FEATURE EXTRACTION AND SELECTION HYBRID ALGORITHM FOR HYPERSPECTRAL IMAGERY CLASSIFICATION

Sen Jia[a], Yuntao Qian[b], Jiming Li[b], Weixiang Liu[a] and Zhen Ji[a]

[a]Texas Instruments DSPs Lab, College of Computer Science and Software Engineering
Shenzhen University Shenzhen, China - (senjia, wxliu, jizhen)@szu.edu.cn
[b]College of Computer Science, Zhejiang University, Hangzhou, China - (ytqian, ljming)@zju.edu.cn

1. INTRODUCTION

Hyperspectral sensors collect multichannel contiguous narrow spectral band imagery, spanning from the visible to the infrared portion of the electromagnetic spectrum [1]. Due to the difficulty of constructing a standard spectral library and the unsatisfactoriness of spectral matching techniques, classification of hyperspectral images is usually addressed in feature space. Hence, dimensionality reduction (DR) is an essential part, which is generally adopted to preprocess the hyperspectral data.

There are two methods to reduce the dimension of hyperspectral data. One is feature extraction, which transforms the original data onto the destination feature space through projections like PCA or ICA. These projections can find out an optimal solution by implementing data variance or independence criteria. In the past decades, wavelet analysis has become a prominent feature extraction method in signal processing [2]. It is a multi-resolution approach, which decomposes a signal by projecting it onto a scaled and translated version of a prototype mother wavelet. Once a mother wavelet is selected, signal is transformed into approximate and detail coefficients by discrete wavelet transform (DWT). Since the approximate coefficients contain the most smoothed version of the signal, DWT can be a productive tool for hyperspectral feature extraction.

The other kind of method is feature or band selection [3]. Compared to feature extraction, feature selection methods identify absorption bands which is a subset of the original spectral bands that contains most of the characteristics. The advantage of feature selection is that it can preserve the relevant original information from the data [4]. Unsupervised feature selection can be considered as a data clustering problem. In the clustering-based feature selection approaches, each feature is considered as a data point, the pairwise similarity or correlation between two features is measured, and then the features are grouped into several clusters.

Recently, we have introduced a new clustering algorithm, named affinity propagation (AP), for hyperspectral feature selection [5]. AP is proposed by Frey and Dueck [6]. It initially considers all data points as potential cluster exemplars, and then exchanges messages between data points until a stable state is reached. Clusters are formed by assigning each data point to its most similar exemplar. The advantages of AP over the k-means and hierarchical clustering-based feature selection methods are a) it identifies the representative of each cluster from the original data set instead generates a new centroid or mean point. b) AP is not sensitive to the initial selection of exemplars because it consider all data points as potential exemplars. c) the parameter of "preference" controls the possibility of a data point being exemplar, which also governs the number of clusters.

In this paper, the feature extraction and selection algorithms are combined together to obtain the representative features of hyperspectral data. Firstly, DWT-based feature extraction is applied to spectral signatures to acquire the wavelet coefficients on

This work was supported by National Natural Science Foundation of China (60902070 and 60903113) and Doctor Starting Project of Natural Science Foundation of Guangdong Province, China (9451806001002287).

different scales. Secondly, AP-based feature selection is utilized to choose the exemplars from the extracted features. Based on the reduced feature vectors, K-nearest neighborhood (KNN) classification algorithm [7] is adopted to examine the efficiency of the chosen exemplars. The proposed algorithm not only decreases the impact of noise in the first step, but also guarantees the representativeness of the obtained features in the second step, making the classification results more accurate. Experiments on real hyperspectral data verify the feasibility of the proposed algorithm.

The rest of this paper is organized as follows. Section 2 and Section 3 presents a brief overview of DWT and AP respectively. In Section 4, the experiment results on real hyperspectral data set are demonstrated. At last, Section 5 makes concluding remarks and suggests the future work.

2. DWT-BASED FEATURE EXTRACTION

The discrete wavelet transform of a signal is defined as an inner product of the signal and wavelet bases. The fine-scale and large-scale information of the signal can be simultaneously investigated by projecting it onto a set of wavelet bases with various scales. The decomposition can be repeated, with successive approximations being decomposed in turn, so that one signal is broken down into a number of components. Initially, the original signal is the input of the filters. Theoretically, the number of elements in the lower scale is half of the number of elements in the upper scale. Thus, the use of these linear wavelet features is also associated with a dimensionality reduction of signal.

Concerning hyperspectral imagery, it is a three-dimensional array with the width and length corresponding to spatial dimensions and the spectral bands to the third dimension, which are denoted by M, N and L in sequence. $\mathbf{R}$ is the image cube with each band $\mathbf{R}_l \in \mathbb{R}^{M \times N}$ being an image matrix. Generally, much of the noise is contained in the detail coefficients, so the approximation coefficients are used as the input of the second feature selection step.

3. AP-BASED FEATURE SELECTION

AP is known as a message-passing algorithm. Two kinds of message-passing procedures termed "responsibility" and "availability" are used to exchange message between each point i and each candidate exemplar k. Initially, the availabilities are set to zero. Then the responsibilities and availabilities are computed using the following rules

$$r(i,k) \leftarrow s(i,k) - \max_{k' \mathrm{s.t.} k' \neq k} \{a(i,k') + s(i,k')\} \tag{1}$$

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i' \mathrm{s.t.} i' \notin \{i,k\}} \max\{0, r(i',k)\} \right\} \tag{2}$$

In addition, the self-availability $a(k,k)$ is updated differently

$$a(k,k) \leftarrow \sum_{i' \mathrm{s.t.} i' \neq k} \max\{0, r(i',k)\} \tag{3}$$

The simple update rules in (1) and (2) will often lead to oscillations caused by "overshooting" the solution, so the two messages are "dumped" as follows:

$$\mathrm{M}_{new} = \lambda \times \mathrm{M}_{old} + (1-\lambda) \times \mathrm{M}_{new} \tag{4}$$

where M_{new} and M_{old} are respectively the message values from the previous and current iteration, and the damping factor λ is between 0 and 1. In all of our experiments, we use a default damping factor of $\lambda = 0.5$.

After obtaining the extracted features by DWT in the first step, feature selection is applied using the AP algorithm. AP starts with the construction of a similarity matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$, in which the element $\mathbf{S}(i,k)(i \neq k)$ measures how well the feature k

represents feature i. A common choice for similarity is negative Euclidean distance:

$$\mathbf{S}(i,k) = -\|\mathbf{R}_i - \mathbf{R}_k\|^2 = -\sum_{m,n=1}^{M,N} (\mathbf{R}_{mni} - \mathbf{R}_{mnk})^2 \tag{5}$$

As for $\mathbf{S}(k,k)$, it has its particular meaning in AP, which reflects the *a priori* suitability of feature k to serve as an exemplar, which is referred to as "preference". Instead of fixing the number of exemplars in advance, the preferences are used to control how many features are selected as exemplars. In most cases, because no prior inclination is available for particular features to be the exemplars, the preferences of all features are set to the same value.

In view of classification, the prior inclination for a particular feature k to be the exemplar, represents the effect of feature k as an individual one in classification, named the discriminative capability. In the unsupervised learning case, the feature with greater deviation from its associated Gaussian distribution has stronger discriminative capability. This is a quantitative measure of nongaussianity of the feature, which can be estimated by kurtosis [8].

$$kurt(t) = E\{t^4\} - 3(E\{t^2\})^2 \tag{6}$$

Kurtosis is zero for a Gaussian random variable. For most (but not quite all) nongaussian random variables, kurtosis is nonzero, and can be both positive or negative. Therefore, the square of kurtosis is used to compute $\mathbf{S}(k,k)$

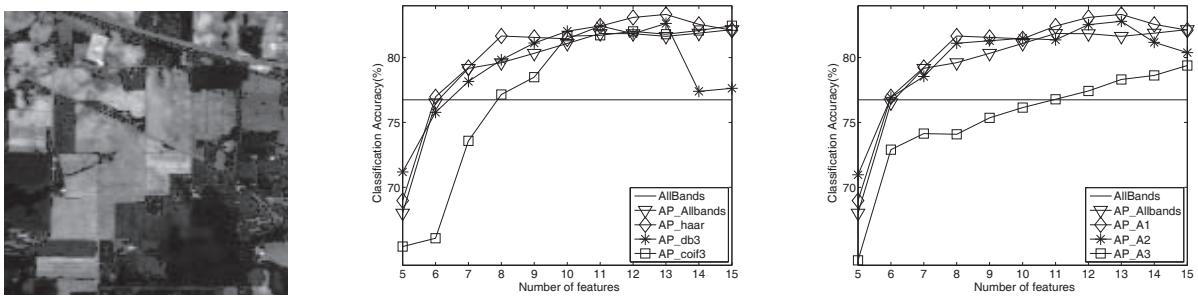
$$\mathbf{S}(k,k) = \alpha(kurt(t_k))^2 \tag{7}$$

where α is a negative value for controlling the number of exemplars. In short, AP not only considers the the discriminative capability of each individual feature through $\mathbf{S}(k,k)$, but the correlation/similarity among features as well through $\mathbf{S}(i,k)$, so that the exemplars are those features that have both of higher discriminative capability and less correlation.

4. EXPERIMENTAL RESULTS

To examine the performance of the proposed classification system, different wavelet bases and the approximation coefficients at different scales are used as input of the second feature selection step on real hyperspectral remote sensing data set, which is a section of a scene taken over northwest Indiana's Indian Pines (145×145 pixels, 220 bands) by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor in 1992 [9]. Figure 1(a) shows the 25th band of the data. In the experiment, 200 out of the 220 bands are used, discarding the lower signal-to-noise (SNR) bands (104-108, 150-163, 220). Meanwhile, KNN classifier is adopted. The class of a new sample is determined by the labels of k training data points that are nearest this sample. In our experiments, the number of neighbors k in KNN is set to be 3.

Figure 1(b) shows the classification results using wavelet approximation coefficients at scale 1 (noted as A1) extracted from several different mother wavelets, including Haar, Daubechies 3 (db3) and Coiflet 3 (coif3). The curves with legend "AllBands" and "AP_AllBands" represent the classification accuracy using all bands and AP-chosen bands, respectively. It can be seen that the results with different mother wavelets are comparable. Moreover, the accuracies with Haar wavelet are more accurate than the other methods in most cases. Therefore, the results using approximation coefficients at different scales with Haar wavelet are displayed in Figure 1(c). For the four schemes, the classification rates basically improve along with the increase of the number of features. Likewise, the results using A1 give the best classification accuracy in most instances, indicating the necessity of introducing the DWT-based feature extraction.



(a) (b) (c)

Fig. 1. a) Indian pines AVIRIS dataset, band 25, b) classification accuracy using wavelet approximation coefficients at scale 1 with different mother wavelets, c) classification accuracy using approximation coefficients at different scales with Haar wavelet.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we present a feature extraction and selection hybrid algorithm for classification of hyperspectral data. Discrete wavelet transform and affinity propagation are respectively utilized to accomplish the two tasks. On the one hand, DWT-based feature extraction eliminates the noise contained in the data; on the other hand, AP-based feature selection has several advantages over other feature selection methods. A comparison using different mother wavelets and the approximation coefficients at different scales are conducted. Experimentations achieved on real hyperspectral data show that the proposed hybrid classification approach has better performance than that without feature extraction, indicating the efficiency of the hybrid algorithm.

As future work, the detail coefficients should be considered. We only use the approximation coefficients in this study for simplicity. In fact, the detail coefficients may contain some important information for the targets that could not be classified by the approximation coefficients alone. Hence, cascade combination technique should be applied for further research [10].

6. REFERENCES

[1] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*, Hoboken, NJ: Wiley, 2004.

[2] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 7, pp. 674–693, July 1989.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] C.-I Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[5] S. Jia, Y. Qian, and Z. Ji, "Band selection for hyperspectral imagery using affinity propagation," in *Proc. DICTA '08.Digital Image Computing: Techniques and Applications*, 1-3 Dec. 2008, pp. 137–141.

[6] J. F. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, Feb. 2007.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[8] J.-F. Cardoso, "Dependence, correlation and gaussianity in independent component analysis," *Journal of Machine Learning Research*, vol. 4, pp. 1177–1203, 2003.

[9] "AVIRIS NW indiana's indian pines 1992 data set [online]," Available: ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C (original files) and ftp://ftp.ecn.purdue.edu/pub/biehl/PC-MultiSpec/ThyFiles.zip (ground truth).

[10] J. Gama, *Combining classification algorithms*, Ph.D. thesis, University of Porto, 2000.