Structured sparse model based feature selection and classification for Hyperspectral imagery

Author

Qian, Yuntao, Zhou, Jun, Ye, Minchao, Wang, Qi

Published

2011

Conference Title

2011 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS)

DOI

10.1109/IGARSS.2011.6049463

STRUCTURED SPARSE MODEL BASED FEATURE SELECTION AND CLASSIFICATION FOR HYPERSPECTRAL IMAGERY

Yuntao Qian [1], *Jun Zhou* [2,3], *Minchao Ye* [1], *and Qi Wang* [1]

[1] College of Computer Science, Zhejiang University, Hangzhou 310027, China
[2] Canberra Research Laboratory, National ICT Australia , Canberra, ACT 2601, Australia
[3] College of Engineering and Computer Science, The Australian National University Canberra, ACT 0200, Australia

ABSTRACT

Sparse modeling is a powerful framework for data analysis and processing. It is especially useful for high-dimensional regression and classification problems in which a large number of feature variables exist but the amount of training samples is limited. In this paper, we address the problems of feature description, feature selection and classifier design for hyperspectral images using structured sparse models. A linear sparse logistic regression model is proposed to combine feature selection and pixel classification into a regularized optimization problem with the constraint of sparsity. To explore the structured features, three-dimensional discrete wavelet transform (3D-DWT) is employed, which processes the hyperspectral data cube as a whole tensor instead of adapting the data to a vector or matrix. This allows more effective capturing of the spatial and spectral structure. The structure of the 3D-DWT features is imposed on the sparse model by group LASSO which selects the features on the group level. The advantages of our method are validated on the real hyperspectral data.

Index Terms— Hyperspectral imaging, Structure sparse models, Classification, Feature selection

1. INTRODUCTION

Classification of hyperspectral remote sensing data is not a trivial task due to the high variations of the spectral signature of identical material and the disequilibrium between small size of labeled samples and high-dimensionality of data. Most classification methods directly use the raw spectral signatures as features, and then apply different classification methods. The problem here lies in that the raw spectral features are very sensitive to noise corruption and the influence from the environment during the image capture process, and their dimensionality is high.

To solve this problem, many research efforts have focused on discriminative and robust feature extraction and selection. This is due to the fact that for classification, many image features can be extracted but useful structure is often contained in a subset of features, i.e. only part of the features are useful for discriminating different surface materials. However, most feature selection methods can be seen as an independent preprocessing step before classifier learning [1], which does not guarantee the consistency between two steps. Moreover, the prior information about the structure of features is often ignored in the feature selection step.

In this paper, our goal is to develop a unified framework for feature selection and hyperspectral pixel classification, which relies on structured sparse modeling and regularized optimization. Sparse signal modeling assumes that a signal can be efficiently represented by a sparse linear combination of atoms from a given or learned dictionary. The cardinality of the selected atoms is significantly smaller than the size of the dictionary and the dimension of the signal [2]. In the case of hyperspectral classification, the feature descriptor can be considered as the dictionary, the selected features as active atoms, and the classifier as linear sparse combination. The constraint of sparsity is imposed on the model as a regularizer, and the learning of sparse model is always a convex optimization problem.

Recently, sparse signal modeling has been extended to a more general case to deal with structured dictionary [3, 4]. Structured dictionary indicates that prior knowledge about atoms of the dictionary can be extracted. For example, if wavelet decomposition based subcomponents are used as the atoms, its wavelet tree can be seen as the structure of dictionary. The information about the structure of atoms can be mapped into the sparse model by various regularization schemes. In this framework, the above discussed problems in the current classification techniques can be overcome to a certain extent.

This work was supported by the National Natural Science Foundation of China No.60872071 and China-Australia Special Fund for Science and Technology Cooperation No.61011120054

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program

2. STRUCTURED SPARSE LOGISTIC REGRESSION

2.1. Sparse Group LASSO

Assume a prediction problem with N instances having outcomes $y_1, y_2, \cdots, y_N$ and features x_{ij}, where $i = 1, 2, \cdots, N$ and $j = 1, 2, \ldots, p$. Let $\mathbf{X}$ denotes the $N \times P$ input matrix, and $\mathbf{Y}$ denotes the $N \times 1$ output vector. The general linear regression model is given by

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where β is a vector of coefficients corresponding to the features, and ε is the noise. In order to estimate β, traditional optimization methods such as least square can be used, whose prediction performances, however, may not be good in many cases. Further, the physical interpretation of the solution may not be clear. Therefore, various constraints on β are widely studied.

One of the constraints commonly used is sparsity. In the case of hyperspectral imagery classification, sparsity means that only part of features are useful for discriminating the surface materials. The most popular sparse regression model is the LASSO proposed by Tibshirani [2], which is a regularized least square method imposing an L_1 penalty on the regression coefficients. It is defined as

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda|\beta|_1 \tag{2}$$

where $|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$. Owing to the nature of L_1 norm, the LASSO makes predication and variable selection simultaneously. It should be noted here that the LASSO considers all variable to be independent and non-correlative. Thus, it selects individual input variable.

In many cases sparsity alone may not be sufficient to obtain stable and desired solution. The prior knowledge of the structure of variables, such as grouping or hierarchical structure, can be supplementary to further improve the performance of sparse regularization. For this purpose, group LASSO has been proposed to use groups of the input variables instead of individual variables as a unit of variable selection [3], which is given by

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda \sum_{g=1}^{G} |\beta_{\Omega_g}|^2 \tag{3}$$

where $|\beta_{\Omega_g}|^2 = \sum_{j \in \Omega_g} \beta_j^2$. Such solution incorporates the grouping structure of input variables in the LASSO, while inducing sparsity in the group level and smoothness in the individual variable level.

A further extension of the group LASSO, namely the sparse group LASSO, combines the LASSO and group LASSO. It makes the coefficients β sparse not only between groups, but also in individual variables of each group [4], such that

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda_1 \sum_{g=1}^{G} |\beta_{\Omega_g}|^2 + \lambda_2|\beta|_1 \tag{4}$$

when $\lambda_1 = 0$, sparse group LASSO reduces to the LASSO, and when $\lambda_2 = 0$, it reduces to the group LASSO.

2.2. Constrained Logistic Regression

Classification is a special regression problem with discrete output. In binary classification, logistic regression models the conditional probability $p_\beta(y_i \mid \mathbf{x}_i)$ by

$$\log\left\{\frac{p_\beta(y_i \mid \mathbf{x}_i)}{1 - p_\beta(y_i \mid \mathbf{x}_i)}\right\} = \beta\mathbf{x_i} \tag{5}$$

Then we can obtain

$$P(y = y_i|x_i) = \frac{1}{1 + \exp(-y_i\beta_j\mathbf{x}_i)} \tag{6}$$

The maximum likelihood estimation of β is

$$\hat{\beta} = \arg\max_{\beta} \left\{\sum_{i=1}^{N} -\log(1 + \exp(-y_i\beta_j\mathbf{x}_i))\right\} \tag{7}$$

Adding the structured sparse constraint, the structured sparse logistic regression can be defined as [5]

$$\hat{\beta} = \arg\min_{\beta} f_1(\beta) + \lambda f_2(\beta) \tag{8}$$

where $f_1(\beta) = \sum_{i=1}^{N} \log(1 + \exp(-y_i\beta_j\mathbf{x}_i))$ and $f_2(\beta)$ is the structured sparse constraint of β, which is the same as that in the corresponding regression model. It should be noted that the one-to-all scheme can be used to deal with the multiclass problem.

In the last decade, optimization of problems in the form of (2), (3), (4) and (8) have been deeply studied which leads to very efficient solutions. The most popular algorithm is the block coordinate decent, which minimizes a multi-variate objective function by partitioning the parameters/coordinates into several blocks and circularly optimizes the blocks until convergence [6]. In each iteration, one of the coordinate blocks is updated while the other coordinates remaining fixed. Block coordinate decent algorithm for the logistic group LASSO is outlined below.

Algorithm: Logistic group LASSO using block coordinate decent

1. Let $\beta \in \mathbf{R}^p$ be an initial parameter vector
2. Iterate over groups β_{Ω_g}, and $g = 1, 2, \cdots, G$
   (a) if $f_1(\beta_{-\Omega_g}) \leq \lambda$
   $\beta_{\Omega_g} \leftarrow \mathbf{0}$
   where $\beta_{-\Omega_g}$ is the parameters not in group Ω_g

(b) else

$$\beta_{\Omega_g} \leftarrow \arg\min_{\beta_{\Omega_g}} f_1(\beta) + f_2(\beta)$$

where $f_2(\beta) = \lambda \sum_{g=1}^{G} |\beta_{\Omega_g}|^2$

3. Until convergence criterion is met

In step 2, we first check the importance of a group Ω_g. If the importance is lower than a threshold λ, zero vector is assigned to the parameters β_{Ω_g} in this group. On the contrary, if the group is important, an optimal solution is found with respect to β_{Ω_g}.

3. 3D-DWT BASED STRUCTURED FEATURE EXTRACTION

To capture the structured features in hyperspectral imagery, three-dimensional discrete wavelet transform (3D-DWT) is used. Formally, standard wavelet transform can decompose a function $f(x)$ into a linear combination of wavelet and scaling functions $\psi(x)$ and $\varphi(x)$. In discrete wavelet transform (DWT), wavelet and scaling functions are represented by the filter bank $(\widetilde{G}, \widetilde{H})$ given by the lowpass and highpass filter coefficients $g[k]$ and $h[k]$ respectively. Multidimensional DWT can be carried out by a series of one-dimensional DWT, and 2^n (n is the number of dimensions) subcomponents are produced at each level.

Here, we used the Haar set as the wavelet filter bank $(\widetilde{G}, \widetilde{H})$. The Harr wavelet decomposes the hyperspectral data cube into two levels. In order to assign every pixel a coefficient vector in each subcomponent, the down-sampling step is removed. Altogether, 15 subcomponents $\mathbf{W}_1, \ldots, \mathbf{W}_{15}$ are produced, and each has the same size as the original cube.

The wavelet coefficients of each pixel (i, j) can be directly used as its feature vector.

$$\mathbf{s}(i,j) = (\mathbf{W}_1(i,j,\cdot), \mathbf{W}_2(i,j,\cdot), \cdots, \mathbf{W}_{15}(i,j,\cdot)) \quad (9)$$

In order to account for the texture characteristics of spatial distribution in the hyperspectral imagery [7], we apply a mean filter to the computed feature vectors

$$\hat{\mathbf{s}}_n(i,j,\cdot) = \frac{1}{9} \sum_{a=i-1}^{i+1} \sum_{b=j-1}^{j+1} |W_n(a,b,\cdot)| \quad (10)$$

Finally, the 3D-DWT based texture features are computed

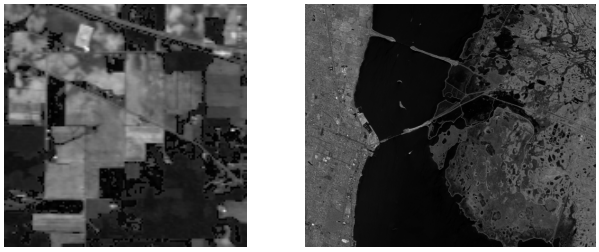
$$\hat{\mathbf{s}}(i,j) = (\hat{\mathbf{s}}_1(i,j,\cdot), \hat{\mathbf{s}}_2(i,j,\cdot), \cdots, \hat{\mathbf{s}}_{15}(i,j,\cdot)) \quad (11)$$

For 3D-DWT based texture features, each decomposed subcomponent (sub-cube) is defined as a group. The texture features of each pixel in the subcomponent form a group of input variables. This generates 15 groups of input variables, and the number of variables in each subcomponents is the same as the number of spectral bands in the original data cube.

4. EXPERIMENTS

We employed two real-world remote sensing hyperspectral images acquired by NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) instrument, so as to evaluate the performance of the proposed method. The first image was acquired over the Indian Pine Test Site in Northwestern Indiana in 1992 [8], whose 70th band is shown in Fig. 1(a). This image contains 16 land-cover classes and 10366 labeled pixels, while 185 bands were used for the experiments. The second image was acquired over the Kennedy Space Center (KSC), Florida in 1996 [9]. Its 50th band image is shown in Fig. 1(b). There are 13 land-cover classes with 5211 labeled pixels, while 176 bands were used for the analysis. In the experiments, we finished the following tasks:

- Comparing 3D-DWT based structured features against the raw spectral signatures,
- Comparing the structured sparse method (the logistic group LASSO model) against the support vector machines (SVM) with linear and radial basis kernels, which are denoted by SVM(lin) or SVM(rbf), respectively.
- Testing the validity of feature selection using the structured sparse method.



(a) (b)

Fig. 1. (a) Band 70 of AVIRIS Indian Pine. (b) Band 50 of AVIRIS KSC.

In order to simulate the real-world situations where only few labeled samples are available, we evaluated the performance of classification in two settings: 5% and 25% of the labeled pixels were randomly selected as the training samples, while the rest were used as the testing set. The sparsity of the model was controlled by a single model parameter λ, which measures the ratio of the unselected features (whose coefficients are zero) over all features.

Table 1 shows the average classification accuracies for the AVIRIS test data. From this table, several observations can be made. Firstly, 3D-DWT features are generally much better than the raw spectral features, which shows structured feature extraction is very useful for hyperspectral data. Secondly, in most cases, the results obtained by SVM(rbf) are better

Table 1. Average accuracy results on AVIRIS datasets at different sampling rates.

Features	Classifiers	Indian Pine		KSC	
		5%	25%	5%	25%
Spectrum	LASSO	0.670	0.783	0.849	0.911
	SVM(lin)	0.637	0.770	0.795	0.885
	SVM(rbf)	0.742	0.861	0.865	0.923
3D-DWT	Group LASSO	0.879	0.943	0.897	0.975
	SVM(lin)	0.883	0.976	0.777	0.940
	SVM(rbf)	0.901	0.979	0.845	0.941

than those from the structured sparse method and SVM(lin). This is due to the nonlinear characteristics in the hyperspectral data. However, in the case of KSC data with 3D-DWT features, group LASSO performs the best. Thirdly, on KSC data, the structured sparse method is better than SVM(lin) and is close to SVM(rbf) in accuracy. However, its computational cost is less than SVM(rbf) because it is a linear solution.

To illustrate the effectiveness of the feature selection step of the proposed method, Fig. 2 shows the results on classification accuracies versus sparsity. In the figure, the sparsity of the model is displayed in the blue curve, which is controlled by λ. We used the selected features as the inputs to SVM(rbf) so as to investigate the validity and discriminative power of the selected features. It can be seen that SVM(rbf) with selected features can produce close or better classification results than using all features. This observation suggests that the feature selection by structured sparse method is very effective, which can reduce feature dimensions while preserving the discriminative ability of the features.

5. CONCLUSIONS

In this paper, we applied structured sparse model to feature selection and pixel classification for hyperspectral imagery. For these purposes, a logistic group LASSO model with 3D-DWT based texture features were developed. This method allows integration of the feature selection and classification steps into a unified model by minimizing the combined empirical loss and penalization on sparsity while taking into account the structure of the input features. We have illustrated the advantages of this method on real-world data and compared our method against raw spectral features based sparse model and SVMs.

6. REFERENCES

[1] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Computer Vision*, vol. 3, no. 4, pp. 213–222, 2009.

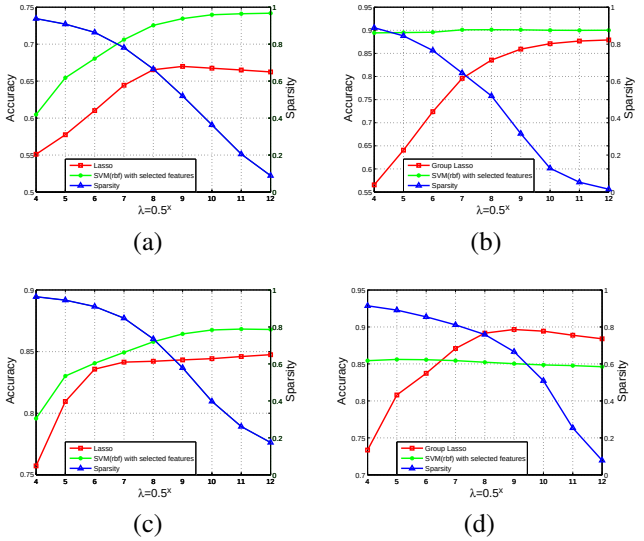


Fig. 2. Results on AVIRIS datasets with 5% sampling rate. (a) Indian Pine with spectral features. (b) Indian Pine with 3D-DWT features. (c) KSC with spectral features. (d) KSC with 3D-DWT features.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

[3] M. Yuan and Y. Li, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B*, vol. 68, pp. 49–67, 2006.

[4] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 37, no. 6A), pp. 3468–3497, 2009.

[5] L. Meier, S. Geer, and P. Buhlmann, "The group lasso for logistic regression," *J. R. Statist. Soc. B*, vol. 70, pp. 53–67, 2008.

[6] P. Tseng, "Convergence of block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, 2001.

[7] X. Zhang, N.H. Younan, and C.G. O'Hara, "Wavelet domain statistical hyperspectral soil texture classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 615 – 618, 2005.

[8] "AVIRIS NW indiana's indian pines 1992 data set," http://dynamo.ecn.purdue.edu/biehl/MultiSpec.

[9] J. Ham, Y. Chen, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.