Gradient Optimization for multiple kernel’s parameters in support vector machines classification

A. Villa†, M. Fauvel†, J. Chanussot†, P. Gamba* and J. A. Benediktsson○

†GIPSA-lab, Signal & Image Dept., Grenoble Institute of Technology - INPG, France.
*Telecommunications & Remote Sensing Laboratory, Dept. of Electronics - University of Pavia, Italy.
○Dept. of Electrical and Computer Engineering, University of Iceland, Iceland.

Abstract—The subject of this work is the model selection of kernels with multiple parameters for support vector machines (SVM), with the purpose of classifying hyperspectral remote sensing data. During the training process, the kernel parameters need to be tuned properly. In this work a gradient descent based algorithm is used to estimate the parameters. The selection of multiple parameters is addressed, and an approach based on the analysis of the variance values of individual bands was proposed. Several state of the art kernels were tested. Experiments were conducted on real hyperspectral data. Results obtained with the different approaches/kernels were compared statistically, and showed good results in terms classification accuracies and processing time.

Index Terms—SVM, hyperspectral data, kernel methods, model selection.

I. INTRODUCTION

The classification of remote-sensing data has become an increasingly challenging problem, due to recent advances in remote sensor technology. One of the major difficulties in hyperspectral classification is the curse of dimensionality. The high number of features in a hyperspectral image is a major drawback for several reasons, such as the large number of training samples required and the Hughes’phenomenon [1]. Classification of this tremendous amount of data is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications. Since the beginning of 21st century, classifiers based on statistical learning theory have shown remarkable abilities to deal with both high-dimensional data and a limited training set. One of the best-known methods is the support vector machines (SVM), a kernel-based method which has found applications in many pattern recognition problems [2], [3]. Recent remote sensing literature has shown that SVM methods generally outperform both traditional classifiers and more advanced ones in classification problems involving hyperspectral images [4], [5]. However, there are still open issues that could allow further improvement of their performances such as model selection. The subject of this works is the model selection of kernels for the purpose of classification of hyperspectral remote sensing data. The selection of multiple parameters is addressed: Kernels with one parameter for each spectral band were tested (multi-sigma approach) and following [6] we studied covariance matrix to find groups of bands with similar variance value (grouped-sigma approach), using a gradient descent based approach to automatically optimize them [7]. Two different kernels were analysed in this work: The Gaussian kernel and the polynomial one. To assess the effectiveness of the method, two different hyperspectral data sets were used: An image of the engineering school of Pavia taken from airborne sensor ROSIS and a small segment of a satellite sensor AVIRIS image, taken over an agricultural area of Indiana.

II. SUPPORT VECTORS MACHINES

A. Generalities

The SVM is surely one of the most commonly used kernel learning algorithm. It performs robust non-linear classification of samples using the kernel trick [8]. Given a training set $S = \{(x^1, y_1), \ldots, (x^\ell, y_\ell)\} \in \mathbb{R}^n \times \{-1; 1\}$, the decision function is found by solving the convex optimization problem

$$\max_\alpha g(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \bar{k}(x^i, x^j)$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{\ell} \alpha_i y_i = 0$.

where $\alpha$ are the Lagrange coefficients, $\bar{k}(x_i, x_j) = k(x_i, x_j) + \delta_{ij}/C$, $k$ the kernel function, $C$ a constant that is used to penalize the training errors, $\delta_{ij}$ a function such that $\delta = 1$ if $i=j$, $\delta = 0$ otherwise. To be an acceptable kernel, $k$ should be a positive semi-definite function [8]. A short comparison of kernels for remotely sensed image classification can be found in [9].

When the optimal solution of (1) is found, the classification of a sample $x$ is achieved by looking to which side of the hyperplane it belongs:

$$y = \text{sgn}\left(\sum_{i=1}^{\ell} \alpha_i y_i k(x^i, x) + b\right).$$

(2)

Over the several multi-class strategies, one versus one [10] was the used approach in our experiments.

B. Kernel selection

Learning machines with different types of non linear decision function in the input space can be constructed using
different kernels. Much of the flexibility and classification power of support vector machine resides in the choice of kernel and in the selection of the kernel parameters; an accurate selection of kernel has a considerable influence on the learning capacity. In our experiments two classical kernels commonly used in remote sensing applications have been compared:

1) Gaussian kernel:

$$k_{\sigma}(x^i, x^j) = \exp\left(\frac{\|x^i - x^j\|^2}{2\sigma^2}\right).$$  \hspace{1cm} (3)

2) Polynomial kernel:

$$k_{\sigma}(x^i, x^j) = \left[\frac{(x^i, x^j)}{\sigma} + 1\right]^d$$  \hspace{1cm} (4)

where $\sigma$ is a parameter controlling the smoothness of the function which needs to be tuned correctly to fit the data properly and $d$ is the polynomial degree.

### III. HYPERPARAMETERS SELECTION

Hyperparameters selection is classically done by cross-validation: The training set is divided into $N$ parts, then the SVM is trained using $(N-1)$ parts and the obtained parameters are tested on the remaining part. This is done for all the parts and then the average accuracy is reported. Consequently, the parameters which lead to the best accuracies are selected. The major problem of cross validation is it can be very time consuming and so it is possible to tune only few parameters. In this paper another approach developed by Chapelle et al. [7] is investigated. It automatically tunes multiple parameters by minimizing some estimates of the generalization error of SVM using a gradient descent algorithm over the set of parameters.

This approach allows not only to find the hyperplane which maximizes the margin but also the values of the mapping parameters that yield best generalization error.

Since the true risk is not accessible, the idea of Chapelle is building an estimate of this risk $T$, which is a function of the kernel’s parameters. The minimization of this function yields the optimum parameters $p$ for the SVM. The radius margin bound is commonly used:

$$T(p) := \|w\|^2 / R^2$$  \hspace{1cm} (5)

where

$$\|w\|^2 = 2 \sum_{i=1}^{l} \alpha_i - \sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j k(x^i, x^j)$$  \hspace{1cm} (6)

and $R$ is the radius of the smallest sphere which contains the data:

$$R^2 = \sum_{i=1}^{l} \beta_i k(x^i, x^i) - \sum_{i,j=1}^{l} \beta_i\beta_j k(x^i, x^j).$$

The objective is to choose the values of the parameters such that $g(\alpha)$ is maximized and $T$, the model selection criterion, is minimized. Chapelle proposed an algorithm that alternates the SVM optimization with a gradient step in the direction of the gradient of $T$ in the parameter space. This can be achieved by the following procedure:

1) Initialize parameters $p$ to some value
2) Using a standard SVM algorithm, find the maximum of the quadratic form $g(\alpha)$

$$\alpha = \text{argmax} \ g(\alpha)$$  \hspace{1cm} (7)

3) Update the parameters $p$ such that $T$ is minimized. This is typically achieved by a gradient step
4) Go to step 2 until the minimum of $T$ is reached

In [11], this approach was positively used in a genetic optimization framework to estimate the best SVM parameters in a completely automatic way. In [12] gradient descent was favourably reviewed for the weight selection for spectral weighted kernels and also used for feature selection. Here the possibility to evaluate many parameters with gradient descent for classification purpose was investigated. For example, in the case of Gaussian kernel, a value of $\sigma^2$ per dimension was selected:

$$k_{\sigma}(x^i, x^j) = \exp\left(-\sum_{p=1}^{n} \frac{(x_{i}^{p} - x_{j}^{p})^2}{2\sigma^2_{p}}\right)$$  \hspace{1cm} (8)
IV. EXPERIMENTAL RESULTS

Two hyperspectral data sets were considered in this work. The first one is an airborne data from the ROSIS-03 with 115 spectral bands in the spectral range from 0.43 to 0.86 μm. The spatial resolution is 1.3m per pixel. The original data set is 610 by 340 pixels. Twelve data channels were removed due to noise. The remaining 103 spectral dimensions were processed. Nine classes of interest were considered. The second data set is a small segment of an AVIRIS data set over the agricultural area of Indiana. The data set is 145 by 145 set of pixel; it is composed of 220 spectral channels (spaced at about 10 nm) acquired in the 0.4-2.5 μm region. All 220 bands were processed, without removing noisy channels. Sixteen ground-truth classes were considered. Different training sets were randomly constructed from the reference data with 20 pixels by class; each experiment was repeated five times and the average results reported. Information about the training and test size of each data set are presented in Table I.

A. Single sigma

The first experiments were carried out computing one value of sigma for each decision function. At first, for the polynomial kernel, all three parameters C, σ and p were optimized using gradient descent method. Results showed a problem in the optimization of the polynomial degree. Therefore eight more polynomials have been evaluated, with fixed degree p=1,...,8, and gradient descent optimization of σ and C. The results presented in Table II show that the two kernels provide almost the same results in the case of the ROSIS data set, with an Overall Accuracy (OA) of about 78% and a κ of 72%. Processing time was much shorter with the Gaussian kernel. Analysis of the AVIRIS data set showed a considerable decrease of the OA and κ because of the larger number of classes. The polynomial kernel provides a strong improvement of classification accuracy (about 15% of OA and 18% of κ), with a slightly increased processing time.

B. Data scaling

The influence of data scaling in the classification accuracy and processing time was then investigated. Considering $x^i_p$ the $i$th value of the component $p$, $x^{min}_p$ and $x^{max}_p$ are the minimum and the maximum value of $x^i_p$. $\mu_p$ is the mean value and $\sigma_p$ the standard deviation of the band $p$, the tested algorithms were:

1) Data stretching: Stretches all data between 0 and 1 before the training process and the optimization of the parameters

$$x^i_p = \frac{x^i_p - x^{min}_p}{x^{max}_p - x^{min}_p} \quad (9)$$

2) Standardization: Subtracts from each sample value the mean value of the samples of its own band and divides by their standard deviations

$$x^i_p = \frac{x^i_p - \mu_p}{\sigma_p} \quad (10)$$

In the case of the ROSIS data set, scaling provided accuracy improvement for both tested kernels; while changes obtained with the Gaussian kernel were not statistically significant (less than 0.5% of OA), the polynomial kernel allowed a considerable improvement of OA and κ and a considerable saving of time with both scaling algorithm. In particular, standardization led to the best accuracies, since it provides an OA of 84% and κ of 80%. Table III shows complete results of analysis for the ROSIS data set. Data scaling on the AVIRIS data set provided a decrease in the classification accuracy with both tested kernels.

C. Multi sigma

The following experiments evaluated one parameter $\sigma_i$ for each spectral band (8). This way 103 values were selected for the ROSIS data set and 220 for the AVIRIS data set. For both the considered data sets, Gaussian kernel with multi-sigma approach led to a decrease of the classification accuracy; the polynomial kernel provided an improvement of OA and κ in the processing of the AVIRIS data set. Experiments gave conflicting results: A slight decrease in the κ corresponded to an improvement of the OA. Simple comparison of these

### Table I

<table>
<thead>
<tr>
<th></th>
<th>ROSIS</th>
<th>AVIRIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td>Total</td>
<td>3921</td>
<td>42776</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th></th>
<th>Rosis</th>
<th>Aviris</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Kernel</td>
<td>Gaussian</td>
<td>Polynomial</td>
</tr>
<tr>
<td>AA</td>
<td>85.78%</td>
<td>85.58%</td>
</tr>
<tr>
<td>OA</td>
<td>77.89%</td>
<td>77.31%</td>
</tr>
<tr>
<td>κ</td>
<td>72.34%</td>
<td>71.72%</td>
</tr>
<tr>
<td>Time</td>
<td>11 min</td>
<td>19 min</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Scaling method</th>
<th>Nothing</th>
<th>Data stretching</th>
<th>Standardization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian kernel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>85.78%</td>
<td>85.81%</td>
<td>85.16%</td>
</tr>
<tr>
<td>OA</td>
<td>77.89%</td>
<td>77.92%</td>
<td>77.50%</td>
</tr>
<tr>
<td>κ</td>
<td>72.34%</td>
<td>72.38%</td>
<td>71.79%</td>
</tr>
<tr>
<td>Time</td>
<td>12 min</td>
<td>25 min</td>
<td>18 min</td>
</tr>
<tr>
<td>Polynomial kernel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>85.58%</td>
<td>88.47%</td>
<td>89.38%</td>
</tr>
<tr>
<td>OA</td>
<td>77.31%</td>
<td>79.37%</td>
<td>84.47%</td>
</tr>
<tr>
<td>κ</td>
<td>71.72%</td>
<td>74.29%</td>
<td>80.22%</td>
</tr>
<tr>
<td>Time</td>
<td>58 min</td>
<td>20 min</td>
<td>17 min</td>
</tr>
</tbody>
</table>
values is not appropriate to evaluate the best approach, since it is explicitly based on the assumption that the two training set samples are independent, while in our test we used the same training set for single and multi-evaluated sigma. McNemar’s test was used to evaluate statistical significance of the difference. It is a non-parametric test based upon a standardized normal test statistic, proposed by Foody in [13]. In 4 of 5 analysis of AVIRIS data set, multi-evaluated sigma resulted in a significantly better performance than single-evaluated.

D. Grouped sigma

The last tested approach was a sort of compromise between the ‘multi-evaluated’ and the ‘single-evaluated’ sigma: The so-called ‘grouped-sigma’ approach. An analysis of the covariance matrices (Figure 1) shows groups of contiguous bands with similar values of variance for both data sets. This approach consists of computing a value of sigma for each group. This way two values were selected for the ROSIS data set and seven for the AVIRIS data set. This approach should grant a better generalization capability with respect to ‘single-sigma’ approach, with a shorter processing time respect to the ‘multi-sigma’. Further experiments showed that the choice of the bands gap is not critical, i. e. we have a little range of values for which the SVM performs equally. Experiments carried out on the ROSIS data set provided similar results to the ‘single-sigma’ approach, as could be expected having only two different values of sigma, while an improvement occurred with both kernels for the AVIRIS data set. McNemar’s test was used to compare grouped sigma to the best performing approaches in the analysis of the AVIRIS data set, that is ‘single-sigma’ for the Gaussian kernel and ‘multi-sigma’ for the polynomial. Grouped sigma resulted to perform significantly better in 4 of 5 tests with the Gaussian kernel, while for the polynomial one no significant differences were seen. We can conclude that in the case of a data set where several groups of bands with similar variance values are singled out, ‘grouped sigma’ performs better than traditional approaches, with a saving of computational time with respect to the case of ‘multi-sigma’.

V. CONCLUSIONS

Multiple parameters fitting was investigated for hyperspectral data classification. Two different kernels and several approaches were considered, both yielding good results in terms classification accuracies and elapsed time of data processing. The polynomial kernel outperformed the Gaussian one for both considered data sets. Data scaling resulted beneficial for the classification of the ROSIS data set, where noisy bands had been removed before processing. In particular data standardization proved to grant a substantial improvement of classification accuracy. Furthermore if it was joined with polynomial kernel, a shorter processing time was observed. Selection of multiple parameters was helpful in the case of the AVIRIS data set. There no pre-processing was performed and several groups of bands with similar variances could be singled out. The proposed approach ‘grouped-sigma’ gave the best results with only a slightly increased elapsed time with respect to traditional approaches. The construction of special kernels that can better adapt to the properties of hyperspectral remote sensing data will be explored in our future work.

REFERENCES