Second order statistics for hyperspectral data classification

Saoussen Bahria, Mohamed Limam

SECOND ORDER STATISTICS FOR HYPERSPECTRAL DATA CLASSIFICATION

Saoussen Bahria and Mohamed Limam

LARODEC laboratory- High Institute of Management- University of Tunis- 1007 Tunis- Tunisia

Abstract. Hyperspectral imagery classification, taking into account spectral and spatial features, is a promising task in remote sensing field. In this paper, the incorporation of second order statistics in hyperspectral data classification using support vector machines is proposed. The effect of using the semivariance geostatistic as a spatial feature rather than first order statistics (mean and standard deviation) is tested. The overall classification accuracy is evaluated for the AVIRIS Indian Pines-1992 benchmark data set. Empirical results show that the proposed approach gives better performance than the method based on first order statistics.

Keywords: Classification, spatial statistic.

Résumé. La classification des données hyperspectrales basée sur la combinaison des attributs spectraux et spatiaux, est une question prometteuse dans la télédétection spatiale. Ce papier propose l'incorporation des statistiques de second ordre dans la classification des données hyperspectrales en utilisant la technique des séparateurs à vaste marge ou support vector machines. L'effet de la considération de la semivariance comme attribut spatial au lieu des statistiques de premier ordre (moyenne et écart type) a été testé. La performance de classification a été évaluée pour l'mage hyperspectrale benchmark « AVIRIS Indian Pines-1992 ». les résultats empiriques montrent que la méthode proposée fournit une classification meilleure par rapport à la méthode considérant les statistiques de premier ordre.

Mots-clés. Apprentissage et classification, statistique spatiale.

Bibliographie

[1] CAMPS-VALLS, G., GOMEZ-CHOVA, L., MUNOZ-MARI, J., VILA-FRANCES, J., CALPE- MARAVILLA, J., 2006, Composite kernels for hyperspectral data classification. *IEEE Geosciences and Remote Sensing Letters,* **3**, pp. 93-97.

[2] GUALTIERI, J. A., and CHETTRI, S., 2000, Support vector machines for classification of hyperspectral data. *Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. IEEE 2000 International*. **2**, pp. 813 - 815.

[3] GOOVAERTS, P. 2002, Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data. *Journal of Geographical Systems*, **4**, pp. 99-111.

[4] YANG, F., XIATAO, L., GUANGZHOU, Z., CUIYU, S., XIAONING, S., 2006, Remote sensing image classification based on geostatistics and ANN, *Proceedings of the SPIE*, **6405**, pp. 64052C.

1. Introduction

In the remote sensing literature, many learning methods have been developed to tackle the problem of hyperspectral data classification. Most common approaches are based on spectral properties of the data. Gualtieri and Chettri (2000) discuss the application of support vector machines (SVM) for hyperspectral remote sensing data classification. They demonstrate the effectiveness of the SVM as a new machine learning tool suitable to handle high dimensional data sets, such as hyperspectral images. Kuo and Landgrebe (2001) discuss the problem of high estimation error in hyperspectral data classification due to limited training data. To avoid this problem, they propose a pair of covariance estimators called Mixed- Leave-One-Out Covariance. Liu *et al.* (2006) undertook a broad study of Purdue's Indian Pines test site. They compare classification performance of many supervised approaches including SVM.

Recently, a number of studies focusing on the improvement of hyperspectral data classification by including both spectral and spatial features have been conducted. Goovaerts (2002) presents a methodology to incorporate spatial coordinates of pixels and spectral properties in maximum likelihood classification. Melgani and Serpico (2002) describe a statistical approach for the combination of spectral and spatio-temporal contextual information for remote sensing data classification. Camps-Valls *et al*. (2006) propose a family of composite kernels that combine spatial and spectral information based on first order statistics (mean and standard deviation) as spatial features. Yang *et al*. (2006) combine textural and spectral information using the semivariogram and a back propagation neural network for remote sensing data classification. Huang *et al*. (2008) compare the accuracy of semivariogram geostatistical analysis with the co-occurrence matrix for spatial features incorporation in the remote sensing data classification.

The objective of this paper is to investigate hyperspectral data classification based on the incorporation of spatial features using SVM framework. In previous studies, only first order statistics are used as spatial features to improve traditional spectral classification. However, first order statistics, the mean and the standard deviation, provide simple local statistical information and ignore the spatial

arrangement structure of the data within the moving window. Therefore, in this paper the incorporation of second order statistics in hyperspectral data classification using support vector machines is proposed. The effect of using the semivariance geostatistic as a spatial feature rather than first order statistics is tested.

The paper is organized as follows. Section 2 presents the data and the methodology used in this work. Empirical results are discussed in Section 3. Advantages and limitations of the proposed approach compared to previous classification methods are discussed in Section 4.

2. Data and Methodology

2.1. Dataset: The AVIRIS Indian Pines-1992

The AVIRIS Indian Pines hyperspectral benchmark dataset is available online at http://aviris.jpl.nasa.gov/html/aviris.freedata.html. Ground reference data is available atftp.ecn.purdue.edu/pub/~biehl/PC_MultiSpec/ThyFiles.zip. The scene used for the experiment was acquired by NASA on 12 June 1992. It consists of 145 by 145 pixels (20 m spatial resolution) and 220 spectral bands, with about two-thirds agriculture, and one-third forest or other perennial vegetation. A total of 16 cover classes are identified in the ground reference data set. The scene is termed Indian Pines 1 in the AVIRIS JPL repository. A total of 20 bands [104-108, 150-163, 220] were removed to exclude the water absorption region. The remaining 200 spectral channels were used for the experimental study. A total of 10148 and 2031 pixels were used for training and testing, respectively. The numbers of training and testing samples for the 16 classes are given in table 1.

Table 1. Numbers of training samples for the 16 classes in the AVIRIS Indian Pines data set

Class Names	Training samples	Testing samples
1-Alfalfa	54	10
2-Corn-notill	1423	286
3-Corn-min	834	166
4-Corn	234	46
5-Grass/Pasture	497	99
6-Grass/Trees	747	149
7-Grass/Pasture-mowed	26	5
8-Hay-windrowed	489	97
9-Oats	20	4
10-Soybeans-notill	797	159
11-Soybeans-min	2468	493
12-Soybeans-clean	614	122
13-Wheat	212	42
14-Woods	1294	258
15-Bldg-Grass-Trees	380	76
16-Stone-steel-towers	95	19
Total samples	10184	2031

2.2. Methodology

A three-stage hyperspectral data classification process is proposed. First, the semi-variogram is used to measure the spatial correlation which is integrated as spatial features. Second, a stacked kernel matrix is constructed based on spectral and spatial features vectors. Third, the one-against-one multi-class SVM classification is carried out using the stacked kernel. According to Mather and Pal (2005), the one-against-one approach is the most efficient among the SVM classification methods for remote sensing data. It is called pair-wise classification since it creates an SVM classifier for each possible pair of classes. As a consequence, for an *m*-class problem, *m* (*m*-1)/2 binary classifiers were computed.

Two comparative studies were conducted. First, SVM classifications carried with and without spatial features were compared. Second, the effect of using the semi-variogram values as spatial features, rather than the first order statistics was tested. The proposed approach was compared to the method of Camps-Valls *et al*. (2006) based only on incorporating the first order statistics, the mean and the standard deviation as spatial features. In addition, experimental results were also compared with those reported in previous works dealing with the SVM classification of the same benchmark data set (Gualtieri and Chettri 2000, Liu *et al.*2006, Camps-Valls *et al*. 2006). For this purpose**,** the one-against-one multiclass SVM was carried out based on polynomial and Radial Basis Function (RBF) kernels. The 10-fold cross-validation method was used for the selection of the penalty parameter C, the optimal degree d for the polynomial kernel, and the Gaussian parameter σ for the RBF kernel. The cross-validation procedure provides the following optimal parameters C = 100, d = 2 for the polynomial

kernel, and C = 5000, σ = 40 for the RBF kernel. Then, the optimal selected parameters were introduced to the SVM classifiers. SVM classifications with and without spatial features were compared. Then, the effect of using second order statistics rather than first order statistics was tested. The polynomial kernel was considered for all comparisons in the second experiment since it outperforms the RBF kernel in the first study (results section, table 2).

To measure the spatial variation in the hyperspectral image a variogram geostatistic is used. In a regionalized variable, the variogram is used to quantify spatial correlation. A regionalized variable is a random variable, where location in space or time is known. In this formulation, variables are indexed by their location.

In a remote sensing image, the variogram is experimentally computed as the mean sum of squares of the differences between pairs of values separated by a distance *h* as follows

$$\gamma(h) = 1/2n \sum_{i=1}^{n} [Y(x_i + h) - Y(x_i)]^2 \quad (1)$$

Where γ is the estimated semivariance, n is the number of pairs of observations separated by *h*, x_i is a location, $Y(x)$ is the data value at the location *x*, $Y(x_i + h)$ is the data value at the location $x_i + h$.

The estimation of the semivariogram will be less precise at larger lag distance, since in that case, there are fewer pairs of observations. The common recommendation in geostatistical literature is to use small lag distance, in particular 1 pixel in remote sensing imagery. Consequently, a 5*5 moving window was used to calculate texture semi-variances at a lag of 1 pixel and angle of 0° in the vertical direction. The semi-variance values are then combined with spectral features in the stacked composite kernel. Errors matrices are determined for all classifiers. The overall classification accuracy and the concordance coefficient (Kappa) of all classifiers are derived from error matrices. Moreover, statistical significance of the differences between the different classifiers is assessed based on the Z statistics at the 95 percent confidence level.

For the experimental implementation, the OSU-SVM toolbox (http://www.eleceng.ohio-state.edu /~maj/osu_svm) is used with MATLAB 6.5 numerical workstation on a Pentium III 2.1 GHz personal computer with 3 GB of memory.

3. Results

The performance of SVM classification with and without spatial features is investigated. As shown in table 2, spatial features improve SVM hyperspectral data accuracy for the

two tested kernels. However, in our study the polynomial kernel performs better than the RBF kernel.

The overall classification accuracies and kappa coefficients for the different classifiers are reported in table 3. Numerically compared, the proposed approach gives superior classification accuracy than previous method using first order statistics as spatial features, reported by Camps-Valls et *al*. (2006). As a consequence, the semi-variogram geostatistical measure is more informative than the first order statistics and leads to better classification results with 96.55 % overall accuracy. Statistically, the proposed approach gives the highest kappa value 0.96 rather than 0.93 for the spatial-spectral SVM classifier with first statistics as spatial features and 0.87 for the spectral-SVM classifier.

Table 2. Overall accuracy (%) and kappa statistics (in parentheses) of the SVM classification carried out with and without spatial features on the AVIRIS Indian Pines data set

Classification	SVM without spatial features	SVM with spatial features
SVM-Polynomial kernel (C = 100, d = 2)	**87.54 (0.86)**	**96.55 (0.96)**
SVM-RBF kernel (C = 5000, σ = 40)	86.10 (0.85)	95.80 (0.95)

Table 3. Overall classification accuracy (%) and Kappa coefficient for different classifiers

Classifications	Overall accuracy	Kappa statistic
Spectral SVM		
Gualtieri and Chettri (2000)	87.30	0.86
Camps - Valls *et al.* (2006)	88.55	0.87
Liu *et al.* (2006)	93.77	0.93
Spatial-Spectral SVM		
Camps - Valls *et al*. (2006): mean and standard deviation as spatial features	94.21	0.93
Proposed Work: semivariance geostatistic as spatial features	**96.55**	**0.96**

4. Conclusion and discussion

In this work the incorporation of spatial arrangement structure of hyperspectral data in the stacked kernel based SVM classification is tested. The positive effect of using the semivariance values as spatial features rather than first order statistics is illustrated. The experimental study carried out with the AVIRIS Indian Pines-1992 benchmark data set shows that the semivariogram is more informative than first order statistics as spatial features and allows better classification accuracy.

Further improvements of hyperspectral data classification could be investigated by studying the effect of introducing more than one spatial feature. Moreover, applying multiple window sizes for the extraction of texture features could improve the classification accuracy.