Dynamic Learning of SCRF for Feature Selection and Classification of Hyperspectral Imagery

Ping Zhong*, Zhiming Qian, and Runsheng Wang

ATR National Laboratory, School of Electronic Science and Engineering,
National University of Defense Technology, 410073, Changsha, Hunan, China
{zhongping,qianzhiming,rswang}@nudt.edu.cn

Abstract. This paper investigates the feature selection and contextual classification of hyperspectral images through the sparse conditional random field (SCRF) model. To relieve the heavy degeneration of classification performance caused by the characteristics of the hyperspectral data and the oversparsity when SCRF selects a small feature subset, we develop a dynamic learning framework to train the SCRF. Under the piecewise training framework, the proposed dynamic learning method of SCRF can be implemented efficiently through separated dynamic sparse trainings of simple classifiers defined by corresponding potentials. Experiments on the real-world hyperspectral images attest to the effectiveness of the proposed method.

Keywords: Conditional random field, classification, feature selection.

1 Introduction

Hyperspectral image analysis is attracting a growing interest in real world applications, such as urban planning, mapping, agriculture, forestry, and disaster prevention and monitoring. Many these applications can be finally transformed into some classification tasks. In the literature, many techniques have been developed for the classification purpose, including support vector machines [1, 2], neural networks[3], graph method[4–6], and others. Many algorithms take into consideration only spectral variations, ignore spatial correlations, and treat each site independently. However, hyperspectral images show strong correlations across spatial and spectral neighbors[6], which have been proved to be very useful for image analysis in both the remote sensing and computer vision communities.

Markov random fields (MRFs) are the classical probabilistic approaches for modeling the contextual information in label images. However, for computational tractability, the observed data are assumed to be conditional independent, which neglects the contextual information in the observed data of a given class. Conditional random fields (CRFs) have recently gained popularity since they have the ability to incorporate contextual information in the labels as well as the

* This research was conducted with support of the NSF of China (Grant No. 60902088 and 61271439) and NDTF Project of ATR Lab. (Grant No. 9140C8004011005).

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 254–263, 2012.

observations[7]. But as for other supervised classifiers, excessive large number of spectral features may bring on the well-known overfitting problem for CRFs[8, 9]. Moreover, it is inefficient to use many irrelevant features due to the increased computational complexity.

Reduction in the number of features thus can be a direct way to overcome the overfitting and save the computational cost. Recently, there have several approaches to select the relevant features for the classical log-linear CRFs with potentials defined as simple linear combinations of features. But for the extended CRF with potentials defined as discriminative classifiers, the log-likelihood cannot guarantees to be an additive function of features. Thus it may be difficult to use the methods directly to select features for the extended CRFs. In contrast, we addressed the feature selection problem during training by adding a sparsity-promoting regularizer to the log-likelihood in the form of a log Laplacian prior on the model parameters[9]. The trained sparse model is named sparse CRF (SCRF) model.

In this work, we go one step further to demonstrate that as the generalized linear models (GLMs), SCRFs may suffer from the heavy degeneration of classification performance when they select small feature subset. This work develops a dynamic learning method of the SCRF (D-SCRF, for short) to relieve the negative effects of the problem on the classification performance. Moreover, we will show that under the piecewise training framework, the dynamic learning of SCRF can be efficiently implemented through two separated dynamic trainings of Sparse Multinomial Logistic Regression (SMLR) models.

2 SCRF for Feature Selection and Classification

In hyperspectral image classification, the observed data y is considered to be a set of spectral vectors $\{y_1, y_2, ..., y_I\}$, where $y_i = [y_{i1}, y_{i2}, ..., y_{iD}]^T$ denotes a spectral vector associated with an image site $i \in S$. D is the number of spectral bands and $S = \{1, 2, ..., I\}$ is the set of image sites. The label set is given by $x = \{x_1, x_2, ..., x_I\}$, where $x_i \in \{1, 2, ..., L\}$ and L is the number of classes.

The CRF for hyperspectral image classification directly models the posterior as

$$P(x|y,\theta) = \frac{1}{Z} \exp\left\{ \sum_{i \in S} \phi_i(x_i, y, w) + \sum_{i \in S} \sum_{j \in \eta_i} \xi_{ij}(x_i, x_j, y, v) \right\} \tag{1}$$

where Z is a normalizing constant known as the partition function. The unary clique potential $\phi_i(.)$ is defined as multinomial logistic regression (MLR) model:

$$\phi_i(x_i, y, w) = \sum_{l=1}^{L} \delta(x_i = l) \log P(x_i = l|y, w) \tag{2}$$

where

$$P(x_i = l|y, w) = \begin{cases} \frac{\exp\left(w_l^T y_i\right)}{1 + \sum_{k=1}^{L-1} \exp\left(w_k^T y_i\right)} & if\ l < L \\ \frac{1}{1 + \sum_{k=1}^{L-1} \exp\left(w_k^T y_i\right)} & if\ l = L \end{cases} \tag{3}$$

w_k is the parameter vector $[w_{k1}, ..., w_{kD}]^T$ for $k\,th$ class. The pairwise clique potential $\xi_{ij}(.)$ is defined as a generalization of the Ising model[10]:

$$\xi_{ij}(x_i, x_j, y, v) = \sum_{k,l \in \{1,...,L\}} v_{kl}^{\mathrm{T}} \mu_{ij}(y)\, \delta(x_i = k)\, \delta(x_j = l) \tag{4}$$

where v_{kl} is the parameter vector and $\mu_{ij}(y)$ is a spectral feature vector obtained by concatenating all elements of two vectors y_i and y_j.

The parameters $\theta = \{v, w\}$ is said to be sparse if and only if many of its entries are exactly zero. The sparsity is associated with the definition of feature selection. So the feature selection can be implemented by the sparse trainings of the model parameters. Let $\{\tilde{x}, \tilde{y}\} = \{\tilde{x}_c, \tilde{y}_c\}_{c \in \tilde{C}}$ be the selected training samples. The sparse training is implemented as a maximum a posteriori (MAP) estimate

$$\tilde{\theta} = \arg\max_{\theta} Q(\theta) = \arg\max_{\theta} \left(L(\theta) - \lambda_\theta \|\theta\|_1\right) \tag{5}$$

where $\|\theta\|_1 = \sum_n |\theta_n|$ denotes the l_1 norm of the parameters θ in the sparsity-promoting Laplacian distribution and $L(\theta)$ is the log-likelihood.

3 Dynamic Learning of SCRF

The sparsity of the parameter set θ is controlled by the regularization parameter λ_θ. The larger is λ_θ, the greater is sparsity. Excessively large values of $\lambda_\theta$will result in under-fitting, while excessively small values of λ_θ could result in over-fitting. In the literature of l_1 regularization, the cross-validation method is usually used to select the optimum λ_θ from predefined values [11], which are fixed through the whole training procedures. However, as for the generalized linear model (GLM), the fixed-value-based method may bring two problems for the SCRF. Firstly, to select relative small feature subset, SCRF should be trained with large values of λ_θ. But the fixed excessively large parameter can result in the over-sparsity. Secondly, each band of hyperspectral data contains some information but only some of the bands have significant effects on output. Such characteristics also prevent the optimal λ_θ derived from fixed-value-based methods from obtaining high level of performances[12].

Both the problems are derived essentially from negative effects of the too many irrelevant or weakly relevant features on the classifier. So a direct method dealing with the problems is to get rid of the obvious irrelevant features on the basis of their relevance or discriminant powers with regard to the targeted classes before training. But the primary feature selection procedure is not correlated to the SCRF model. In contrast, we develop a dynamic learning method to incorporate the primary feature selection procedure into the training of SCRF. As mentioned earlier, the larger is λ_θ, the greater is the sparsity, which means more features are discarded. Based on this conclusion, the dynamic learning makes the λ_θ vary during iterative training: the large values of the λ_θ are utilized to get rid of the obvious irrelevant or weakly relevant features at the earlier iterations; then the

later iterations arrives the convergence and obtains the superior classifier. For the remainder of this work, the variable parameter is denoted as λ_θ^α, and then we get the objective function of dynamic learning framework as

$$Q^\alpha(\theta) = L(\theta) - \lambda_\theta^\alpha \|\theta\|_1 \tag{6}$$

3.1 Piecewise Implementation of Dynamic Learning

Because $\|\theta\|_1 = \sum_n |\theta_n|$ is a nondifferentiable term at the origin, the usual gradient-based methods cannot be directly utilized to maximize the objective function. In this work, we develop an efficient sparse training method under the piecewise training framework. Firstly, $L(\theta)$ is divided according to the types of the cliques. Let $\tilde{C}_m$ be the set of the type of cliques with m sites selected for model training. Then the divided graph factor a is a clique c in the set $A = \left\{\tilde{C}_m\right\}_{m=1,2,\ldots} \triangleq \tilde{C}$, and consequently, the divided factor $f_a(\tilde{x}_a, y)$ of $L(\theta)$ is exactly the potential $\psi_c(\tilde{x}_c, \tilde{y}, \theta)$. Finally, the piecewise dynamic training of SCRF with the special division is to maximize the objective function

$$Q_{PW}^\alpha(\theta) = \sum_{c\in\tilde{C}} \log \frac{\psi_c(\tilde{x}_c, \tilde{y}, \theta)}{\sum_{x_c} \psi_c(x_c, \tilde{y}, \theta)} - \lambda_\theta^\alpha \|\theta\|_1 \tag{7}$$

Consider only up to pairwise clique potentials, then Eq. (7) can be rewritten as

$$\begin{aligned} Q_{PW}^\alpha(w, v) = &\underbrace{\left(\sum_{i\in\tilde{C}_1} \log \frac{\exp\{\phi_i(\tilde{x}_i, \tilde{y}, w)\}}{\sum_{x_i} \exp\{\phi_i(x_i, \tilde{y}, w)\}} - \lambda_w^\alpha \|w\|_1\right)}_{Q_w^\alpha} \\ &+ \underbrace{\left(\sum_{(i,j)\in\tilde{C}_2} \log \frac{\exp\{\xi_{ij}(\tilde{x}_i, \tilde{x}_j, \tilde{y}, v)\}}{\sum_{x_i, x_j} \exp\{\xi_{ij}(x_i, x_j, \tilde{y}, v)\}} - \lambda_v^\alpha \|v\|_1\right)}_{Q_v^\alpha} \end{aligned} \tag{8}$$

Eq. (8) shows that under piecewise training framework with the special division, D-SCRF can be trained by independently dynamic training the local sparse classifiers over each kind of cliques.

In the first term in Eq.(8), the unary potential modeled as MLR in Eq. (3) has the normalization condition as $\sum_{l=1}^{L} P(x_i = l|y, w) = 1$. So the denominator of the first term in Eq. (8) is just the constant one. We then immediately have

$$Q_w^\alpha = \sum_{i\in\tilde{C}_1} \log P(\tilde{x}_i|\tilde{y}, w) - \lambda_w^\alpha \|w\|_1 \triangleq L_{MLR}(w) - \lambda_w^\alpha \|w\|_1 \tag{9}$$

Since $P(\tilde{x}_i|\tilde{y}, w)$ is defined as MLR (Eq. (4)), $L_{MLR}(w)$ is log-likelihood of MLR and then Eq. (9) is exactly the objective function of D-SMLR [13].

In the second term in Eq.(8), Q_v^α can be written as

$$Q_v^\alpha = \sum_{i,j \in \tilde{C}_2} \log P\left(\tilde{x}_i, \tilde{x}_j | \mu_{ij}\left(\tilde{y}\right), v\right) - \lambda_v^\alpha \left\|v\right\|_1 \triangleq L_{MLR}\left(v\right) - \lambda_v^\alpha \left\|v\right\|_1 \tag{10}$$

where

$$P\left(\tilde{x}_i = k, \tilde{x}_j = l | \mu_{ij}\left(\tilde{y}\right), v\right) = \frac{\exp\left(v_{kl}^{\mathrm{T}} \mu_{ij}\left(\tilde{y}\right)\right)}{\sum_{m=1}^{L} \sum_{n=1}^{L} \exp\left(v_{mn}^{\mathrm{T}} \mu_{ij}\left(\tilde{y}\right)\right)} \tag{11}$$

Eq. (11) shows that $P\left(\tilde{x}_i, \tilde{x}_j | \mu_{ij}\left(\tilde{y}\right), v\right)$ acts as a MLR model with L^2 classes, and then $L_{MLR}\left(v\right)$ is also the log-likelihood of MLR and Eq. (10) is exactly the objective function of D-SMLR.

Therefore, we can draw the conclusion that with the potentials defined as Eq. (3) and (4), the dynamic training of the SCRF can be implemented as exactly two kinds of dynamic sparse MLR (D-SMLR) models under the piecewise training framework. The D-SMLR is implemented through changing the hyperparameter λ_θ^α under the iterative training framework.Then the varied hyperparameter is relevant to the iterations and λ_θ^α can be further denoted as $\lambda_\theta^{(t)}$. In this work, we use the following function of varied hyperparameter with the variable t

$$\lambda_\theta^{(t)} = \rho_{\theta,1} * \beta^t + \rho_{\theta,2} \tag{12}$$

where $0 \leq \beta < 1$, $\rho_{\theta,1}$ and $\rho_{\theta,2}$ are positive constants. More details of derivation of the D-SMLR algorithm can be found in [13].

3.2 Model Combination in Inference

We noted that the D-SCRF training through independent D-SMLR trainings may leads to problems with over-counting during inference[14].We introduce scalar powers for each term, and then combine the independently trained models during inference as

$$P\left(x|y\right) \propto \exp\left\{\gamma_1 \left[\sum_{i \in S} \phi_i\left(x_i, y, \tilde{w}\right)\right] + \left[\sum_{i \in S} \sum_{j \in \eta_i} \xi_{ij}\left(x_i, x_j, \mu_{ij}\left(y\right), \tilde{v}\right)\right]\right\} \tag{13}$$

where $\tilde{w}$ and $\tilde{v}$ are the optimal D-SMLR parameters learned independently, and γ_1 is the fixed power for the unary potential. The inference of the form (13) can be efficiently implemented by loopy belief propagation (LBP).

4 Experimental Results

4.1 Data Set for Experiments

The proposed algorithm was tested on real world hyperspectral image. The data consist of a 145x145 pixels portion of an AVIRIS image acquired over NW Indian

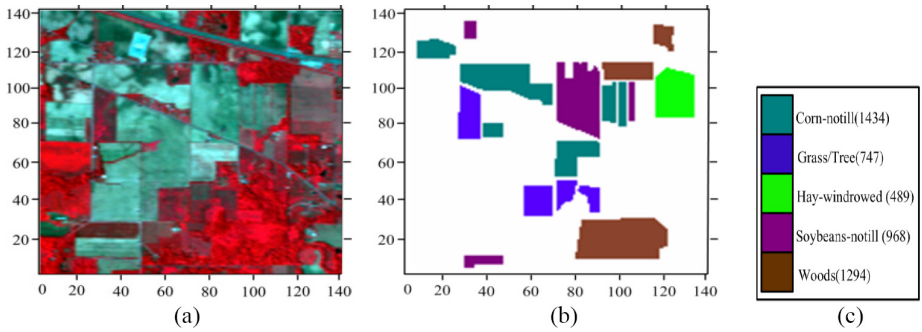

Fig. 1. Indian Pine data set. (a) is original image produced by the mixture of three bands. (b) is ground truth with five classes. (c) is map colour and number of samples.

Table 1. Number of total, training, and test Samples in Indian Pine data set

class Name	total	training	test
corn-notill	1434	500	934
grass/Tree	747	260	487
hay-windrowed	489	172	317
soybeans-notill	968	340	628
woods	1294	452	842
total	4932	1724	3208

Pine in June 1992[15]. In our experiments, all of the 220 original spectral channels were employed and five classes were selected from the efficiency point of view only (see Fig. 1). We randomly select the spatially joint pairwise pixels to create the training dataset. The details of training and testing pixels for each class are listed in table 1.

4.2 Convergence

At first, we evaluate performances of the dynamic training method. The convergence property of the training method is illustrated in Fig. 2 through the plots of gradients with change of iteration times. Since there are total 880 w_{ij} $(i = 1, ..., 220, j = 1, ..., 4)$ and 2200 v_{ij} $(i = 1, ..., 440, j = 1, ..., 5)$ in this experiment setup, it is impossible to demonstrate the gradients of all parameters. Without losing generality, we present only the gradients of the parameters corresponding to the first dimension in the feature vectors. As shown in Fig. 2, both the training processes show convergences with more than 100 iterations.

4.3 Classification Behavior with Different Number of Selected Features

Then, we present the classification performances of SCRF and D-SCRF with the different number of selected features. The SCRF is also trained by the piecewise

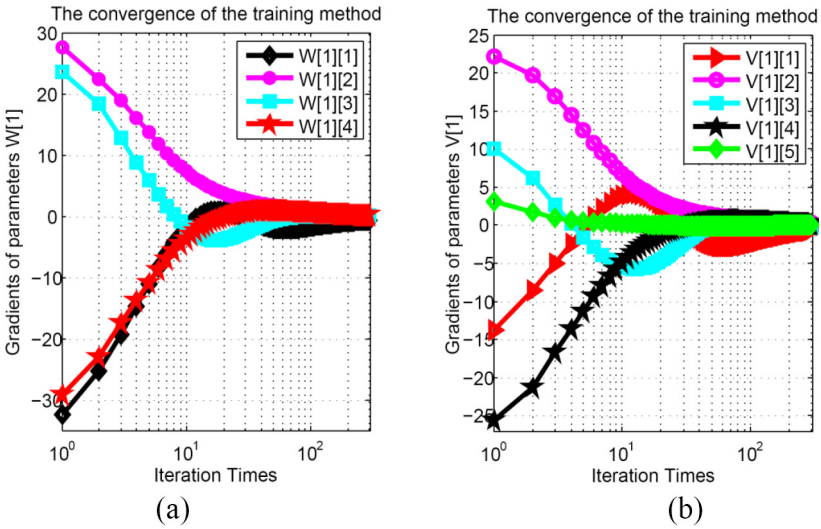

Fig. 2. Convergence of the training method. (a) is the plots of the gradients of $\{w_{1j}, j = 1, ..., 4\}$. (b) is the plots of the gradients of $\{v_{1j}, j = 1, ..., 5\}$.

training method presented in Section 3.1. At first, we demonstrate the classification behaviors of the two models with only unary (setting v as 0) or pairwise (setting w as 0) clique potentials respectively. Then we combine the unary and pairwise clique potentials to get the full SCRF and full D-SCRF through Eq. (13). Similar to that in work[14], the power parameter γ_1 in Eq. (13)was learned as 0.1 through cross validation.

In all the figures, SCRFs and D-SCRFs show similar classification accuracies when relatively large numbers of features are selected. However, with the decreasing number of selected features, the plot of the SCRFs drops sharply for the undersparsity and the characteristics of the hyperspectral data, while the D-SCRFs show more stable classification performance. This means that the D-SCRFs relieves the heavy degeneration of classification performance caused by the undersparsity in the SCRFs and can be more fit for the feature selection in the classification of hyperspectral data. Fig. 3(a) also demonstrates that the full CRFs show better results than the corresponding CRF models with only unary (Fig. 3(b)) or pairwise (Fig. 3(c)) clique potentials since the full CRFs combine their strengths.

4.4 Quantitative Evaluation

Table 2 presents the performances of SCRF and D-SCRF with the selected 5% of total features over the Indian Pine. The SCRF obtained 90.85% overall classification accuracies, in contrast the D-SCRF achieved higher 93.56% accuracies. The inspection of the accuracy for each class confirms that except the grass/tree, D-SCRF obtained higher accuracies than SCRF for other classes. The higher

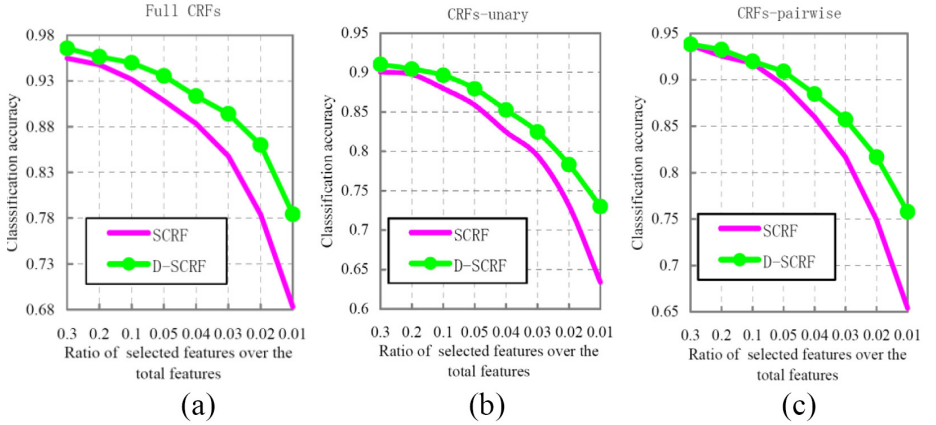

Fig. 3. Classification accuracies of different SCRFs and D-SCRF against number of selected features. (a) is results of full SCRF and full D-SCRF. (b) and (c) show the results of SCRFs and D-SCRFs with only unary and pairwise clique potentials respectively.

accuracy of SCRF for the grass/tree class may derive from the fact that the D-SCRF used the same varied hyperparameter for all the classes. The D-SCRF model can further improve the classification accuracy of each class by setting different varied hyperparameters for different classes. We also give the performance of D-SMLR, which uses only single site spectral data to predict the corresponding label, with the selected 5% of total features and compare it with the SCRF and D-SCRF. It can be noted from table 2 that the classification accuracies of both the SCRF and D-SCRF are much higher than the 87.87% accuracy of D-SMLR. This comparison demonstrates the importance of contextual information for the hyperspectral image classification.

Table 2. Classification Accuracies of D-SMLR, SCRF and D-SCRF

class	D-SMLR	SCRF	D-SCRF
corn-no till	81.91	85.97	90.26
grass/trees	97.74	98.77	98.36
hay-windrowed	98.06	99.05	99.68
soybeans-no till	71.34	77.39	84.55
woods	97.28	98.57	98.81
overall accuracy	87.87	90.85	93.56

5 Conclusion

In this work, we investigated the D-SCRF on the feature selection in the contextual classification of hyperspectral data and developed a dynamic learning framework to relieve heavy degeneration of classification performance caused by

over-sparsity in SCRF and the characteristics of the hyperspectral data. The results on real-world hyperspectral data validate the efficiency and effectiveness of the D-SCRF. The experimental results of current form also indicate several future works. We developed the dynamic training framework to use the varied hyperparameters and thus can relieve the heavy degeneration of classification performance. But the optimality of the varied hyperparameters is difficult to be investigated. In the future, we hope to develop the methods to select the optimal hyperparameters, or to use the adaptive sparseness methods to avoid the adjusting or estimating of the hyperparameters[16].

References

1. Chi, M., Bruzzone, L.: Semisupervised Classification of Hyperspectral Images by SVMs Optimized in the Primal. IEEE Trans. Geosci. Remote Sens. 45, 1870–1880 (2007)
2. Muñoz-Marí, Bruzzone, L., Camps-Valls, G.: A Support Vector Domain Description Approach to Supervised Classification of Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 45, 2683–2692 (2007)
3. Ashish, D., McClendon, R.W., Hoogenboom, G.: Land-use classification of multispectral aerial images using artificial neural networks. Int. Jour. Remote Sens. 30, 1989–2004 (2009)
4. Camps-Valls, G., Marsheva, T.V.B., Zhou, D.: Semi-Supervised Graph-Based Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 45, 3044–3054 (2007)
5. Zhong, P., Wang, R.: Modeling and Classifying Hyperspectral Imagery by CRFs with Sparse Higher Order Potentials. IEEE Trans. Geosci. Remote Sens. 49, 688–705 (2011)
6. Zhong, P., Wang, R.: Learning conditional random fields for classification of hyperspectral images. IEEE Trans. Image Process. 19, 1890–1907 (2010)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, pp. 282–289 (2001)
8. Zhong, P., Wang, R.: A multiple Conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Trans. Geosci. Remote Sens. 45, 3978–3988 (2007)
9. Zhong, P., Wang, R.: Learning Sparse CRFs for Feature Selection and Classification of Hyperspectral Imagery. IEEE Trans. Geosci. Remote Sens. 46, 4186–4197 (2008)
10. Kumar, S.: Models for learning spatial interactions in natural images for context-based classification. PhD thesis. Carnegie Mellon University (2005)
11. Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Machine Intell. 27, 957–968 (2005)
12. Ng, A.Y.: Feature selection, L1 vs. L2 regularization, and rotational invariance. In: International Conference on Machine Learning (2004)
13. Zhong, P., Zhang, P., Wang, R.: Dynamic learning of sparse multinomial logistic regression for feature selection and classification of hyperspectral data. IEEE Geosci. Remote Sens. Lett. 5, 280–284 (2008)

14. Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. Jour. Comp. Vision. 81, 2–23 (2009)
15. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley, Hoboken (2003)
16. Figueiredo, M.A.T.: Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Machine Intell. 25, 1150–1159 (2003)