Noise-adjusted Sparsity-Preserving-based Dimensionality Reduction for Hyperspectral Image Classification

Nam Ly, Qian Du, James E. Fowler
Department of Electrical and Computer Engineering, Mississippi state University,USA
nhl12@msstate.edu, du, fowler@ece.msstate.edu

Abstract

In this paper, we investigate the performance of a sparsity-preserving graph embedding based approach, called l^1 graph, in hyperspectral image dimensionality reduction (DR), and propose noise-adjusted sparsity-preserving(NASP) based DR when training samples are unavailable. In conjunction with the state-of-the-art hyperspectral image classifier, support vector machine with composite kernels (SVM-CK), the experimental study show that NASP can significantly improve the classification accuracy, compared to other widely used DR methods.

1 Introduction

Hyperspectral imaging (HSI) is a relatively recent technology in which the airbone remote sensors capture the reflected energy in hundreds to thousands narrow spectral bands in each spatial location in the image scene. It is well known that the high data dimensionality results in the problem of *curse of dimensionality* (or the Hughes phenomenom). Dimensionality reduction (DR), therefore plays a critical step in most of the HSI analysis, especially in classifiation task when the number of labeled training samples is limited. Commonly used DR techniques include unsupervised approaches, such as principal component analysis (PCA), noise-adjusted principal component analysis (NAPCA) [7], as well as supervised approaches, such as Linear Discriminant Analysis (LDA) [1].

Recently, Yan *et al.* [11] proposed a general approach known as graph embedding to unify all of the linear DR algorithms within a common framework. In graph embedding, graph construction becomes critical: an appropriate graph provides a high level of dimensionality reduction and preserves the important information, such as anomalous pixels, manifold,and multimodal structures. Two popular graph structures: $k-$nearest neighbor and $\epsilon-$radius ball [6] connect points within its k-nearest samples or surrounding ϵ ball, respectively, the graph-weight then calculated based on pairwise Euclidean distance [2]. It is simple to construct those graphs but the parameters-the size of sample neighborhood k and the radius of the ball ϵ- are very sensitive to noise and difficult to be identified in real world applications. Departing from traditional graph construction, Cheng *et al.* [4] exploits the merits of sparse representation to construct a novel graph, called l^1-graph, which inherits many advantages of sparse reconstruction and creates an adaptive and nonparametric graph. Motivated by the promising results of Sparsity Preserving projection method (SP) derived from l^1-graph [4], we propose an unsupervised dimensionality reduction algorithm, Noise Adjusted Sparsity Preserving (NASP), to reduce the dimension of HSI before employing an classifier such as Composite Kernels-based Support Vector Machine (SVM-CK). Unlike PCA or NAPCA, NASP is designed to preserve the underlying multimodal structure and the sparsity of non-Gaussian class distribution in the projected domain. Compared to the original SP-based DR method, NASP is expected to better preserve the underlying data structure after noise is better taken care of.

The remainder of the discussion is organized as follows. Section 2 describes the graph-embedding and proposed DR algorithm: NASP. Section 3 presents the SVM-CK-based HSI classifier, and NASP-SVM-CK in conjunction with the NASP-based DR process. Section 4 reports real-data experimental results. Finally, several concluding remarks are made in Sec. 5.

*This material is based upon work supported by the National Science Foundation under Grant No. CCF-0915307

2 Graph Embedding and NASP

Let a hyperspectral data matrix be represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]$ and $\mathbf{x}_i \in \mathbb{R}^N$, where M is the number of pixels and N is the number of bands. From the point of view of graph embedding, let G=$\{\mathbf{X}, \mathbf{W}\}$ be an undirected weighted graph with vertex set $\mathbf{X}$ and similarity (weight) matrix $W \in \mathbb{R}^{M\times M}$ where W_{ij} measures the similarity between two vertices, $\mathbf{x}_i$ and $\mathbf{x}_j$. The diagonal matrix D and the Laplacian matrix L of a graph G are defined as: $L = D - W$, $D_{ii} = \sum_{j\neq i} W_{ij}$, $\forall i$. In case of linear projection, the reduced dimensional feature space is derived by linear projection: $\mathbf{y} = P^T\mathbf{x}$, where P is the unitary projection matrix, then the objective function becomes:

$$P^* = \arg \min_{\substack{P^T X B X^T P = q \\ \text{or } P^T P = q}} \sum_{i\neq j} \|P^T\mathbf{x}_i - P^T\mathbf{x}_j\|^2 W_{ij} \tag{1}$$

where q is a constant and B is a penalty graph. In light of linearization of graph embedding, PCA has intrinsic graph connecting all of the data pairs with equal weiths and constrained by scale normalization on the projection vector ($W_{ij} = 1/M, i \neq j; B = I$).The quality of the aforementioned DR techniques, therefore, completely relies on the construction of intrinsic graph. However, the two existed famous graphs: k-nearest-neighbor method, and ϵ ball method have some limitations:($\mathbf{1}$)Sensitive to data noise: Due to the graph-weight construction (Gaussian-kernel [2] or l^2-reconstruction [9]) is found on pair-wise Euclidean distance, they are very sensitive to noise.($\mathbf{2}$) Datum-non adaptive neighborhood: to determine the neighborhood of each sample, both methods use fixed global parameter: k or radius-ϵ, and hence fail to offer datum-adaptive neighboorhood. To overcome those shortcommings, the l^1-graph [4] has been proposed by utilizing the recent advances in sparse coding [5].

An l^1-graph [4] exploits the sparse representation (SR) of each pixel in term of the rest ones in the training data set. In particular, l^1-graph uncovers the underlying sparse reconstruction relationship of each pixel, and it is desirable to preserve these reconstruction relationships in the reduced-dimensional-feature space. Based on the reconstruction of l^1-graph, the proposed DR algorithm, NASP, is to suppress the noise effect in data structure, which can be performed with two steps: the first step conducts noise-whitening to the original data, and the second step calculates P via (2). Note that in the low-dimensional space, the reconstruction capability is measured by l^2 norm instead by l^1 norm for computational efficiency. Derived from (1),the transformation matrix P can be found via optimization:

$$\begin{aligned} P^* &= \min_{P^T X X^T P = I} \sum_{i\neq j} \|P^T\mathbf{x}_i - P^T\mathbf{x}_j\|^2 W_{ij} \\ &= \min_{P^T X X^T P = I} \operatorname{tr}(P^T X L X^T P) \\ &= \min_P \frac{\operatorname{tr}(P^T X L X^T P)}{\operatorname{tr}(P^T X X^T P)} \end{aligned} \tag{2}$$

where W_{ij} is determined by the reconstruction of l^1-graph and penalty graph $B = I$. This $trace - ratio$ optimization problem is easily solved by the generalized eigenvalue problem as: $XLX^T\mathbf{p}_{k+1-j} = \lambda_j XX^T\mathbf{p}_{k+1-j}$ where $\mathbf{p}_{k+1-j}$ is the eigenvector corresponding to the jth largest eigenvalue λ_j as well as the $(k+1-j)$th column vector of the transform matrix P. The construction of l^1-graph represents four advantages: ($\mathbf{1}$) robust to noise: thanks to the overall SR instead of conventionally pairwise Euclidean distance and the noise-whitenning step, l^1-graph is more robust to noise in many pattern recognition tasks.($\mathbf{2}$) sparsity: recent research on manifold learning [2] shows that the sparse graph conveys valuable information for classification purpose; the sparsity of l^1-graph is automatically determined instead of manually identified as in k-nearest-neighbor and ϵ-ball method. l^1-graph is therefore a non-parametric method.($\mathbf{3}$) datum-adaptive neighborhood: the number of neighbors defined by l^1-graph is adaptive to each sample, which is valuable for aplications with unevenly distributed data, e.g, hyperspectral image with non-homogenous regions. ($\mathbf{4}$) unified construction: conventional graph construction process is typically divided in two steps: the graph adjacency structure and the graph-weight construction; however, those two steps are solved simultaneously in l^1-graph.

It is hence expected that SP and NASP will surpass PCA and NAPCA as a DR projection when data are severely non-Gaussian, which is a prevalent situation in real world applications. In the next section, we will show that the combination of NASP and SVM-CK outperforms the existing DR methods for $spectral - spatial$ HSI classification.

3 Spectral-Spatial Hyperspectral Image Classification Scheme

Neighboring hyperspectral pixels usually belong to the same class because their spectral signatures are highly correlated. A mechanism to incorporate the *contextual* information into *spectral* information can signficantly improve the classification accuracy. In this section, we first introduce the famous $spectral - spatial$

classifiers in the literature: SVM-CK, then outline the proposed algorithms: NASP-SVM-CK.

Support Vector Machine (SVM) and SVM-CK were fully described in [3, 10] as state-of-the-art kernel-based classification techniques. Note that one usually works with the transformed input data, rather than the original input space samples, $\mathbf{x}_i$. A full family of composite kernels for the combination of spectral kernel (K_ω) and spatial (K_s) kernel was described in [3]. Notice that in [3], the authors: (a) Define the spatial features ($\mathbf{x}_m^s$) be the average of the reflectance values in a given window around the pixel $\mathbf{x}_i$ for each band, and let the spectral features $\mathbf{x}_i^\omega$ be the actual spectral signature ($\mathbf{x}_i = \mathbf{x}_i^\omega$). (b) Using polynomial kernel for spectral features and RBF kernel for the spatial features. In this paper, we use weighted summation kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mu K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + (1-\mu)K_\omega(\mathbf{x}_i^\omega, \mathbf{x}_j^\omega)$ with one modification: polynomial kernel is used for spatial features and RBF kernel is used for spectral features. We found that this new composite-kernel gives better result than original one: from 2 to $3(\%)$ higher in overall accuracy in our experiments.

It has been shown in the literature that NAPCA can dramatically outperform PCA as a preprocessing tool in preserving useful information in the original data . We therefore argue that NASP can serve as a better DR method than SP for SVM-CK. In this paper, NASP-SVM-CK, is proposed to solve $spectral-spatial$ HSI classification.

NASP-SVM-CK
algorithm:

1)Conduct noise-whitening to the original data
2)Construct the l^1-graph and graph weight $\mathbf{W}$
3)Derive the projection matrix P from (2)
4)Project training data $\mathbf{Y} = P^T\mathbf{X}$
5)Project test data $\mathbf{E} = P^T\mathbf{A}$
6)Calculate K_s, K_w, and K kernels
7)Class labels $z = \mathrm{SVM}(K, \varsigma, \gamma, \mu)$

In the following section, we compare our proposed algorithm, NASP-SVM-CK with SP-SVM-CK, SVM-CK, PCA-SVM-CK, NAPCA-SVM-CK, and MLR*sub*MLL, [8], a recently proposed supervised $spectral-spatial$ classification.

4 Experimental Results

In this section, we demonstrate the effectiveness of the proposed algorithm on real hyperspectral image,the popular AVIRIS Indian Pines image. The one-against-one strategy is employed for C-class classification using SVM and SVM-CK. To perform SVM, we use the popular tookkit libSVM [1]. The AVIRIS sensor generates 220 bands across the spectral range from 0.2 to 2.4 μm; however, in the experiment, we remove 18 water absorbtion bands. This image has spatial resolution of 20 m per pixel and spatial dimension 145×145. It contains 16 ground-truth classes, for each of the 16 classes, we randomly choose 10% of the labeled samples for training and the rest 90% for testing according to $fair$ and $unfair$ strategies. The parameters of SVM and SVM-CK $(\varsigma, \gamma, \mu) = (256, 0.3536, 0.7)$ are obtained by $ten-$fold cross validation and the parameters of MLR*sub*MLL are derived from [8]. The effects of reduced dimensions on the overall accuracy are shown in Fig. 1. In this figure, SVM-CK is used as a comparison baseline, the proposed algorithms: NASP-SVM-CK is compared to PCA-SVM-CK. The results show that NASP-SVM-CK gave the best overall accuracy at the low number of dimensions.

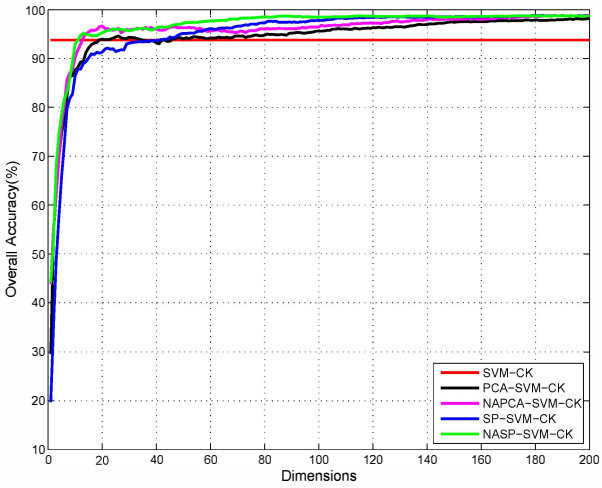

Figure 1: Effects of the number of dimensions in the transformed domain

From Fig. 1, we choose $d = 80$ which fairly represents the classification performance of each algorithm, the classification accuracy for each class, the overall accuracy, average accuracy, and the κ coefficient are shown in Table 1 using different clasisfiers on the test set. The overall accuracy is computed by the ratio between correctly classified test samples and the total number of test samples, and the average accuracy is the mean of the 16 True Positive Rates (TPR), and the Cohen-κ coefficient is computed by weighting the measure accuracies which shows a robust measure of the degree of agreement. In most cases, the proposed NASP-SVM-CK outperforms

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 1: CLASSIFICATION ACCURACY (%) FOR THE INDIAN PINES IMAGE ON THE TEST SET

Accuracy	SVM	MLR*sub*	MLR*sub*MLL	SVM-CK	PCA	NAPCA	SP	NASP
Overall	82.61	80.23	89.98	93.76	94.66	96.43	95.71	**97.73**
Average	84.78	70.43	73.15	92.86	95.47	96.69	97.18	**97.89**
κ	0.8	0.77	0.88	0.92	0.93	0.96	0.95	**0.97**

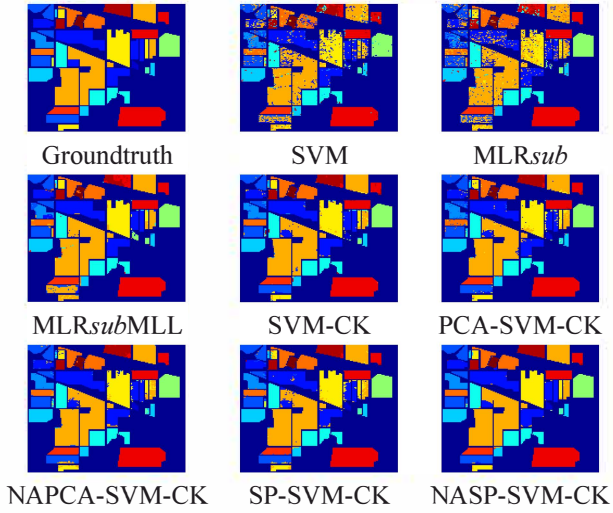

Figure 2: Groundtruth and classification maps of the Indian Pines image

the original SVM-CK, MLR*sub*MLL, PCA-SVM-CK, NAPCA-SVM-CK, and SP-SVM-CK. Overall, NASP-SVM-CK provides the best performance especially in the extreme case, e.g., class 1, 7, and 9 which has only 5, 3, and 2 training samples. This is very important in hyperspectral image analysis due to costly training data collection. The classification maps on labeled pixels obtained from the various algorithms are shown in Fig. 2. One can see that by incorporating the contextual information, the NASP-SVM-CK algorithm provides a much smoother classification map than the other methods.

5 Conclusions

DR has been widely used as a preprocessing step for hyperspectral image analysis. In this paper, we investigate DR under the unified framework of graph embedding. By preserving sparsity property in the original data, the l^1 graph method can outperform other widely used methods, such as PCA and LDA. In this paper, we propose the NASP based DR method to suppress the noise effect. In conjunction with the state-of-the-art hyperspectral image classifier, SVM-CK, where RBF kernel is used in the spectral dimension and polynomial kernel for the spatial dimension, the real data experiment shows that NASP can significantly improve the classification accuracy, compared to SP and other widely used DR methods.

References

[1] T. Bandos, L. Bruzzone, and G. Camps-Valls. Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3):862 –873, March 2009.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[3] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 3(1):93 – 97, January 2006.

[4] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang. Learning with l^1-graph for image analysis. *IEEE Transactions on Image Processing*, 19(4):858 –866, April 2010.

[5] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math*, 59:797–829, 2004.

[6] D. Eppstein, M. S. Paterson, and F. F. Yao. On nearest neighbor graphs. *Discrete & Computational Geometry*, 17(3):263–282, April 1997.

[7] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[8] J. Li, J. Bioucas-Dias, and A. Plaza. Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):809 –823, March 2012.

[9] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[10] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[11] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40 –51, January 2007.