Available online at www.sciencedirect.com
ScienceDirect
Advances in Space Research 41 (2008) 1793–1799
ADVANCES IN SPACE RESEARCH (a COSPAR publication)
www.elsevier.com/locate/asr

Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem ☆

Mingmin Chi [a,*], Rui Feng [a], Lorenzo Bruzzone [b]

[a] Department of Computer Science and Engineering, Fudan University, 220 Han Dan Road, Shanghai 200433, China
[b] Department of Information and Communication Technologies, University of Trento, Italy

Received 1 November 2006; received in revised form 3 February 2008; accepted 6 February 2008

Abstract

With recent technological advances in remote sensing, very high-dimensional (hyperspectral) data are available for a better discrimination among different complex land-cover classes having similar spectral signatures. However, this large number of bands makes very complex the task of automatic data analysis. In the real application, it is difficult and expensive for the expert to acquire enough training samples to learn a classifier. This results in a classification problem with small-size training sample set. Recently, a regularization-based algorithm is usually proposed to handle such problem, such as Support Vector Machine (SVM), which usually are implemented in the dual form with Lagrange theory. However, it can be solved directly in primal formulation. In this paper, we introduces an alternative implementation technique for SVM to address the classification problem with small-size training sample set. It has been empirically proven that the effectiveness of the introduced implementation technique which has been evaluated by benchmark datasets.

Keywords: Primal Support Vector Machine (SVM); Classification; Small-size training dataset problem; Hyperspectral remote-sensing data

1. Introduction

One of the most critical problems relating to the supervised classification of remote-sensing images lies in the definition of a proper size of training set for an accurate learning of classifiers. Since the collection of ground-reference data is an expensive and complex task, in many cases the number of training samples is insufficient for a proper learning of classification systems. This issue is particularly critical when hyperspectral images are considered. Such hyperspectral data are generally made of about 100–200 spectral channels of relatively narrow bandwidths (5–10 nm). Although high-dimensional features are capable of better discriminating among the complex (sub)classes, in the real application, it is difficult and expensive for experts to acquire enough training samples to learn a classifier. Consequently, it is impossible to meet the requirements on the necessary number of training samples since the size of training dataset is relatively fixed.

When the number of (representative) training samples is relatively small with respect to the number of features (and thus of classifier parameters to be estimated), the well-known problem of the *curse of dimensionality* (i.e., the *Hughes phenomenon* Hughes, 1968)[1] occurs. This results in the risk of overfitting of the training data and can lead to poor generalization capabilities of the classifier. Conventional classification methods, such as the Gaussian Maximum Likelihood algorithm, cannot be applied to hyperspectral data due to the high dimensionality of the

☆ Expanded version of a talk presented at COSPAR on terrestrial phenomena and land products from space: validation, application and perspectives (Beijing, China, July 2006).

* Corresponding author. Tel.: +86 21 5566228.
E-mail addresses: mmchi@fudan.edu.cn (M. Chi), fengrui@fudan.edu.cn (R. Feng), lorenzo.bruzzone@ing.unitn.it (L. Bruzzone).

[1] With more discriminative features, classification performance is improved with the increase of the number of labeled samples; if the number of labeled samples is fixed, the performance otherwise decreases.

0273-1177/$34.00 © 2008 COSPAR. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.asr.2008.02.012

data and the relatively small number of available training samples. According to the Gaussian modeling of the statistical distributions of classes, Maximum Likelihood classifiers are widely used for the optimization, where the mean vector and the covariance matrices of classes are estimated in terms of training samples. However, the small ratio between the number of training samples and the number of classifier parameters often results in unstable covariance matrices (which, in some cases, can be singular). This strongly affects the classification accuracy.

By supervised algorithms for classification task, Hoffbeck and Landgrebe (1996) proposed to use regularized covariance matrices by the leave-one-out covariance (LOOC) estimate (Hoffbeck and Landgrebe, 1996). For obtaining a classifier with improved generalization capabilities, in (Tadjudin and Landgrebe, 1999) an LOOC-based regularized estimator was presented in the Bayesian framework. This estimator reduces the number of parameters to be computed, thus reducing the variances of their estimates. The difficulty in using classification methods based upon conventional multivariate statistical approaches is that many of these methods rely on having a non-singular class-specific covariance matrix for all classes (Benediktsson et al., 1995). When working with high-dimensional data sets, it is likely that the covariance matrices will be singular when using a limited (with respect to the number of input bands) amount of training samples. Accordingly, those approaches can result in the overfitting of training data and so lead to a poor generalization.

In order to address the small-size training set problem, in the machine learning community, classification methods in the regularization framework are usually used in high-dimensional input space. One of the most popular methods is Support Vector Machine (SVM) (Vapnik, 1998), a large margin based classifier with a good generalization capacity in the small-size training set problem with high-dimensional input space. Recently, SVMs have been successfully applied in the classification of hyperspectral remote-sensing data (Gualtieri and Cromp, 1998; Melgani and Bruzzone, 2004). However, all of the literatures are focusing on the pursuit of the solution with dual property, i.e., Lagrange theory is applied for the optimization problems (Gualtieri and Cromp, 1998; Melgani and Bruzzone, 2004; Burges, 1998; Critianini and Shawe-taylor, 2000). Nonetheless, SVMs can also be optimized directly on the primal representation (Mangasarian, 2002; Keerthi and DeCoste, 2005; Chapelle, 2007). This is the focus of the paper to propose the usage of the primal SVM for the classification of hyperspectral remote-sensing data with small-size training set. As we know, it is the first time to use the primal SVM for the hyperspectral image classification. In particular, the L_2-norm regularizer and the quadratic loss are taken into account for the objective function, and conjugate descent algorithm is applied on the objective for the optimization problem. It is worth noting that the implementation on such objective is an unconstrained optimization problem. The experimental analysis was carried out on Hyperion hyperspectral remote-sensing data acquired by the NASA EO-1 satellite over the Okavango Delta, Botswana in 2001. The results provided by the proposed implementation technique for the primal SVM were compared with those provided by the state-of-the-art approaches reported in (Chen et al., 2004). On the basis of this comparison, the proposed implementation technique provided comparable accuracy as those by the reference methods.

This paper is organized in four sections. In Section 2, the introduced classification technique, Support Vector Machines (SVM) in the primal representation is presented both in the linear and non-linear cases. Section 3 describes the data set used in the experiments and reports the results obtained by the presented classification technique. Finally, discussion and the conclusion of this work is given in Section 4.

2. Primal Support Vector Machine (SVM)

2.1. Problem formulation

Let us consider a binary classification problem. For the generalization to the multiclass case the reader can refer to (Melgani and Bruzzone, 2004; Hsu and Lin, 2002). Let the given training dataset $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n, \mathbf{X} \in \mathbb{R}^{d\times n}$ be made of n labeled samples in a d-dimensional feature space and the associated labels $\mathbf{y} = (y_i)_{i=1}^n$, $y_i = \{\pm 1\}$.

The notation adopted in the paper is as follows: bold faced variables (e.g., $\mathbf{x}$,$\mathbf{w}$) are used to represent column vectors. Matrices are represented by calligraphic upper-case alphabets (e.g., $\mathbf{K}$). Random variables are represented by low-case alphabets (e.g., y). The symbols $\mathcal{H}$, $\mathbb{R}^d$ denote the Hilbert space, and the d-dimensional vector space, respectively. The symbol $^\top$ denotes the transpose of a vector, $\|\cdot\|^2$ denotes the L_2 norm, and "s.t." represents "subject to".

2.2. The linear case

The standard SVM is a linear inductive learning classifier where data in input space are separated by the hyperplane:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{1}$$

with maximal geometric margin $2/\|\mathbf{w}\|^2$, where $\mathbf{w}$ is a vector, normal to the hyperplane and $|b|/\|\mathbf{w}\|^2$ is the perpendicular distance from the hyperplane to the origin. The objective of the learning phase of standard SVM is to maximize the geometrical margins between classes in the feature space. This is equivalent to minimizing the following objective function:

$$\min_{\mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 \right\} \quad \text{s.t. } \forall_{i=1}^n : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geqslant 1. \tag{2}$$

If the perturbation of noises is considered, (2) becomes as follows:

$$\min_{\mathbf{w},\xi_i}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \quad \text{s.t. } \forall_{i=1}^{n} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geqslant 1 - \xi_i, \xi_i > 0 \tag{3}$$

where ξ_i is a slack variable for the training pattern $\mathbf{x}_i$ and C is the penalty parameter of the loss (that plays the role of tuning the regularization of the problem).

2.2.1. Primal representation

An alternative to the dual SVM (Burges, 1998; Critianini and Shawe-taylor, 2000; Melgani and Bruzzone, 2004) is to use an optimization technique on the original representation directly, such as Newton methods (Mangasarian, 2002; Keerthi and DeCoste, 2005; Chapelle, 2007; Boyd and Vandenberghe, 2002). In this case, the objective function (3) is rewritten without explicit constraints as follows:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} H(y_i f(\mathbf{x}_i)) \tag{4}$$

where $H(y_i f(\mathbf{x}_i))$ is the loss for the training patterns $\mathbf{x}_i \in \mathbf{X}_l$, defined by $H(t) = \max(0, 1-t)^p$ when $p = 1$, a hinge loss is used (cf. Fig. 1a) and if $p = 2$, it is a quadratic loss (Mangasarian, 2002; Chapelle, 2007) (cf. Fig. 1b). It is worthy pointing out that (4) is a strongly convex optimization problem with the quadratic loss and so a global solution can be guaranteed (Mangasarian, 2002). As default, we only consider the quadratic loss for labeled samples for the ease of computation in this paper. With the implementation of optimization on (4), we define a labeled sample $\mathbf{x}_i$ given the vector $\mathbf{w}$, as a support vector if $y_i f(\mathbf{x}_i) < 1$, i.e., the loss on this sample is not equal to zero (Chapelle, 2007). For the simplicity, we ignore the offset b in the following discussion since all the algebra presented below can be extended easily to take it into account. For details, reader can be referred to (Chapelle, 2007, Appendix 2.B).

If the gradient descent is used, provided that $H(\cdot)$ is differentiable, the gradient of (4) with respect to $\mathbf{w}$ is given by:

$$\nabla = \mathbf{w} + C\sum_{i=1}^{n}\frac{\partial H(y_i f(\mathbf{x}_i))}{\partial \mathbf{w}}\mathbf{x}_i y_i \tag{5}$$

where $\frac{\partial H(y_i f(\mathbf{x}_i))}{\partial \mathbf{w}}$ is the partial derivative of $H(y_i f(\mathbf{x}_i))$ with respect to $\mathbf{w}$. At the optimal solution $\mathbf{w}^*$, the gradient vanishes such that $\nabla_{w^*} = 0$. Hence, we have the solution

$$\mathbf{w} = \sum_{i=1}^{n}\beta_i \mathbf{x}_i, \quad \beta_i = -C\frac{\partial H(y_i f(\mathbf{x}_i))}{\partial \mathbf{w}} y_i \tag{6}$$

This implies that the solution is the linear combination of input data. This result is also known as *Representer Theorem* (Kimeldorf and Wahba, 1971). Then, we can replace $\mathbf{w}$ in (4) with (6) as follows:

$$\frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\mathbf{x}_i^\top\mathbf{x}_j + C\sum_{i=1}^{n}H\left(y_i\sum_{j=1}^{n}\beta_j\mathbf{x}_i^\top\mathbf{x}_j\right). \tag{7}$$

It is an unconstrained optimization problem, so we can use any optimization technique (e.g., Newton methods (Keerthi and DeCoste, 2005; Mangasarian, 2002)) to solve (7) with respect to $\boldsymbol{\beta}$.

Once the optimal $\boldsymbol{\beta}^*$ is obtained, we can easily compute the predicted value for given input $\mathbf{x}$ by:

$$f(\mathbf{x}) = \sum_{i=1}^{n}\beta_i^*\mathbf{x}_i^\top\mathbf{x}. \tag{8}$$

Thus, the corresponding labeling is:

$$y = \operatorname{sgn}[f(\mathbf{x})] = \begin{cases} +1 & \text{if } f(\mathbf{x}) \geqslant 0 \\ -1 & \text{otherwise.} \end{cases} \tag{9}$$

2.3. The non-linear case

In practical applications, data usually cannot be linearly separated in the input space. However, due to the "*kernel trick*", a linear support vector machine can still be found but in a higher or infinite dimensional space, where the data are mapped to, for example, *Hilbert space* $\mathcal{H}$. We call it *feature space* through a map ϕ:

$$\begin{aligned} &\phi : \mathcal{X} \to \mathcal{H} \\ &\mathbf{x} \to \phi(\mathbf{x}). \end{aligned} \tag{10}$$

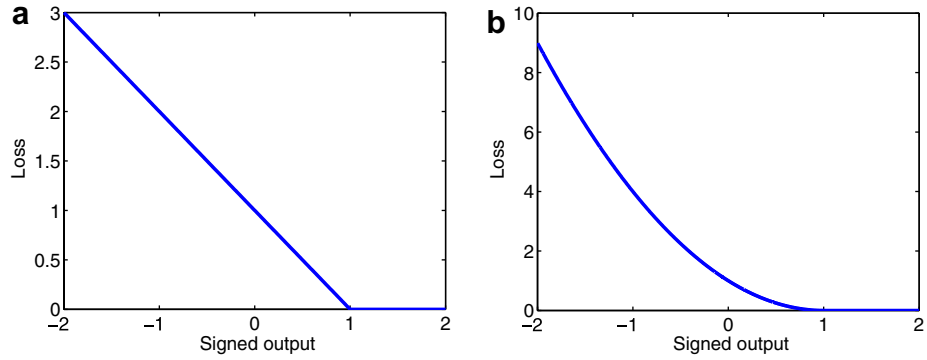

Fig. 1. Loss for the labeled samples in (4), when (a) $p = 1$, a hinge loss $H(t) := \max(1 - t)$ and (b) $p = 2$, a quadratic loss $H(t) := \max(1 - t)^2$.

Then, a linear decision boundary can be constructed in that space.

After mapping, the dot product pair, e.g., $(\mathbf{x}^\top \mathbf{x}')$ in input space $\mathcal{X}$ is represented in the *dot product space* or Hilbert space $\mathcal{H}$ as $(\phi(\mathbf{x})^\top \phi(\mathbf{x}'))$. In order to compute such a form of the dot product, we can use kernel representation

$$k(\mathbf{x}, \mathbf{x}') := \phi(\mathbf{x})^\top \phi(\mathbf{x}') \tag{11}$$

which allows us to compute the value of the dot product in $\mathcal{H}$ without having to explicitly compute the map ϕ. Due to Mercer's Theorem, there exists such a map $\phi(\cdot)$ and so (11) always holds true (Schölkopf and Smola, 2002, pp.36–38).

"Kernel trick" plays a central role on non-linear SVM also in primal formulation. We can deal with the non-linear SVMs in the primal by replacing the inner product $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ with a kernel function, i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$. In the meanwhile, with the Representer Theorem (Kimeldorf and Wahba, 1971) and the reproducing property of kernels (Aronszajn, 1950), we can deal with the non-linear primal SVM problem in the Hilbert space $\mathcal{H}$.

In terms of (4), by the *Representer Theorem* (Kimeldorf and Wahba, 1971) we have:

$$\mathbf{w} = \sum_{i=1}^{n} \beta_i k(\mathbf{x}_i, \cdot), \quad \phi(\mathbf{x}_i) = k(\mathbf{x}_i, \cdot). \tag{12}$$

It is easy to see that the solution is a linear combination over the training samples. The task is reduced to find the optimal $\boldsymbol{\beta}^*$. Combining (4) and (12), we can obtain the following function with the reproducing property $f(\mathbf{x}_i) = \langle f, k(\mathbf{x}_i, \cdot)\rangle_{\mathcal{H}}$ (Schölkopf and Smola, 2002; Aronszajn, 1950) ($\langle \cdot, \cdot \rangle$ denotes the inner product):

$$\begin{aligned} &\frac{1}{2}\sum_{i,j=1}^{n} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) + C\sum_{i=1}^{n} H\left(y_i \sum_{j=1}^{n} \beta_j k(\mathbf{x}_i, \mathbf{x}_j)\right) \\ &\quad = \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} + C\sum_{i=1}^{n} H\left(y_i \mathbf{K}_i^\top \boldsymbol{\beta}\right). \end{aligned} \tag{13}$$

In this case, we solve the optimization of the non-linear primal SVM in terms of (13) with respect to $\boldsymbol{\beta}$. The kernel matrix is defined as $\mathbf{K} := [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$, a symmetric positive definite kernel matrix (which can be viewed as a non-linear similarity measure) and $\mathbf{K}_i = [k(\mathbf{x}_i, \mathbf{x}_j)]_{j=1}^{n} \in \mathbb{R}^{n\times 1}$ is the ith column of $\mathbf{K}$.

If $H(\cdot)$ is differentiable, the optimum value $\boldsymbol{\beta}^*$ can be obtained by gradient descent or Newton method (Chapelle, 2007) with respect to $\boldsymbol{\beta}$ in (13). Finally, we can predict the test sample $\mathbf{x}$ by:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i^* k(\mathbf{x}_i, \mathbf{x}). \tag{14}$$

3. Experiments

3.1. Dataset description

To evaluate the performance of the proposed implementation of the primal SVM, experiments were performed using Binary Hierarchical SVM (BH-SVM) (Dare, 2004), Hierarchical SVM (HSVM) (Chen et al., 2004) and Random Forest Classification and Regression Tree (RF-CART) (Breiman, 2001) on the Hyperion hyperspectral remote-sensing data acquired on Okavango Delta, Botswana. The soft code for the primal SVM with Newton method is available in the website.[2] For the multi-class problem, a one-vs-rest combination strategy (Hsu and Lin, 2002; Melgani and Bruzzone, 2004) is adopted in the primal SVM.

A sequence of data were acquired by the NASA EO-1 satellite over the Okavango Delta, Botswana in 2001. The Hyperion sensor on EO-1 acquired the data at 30 m 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 University of Texas 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, were acquired in May 31, 2001 and consist of 14 identified classes representing the land cover types in seasonal swamps, occasional swamps and drier woodlands located in the distal portion of the Delta.[3] These classes were chosen to reflect the impact of flooding on vegetation in the study area. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table 1. Training data were selected manually using a combination of GPS located vegetation surveys, aerial photography from the Aquarap (2000) project, and 2.6 m resolution IKONOS multispectral imagery.

In the experiments, two kinds of datasets are taken into account in terms of the location distribution of training and test sets (cf. Table 1). Traditionally, the training and test data are spatially co-located and can thus be assumed to be samples from the same distribution. We call such datasets as "Spatially Correlated" (SC) dataset. In this case, we have the totally 3248 samples available. For the dataset, 10 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 labeled data on classifier performance, these training data were then sub-sampled to obtain 10 splits comprised of 50%, 30%, 15%

[2] http://www.kyb.tuebingen.mpg.de/bs/people/chapelle/primal/primal_svm.m.

[3] Available at http://www.csr.utexas.edu/hyperspectral/codes.html.

Table 1
The original training dataset, Spatially Uncorrelated (SU) test dataset and Spatial Correlated (SC) test dataset distribution for Hyperion data of Botswana

Class	Name	Training set	SU test set[a]	SC test set[b]
1	Water	270 (8.31%)	126 (5.05%)	68 (8.31%)
2	Hippo grass	101 (3.09%)	162 (6.5%)	26 (3.18%)
3	Floodplain grasses1	251 (7.74%)	158 (6.34%)	63 (7.7%)
4	Floodplain grasses2	215 (6.63%)	165 (6.62%)	54 (6.6%)
5	Reeds1	269 (8.27%)	168 (6.74%)	68 (8.31%)
6	Riparian	269 (8.27%)	211 (8.46%)	68 (8.31%)
7	Firescar2	259 (7.98%)	176 (7.06%)	65 (7.95%)
8	Island interior	203 (6.26%)	154 (6.17%)	51 (6.23%)
9	Acacia woodlands	314 (9.67%)	151 (6.05%)	79 (9.66%)
10	Acacia shrublands	248 (7.65%)	190 (7.62%)	62 (7.58%)
11	Acacia grasslands	305 (9.38%)	358 (14.35%)	77 (9.41%)
12	Short mopane	181 (5.56%)	153 (6.13%)	46 (5.62%)
13	Mixed mopane	268 (8.27%)	233 (9.34%)	67 (8.19%)
14	Exposed soils	95 (2.92%)	89 (3.57%)	24 (2.93%)

[a] This spatially uncorrelated test set was not included in the training set.
[b] This test set is generated from the spatially correlated labeled set. It covers 25% samples of labeled set.

and 5% of the original labeled data. All classifiers were evaluated using the 10 test sets containing 25% of the original labeled samples (i.e., 818 samples).

In practice, however, it is also useful to estimate how a classifier will perform in areas that are somewhat different, in order to indicate how much additional data labeling and retraining is needed to make the model applicable to much larger areas. In the experiments, a "Spatially Uncorrelated" (SU) test set was also acquired from a geographically separate location at the Botswana site and used to evaluate the classifiers mentioned above. In this case, the same training datasets, i.e., made up of 75%, 50%, 30%, 15% and 5% of the dataset consisting of 3248 labeled data, are used for the learning classifiers and the test set was acquired from the different locations and it composes of 2494 test samples.

Note that before model selection, all the data should be preprocessed with a normalization mechanism in the input space $\mathbf{X}$. It has been shown that normalization is a preprocessing type which plays an important role in support vector machine classifiers. In the following experiments, input features are normalized in a range $[-1\ 1]$.

3.2. Model selection

For the sake of computation complexity, the leave-one-out validation would not be considered in the paper for solving model selection problems. Since the size of labeled samples is limited, the hold-out validation is not reliable in this problem. To compromise both, in this paper, cross validation is utilized to select a proper model in our work. In greater detail, a small-size labeled set is divided to k disjoint folds (in all the experiments, 5-fold cross validation was used). Then, one of k folds is randomly selected as a test set and the remaining $(k-1)$ folds as a training set with the assumption that there exists at least one sample per class. After the learning with k subsets of training samples, average test error can be obtained for a given model. Once all the models are evaluated, the parameters with the lowest average error was selected as final one for prediction.

In the paper, Gaussian RBF kernels are chosen for all the experiments since they are good general purpose kernels. In the primal SVM, two hyperparameters σ and C should be selected by model selection. Table 2 lists the hyperparameters for the SVM in the primal and a sequence of values for the corresponding hyperparameters is selected in the model selection with a grid search strategy. For the results obtained by Binary Hierarchical SVM (BH-SVM), Hierarchical SVM (HSVM) and Random Forest Classification and Regression Tree (RF-CART), we used the results reported in (Chen et al., 2004) for the fair comparison.

3.3. Experimental results

To simulate the small-size training problem, we only conduct experiments on training datasets containing 5% of original labeled samples for the "Spatially Correlated" (SC) and 'Spatially Uncorrelated" (SU) datasets. In the datasets, we have 10 splits of the small-size training set made up of 156 samples. On the analysis of Table 3, one can see that the classification error is already good with 156 training samples in the SC datasets. However, these spatially uncorrelated data have somewhat different characteristics from the training/test data, so the performance of all classifiers is reduced, as expected (Chen et al., 2004). In the following, we only concentrate on these challenging Spatially Uncorrelated (SU) datasets with the 10-split training sets containing 75%, 50%, 30%, 15% and 5% of the original labeled samples.

For a fair comparison, we used directly the classification results provided by the state-of-the-art algorithms for these SU datasets which were obtained in (Chen et al., 2004). The algorithms include BH-SVM, HSVM and RF-CART. Hierarchical Support Vector Machines (HSVM) approach utilizes a tree structure framework and solves a series of max-cut problems to perform the unsupervised class decomposition. Then, a dual SVM classifier is applied at

Table 2
Hyperparameters and a sequence of values for the hyperparameters in the model selection problem with a grid search strategy for the primal SVM

Hyperparameters	Grid searching
σ	$2^0, 2^1, 2^2, 2^3, 2^4, 2^5$
C	$10^0, 10^1, 10^2$

Table 3
Classification results with small-size training set made up of 5% of original samples

	SC datasets	SU datasets
Test Error (%)	10.02	29.28
Std. Dev. (%)	1.19	1.21

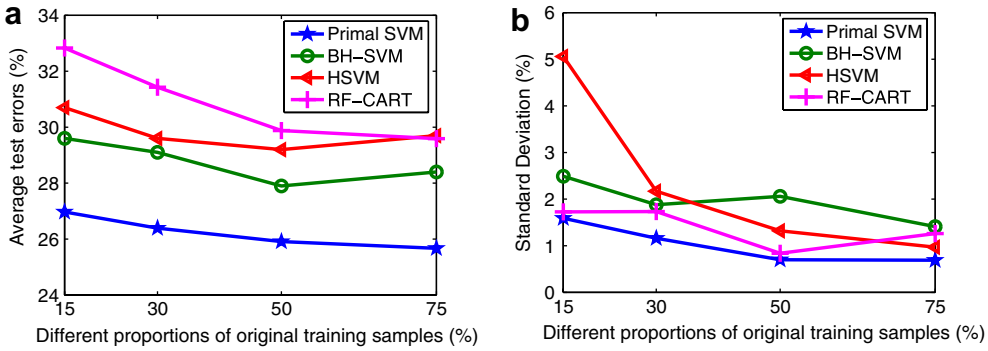

Fig. 2. In terms of different proportions of labeled samples, i.e., 75%, 50%, 30% and 15% of the SU datasets, the classification results provided by the SVM in the primal with comparison to those obtained by the state-of-the-art algorithms: (a) Test errors; (b) Standard deviation.

each internal node to construct the best discriminant function of a binary meta-class problem. The Random Forest Classification and Regression Tree (RF-CART) approach is based on a collection (forest) of CART-tree-like classifier systems, where the trees are grown to minimize an impurity measure based on CART. For the fair comparison, we use the same setting and so the completely the same datasets for our experiments as those used in (Chen et al., 2004).

Fig. 2 shows the classification results versus the different proportions of training samples provided by different algorithms, i.e., the primal SVM (Primal SVM in the figure), BH-SVM, HSVM and RF-CART for the SU datasets. Average classification errors for the SU test data made up of 2494 samples for the 10 experiments conducted with different algorithms are shown in Fig. 2a. The general trend shows that classification accuracies increase as the percentage of training samples involved for all four classifiers except for 75% sampling. However, for the introduced classifier SVM in the primal, the classification accuracies and the corresponding stability (standard deviation decreasing) increase with the higher quantity of training samples in the learning phase. On the analysis of Fig. 2, the primal SVM consistently obtain the best classification performance including average test errors (cf. Fig. 2a) compared to those provided by BH-SVM, HSVM and RF-CART.

4. Discussion and conclusion

In this paper, primal SVM (which is implemented in the primal formulation of optimization problem) has been introduced in the first time for the classification of hyperspectral remote-sensing data. The quadratic optimization problems with inequality constraints in L_2-norm SVM make most of the literature naturally to focus on the usage of Lagrange theory. However, the optimization problems on SVMs can be also carried out directly in the primal representation. Even though primal and dual optimization are equivalent in most cases, both in terms of the solution and time complexity, when it comes to approximate a solution, primal optimization is superior because it is more focused on minimizing what we are interested in: the primal objective function in (4) (Chapelle, 2007). For SVM in the primal with the objective function (4) in mind, the unconstrained problem can be implemented by any optimization technique. This has been verified by semi-supervised SVM, which is an alternative promising technique for the solution of small-size training set problem. Due to the non-convexity of objective function, different optimization techniques can lead to different results (Chapelle and Zien, 2005; Chapelle et al., 2006a,b; Chi and Bruzzone, 2007).

Although dual optimization problem can be written in terms of dot product by the ease of using kernel functions, the primal optimization can solve this problem using representer theorem (Kimeldorf and Wahba, 1971). As pointed out in (Chapelle, 2007), the computation complexity for the primal is $\mathcal{O}(nd^2 + d^3)$, and for the dual $\mathcal{O}(dn^2 + n^3)$. Using conjugate gradient, when $n \ll d$, the primal and dual steps have the same efficiency, on the other hand, the primal converges faster than the dual. For the objective with the quadratic loss, it is a strongly convex optimization problem as pointed out by (Mangasarian, 2002). Of course, we agree that when the kernel is not invertible, the solution is not unique, but we can add an infinitesimally small ridge to avoid this problem.

Further study will focus on the knowledge transfer learning using the primal SVM in the remote-sensing applications.

Acknowledgements

We thank Prof. Melba Crawford of Purdue University, W. Lafayette, IN, USA to kindly provide us datasets used in the experimental part of this paper. This work was supported by a Grant from the National Natural Science Foundation of China (No. 60705008) and by a Grant from the Ph.D. Programs Foundation of Ministry of Education of China (No. 20070246132).

References

Aronszajn, N. Theory of reproducing kernels. Transactions of the American Mathematical Society 68, 337–404, 1950.

Benediktsson, J., Sveinsson, J., Arnason, K. Classification and feature extraction of AVIRIS data. IEEE Transactions on Geoscience and Remote Sensing 33, 1194–1205, 1995.

Boyd, S., Vandenberghe, L. Convex Optimization. Cambridge University Press, Cambridge, UK. Available from:, 2002 http://www.stanford.edu/~boyd/cvxbook.html, 1995.

Breiman, L. Random forests. Machine Learning 45 (1), 5–32, 2001.

Burges, C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167, 1998.

Chapelle, O. Training a support vector machine in the primal. Neural Computation 19 (5), 1155–1178, 2007.

Chapelle, O., Chi, M., Zien, A. A continuation method for semi-supervised svms. In: 23rd International Conference on Machine Learning, 2006a.

Chapelle, O., Schölkopf, B., Zien, A. Semi-supervised Learning. MIT Press, Cambridge, MA, USA, 2006b.

Chapelle, O., Zien, A. Semi-supervised classification by low density separation. In: Tenth International Workshop on Artificial Intelligence and Statistics, 2005.

Chen, Y., Crawford, M., Ghosh, J. Integrating support vector machines in a hierarchical output space decomposition framework. In: Proceeding of IEEE International Geoscience and Remote Sensing Symposium (IGARSS'04), Alaska AK, USA, 2004.

Chi, M., Bruzzone, L. Semi-supervised classification of hyperspectral images by svms optimized in the primal. IEEE Transactions on Geoscience and Remote Sensing 45 (6), 1870–1880, 2007.

Critianini, N., Shawe-taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK, 2000.

Dare, J.A. Support vector machines in a binary hierarchical classifier. Master's thesis, University of Texas at Austin, 2004.

Gualtieri, J., Cromp, R. Support vector machines for hyperspectral remote sensing classification. In: Merisko, R.J. (Ed.), 27'th AIPR Workshop, Advances in Computer Assisted Recognition, vol. 3584. Proceeding of the SPIE, pp. 221–232, 1998.

Hoffbeck, J., Landgrebe, D.A. Covariance matrix estimation and classification with limited training data. IEEE Transactions on Geoscience and Remote Sensing 18 (7), 763–767, 1996.

Hsu, C., Lin, C. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13 (2), 415–425, 2002.

Hughes, G.F. On the mean accuracy of statistical pattern recognition. IEEE Transactions on Information Theory IT 14, 55–63, 1968.

Keerthi, S., DeCoste, D. A modified finite newton method for fast solution of large scale linear svms. Journal of Machine Learning Research 6, 341–361, 2005.

Kimeldorf, G., Wahba, G. Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications 33, 82–95, 1971.

Mangasarian, O.L. A finite newton method for classification. Optimization Methods and Software 17, 913–929, 2002.

Melgani, F., Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing 42 (8), 1778–1790, 2004.

Schölkopf, B., Smola, A. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

Tadjudin, S., Landgrebe, D.A. Covariance estimation with limited training samples. IEEE Transactions on Geoscience and Remote Sensing 37 (4), 2113–2118, 1999.

Vapnik, V.N. Statistical Learning Theory. John Wiley & Sons Inc., New York, 1998.