-CART), binary hierarchical SVM (BH-SVM), and the proposed HSVM. Average classification accuracies for test data for the 10 experiments conducted with each classifier are list in Table I. The overall trend shows that classification accuracies increase as the percentage of training sample increases for all four classifiers. BB-BHC, BH-SVM and HSVM all perform well at 15% sampling

TABLE II
BOTSWANA EXTENDED TEST DATA: ACCURACY (STD. DEV.)

Training %	BB-BHC	RF-CART	BH-SVM	HSVM
15%	66.1(3.07)	67.2(1.73)	70.4(2.49)	69.3(5.06)
30%	63.8(1.87)	68.6(1.73)	70.9(1.88)	70.4(2.17)
50%	63.4(2.33)	70.2(0.84)	72.1(2.06)	70.8(1.32)
75%	63.7(1.33)	70.4(1.26)	71.6(1.41)	70.3(0.97)

rate, which means they mitigate the impact of small sample problem. The HSVM method produces quality classification accuracies among these classifiers on test data, while BH-SVM did slightly better. Classification accuracies on extended test set are presented in Table II. The results show that while the BB-BHC performs well on the test samples, it does not generalize so well on new area. Because RF-CART uses a random forest ensemble method to increase the diversity of classifiers, it performs well on the extended test set. HSVM produced either the best or the same classification results as BH-SVM on the extended test set.

As stated in section II-C, since HSVM uses distance measures to exploit the natural class groupings, the hierarchical structure results in a fast and intuitive SVM training process that requires little tuning. Unlike ECOC decomposition [10] or our previous hierarchical decomposition attempt [9], that require the RBF kernel and tedious tuning to separate unnatural grouping, HSVM uses a linear kernel to search for the best linear decision boundary in each internal node of the hierarchical structure. For a Botswana experiment that has 1619 samples, 14 classes and 145 feature spaces, using a 3GHz Pentium 4 CPU, HSVM finished training and testing in 20 seconds, while BH-SVM took 4800 seconds [9]. Previous study shows that for the same experiment, BB-BHC required 115 seconds and RF-CART took 480 seconds[4].

IV. CONCLUSION

A new Hierarchical Support Vector Machines (HSVM) approach that utilizes a tree structure framework and solves a series of max-cut problems to perform the unsupervised class decomposition has been developed. SVM classifier is applied at each internal node to construct the best discriminant function of a binary meta-class problem.

In this paper, HSVM was evaluated using a series of experiments. HSVM consistently provided good classification results on both test and extended test samples in experiments conducted using 4 different sampling rates and 10 different random samples for each sampling rate. The new HSVM achieves both high classification accuracy and good generalization when sample sizes are small relative to the dimension of the input space and the output space is large.

The HSVM uses distance measures to investigate the natural class grouping and results in an efficient classifier that requires little tuning. The method extends original binary SVM classifier to a fast and multi-classes classifier. The HSVM framework also provides a natural, and intuitive structure. Further study can utilize this hierarchical structure to evaluate possible stopping criteria for mixed-class samples and knowledge transfer problem that applies a classification model to a new area that has a few or no training samples available.

ACKNOWLEDGMENT

This research was supported by the NASA EO-1 Program (Grant NCC5-463), the Terrestrial Sciences Program of the Army Research Office (DAAG55-98-1-0287) and NSF (Grant IIS-0312471). We thank Amy Neuenschwander of the UT Center for Space Research for help in pre-processing the Hyperion data and interpreting the overall classification results. We also thank Jisoo Ham of the UT Center for Space Research for working jointly and help in writing Matlab script for HSVM.

REFERENCES

[1] B. E. Boser, I. Guyon, and V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Computational Learning Theory*, 1992, pp. 144–152.

[2] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," in *Proc. 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 322–330.

[3] M. Skurichina and R. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers," *Pattern Analysis and Applications*, vol. 5, pp. 121–135, 2002.

[4] M. M. Crawford, J. Ham, Y. Chen, and J. Ghosh, "Random Forests of Binary Hierarchical Classifiers for Analysis of Hyperspectral Data," in *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on*, 2003, pp. 337– 345.

[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] C.-W. Hsu and C.-J. Lin, "A comparison on methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415–425, 2002.

[7] S. Kumar and J. Ghosh, "GAMLS: A generalized framework for associative modular learning systems," 1999.

[8] S. Kumar, J. Ghosh, and M. Crawford, "Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis," *International J. Pattern Analysis and Applications*, vol. 5, no. 2, pp. 210–220, 2002.

[9] J. A. Dare, "Support Vector Machines in a Binary Hierarchical Classifier," Master's thesis, University of Texas at Austin, 2004.

[10] S. Rajan and J. Ghosh, "An Empirical Comparison of Hierarchical vs. Two-level approaches to Multiclass Problems," in *Multiple Classifier Systems*, F. Roli, J. Kittler, and T. Windeatt, Eds. LNCS Vol. 3077, Springer, 2004, pp. 283–292.

[11] C. Huang, L. Davis, and J. Townshend, "An Assessment of Support Vector Machine for Land Cover Classification," *Int. J. Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[12] M. J. Todd, "Semidefinite Optimization," Cornell University, Ithaca, NY 14853, Tech. Rep., 2001.

[13] M. X. Goemans and D. P. Williamson, "Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming," *J. Assoc. Comput. Mach.*, vol. 42, pp. 1115–1145, 1995.

[14] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*. Philadelphia: Society for Industrial and Applied Mathematics, 1994.

[15] S. Kullback, *Information Theory and Statistics*. New York: John Wiley and Sons., 1959.

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.