A MULTI-MODAL PATTERN CLASSIFICATION FRAMEWORK FOR HYPERSPECTRAL IMAGE ANALYSIS

Wei Li, Saurabh Prasad, James E. Fowler, Lori M. Bruce

Electrical and Computer Engineering Department and Geosystems Research Institute,
Mississippi State University, Mississippi State, MS 39762

ABSTRACT

Dimensionality reduction is a crucial preprocessing step for effective analysis of high dimensional hyperspectral imagery (HSI). Currently popular dimensionality reduction techniques (such as Principal Component Analysis, Linear Discriminant Analysis and their many variants) assume that the data are Gaussian distributed. The quadratic maximum likelihood classifier commonly employed for HSI analysis also assumes Gaussian class-conditional distributions. In this paper, we propose a classification paradigm that is designed to exploit the rich statistical structure of hyperspectral data. It does not make the Gaussian assumption, and performs effective dimensionality reduction and classification of highly non-Gaussian, multi-modal HSI data. The framework employs Local Fisher's Discriminant Analysis (LFDA) to reduce the dimensionality of the data while preserving its multi-modal structure. This is followed by a Gaussian Mixture Model (GMM) classifier for effective classification of the reduced dimensional multi-modal data. Experimental results on a multi-class HSI classification task show that the proposed approach significantly outperforms conventional approaches.

***Index Terms*—** Hyperspectral data, Dimensionality reduction, Gaussian mixture model

1. INTRODUCTION

Statistical pattern classification systems for the analysis of Hyperspectral Imagery (HSI) typically employ a dimensionality reduction algorithm before performing the classification task. HSI data typically has hundreds (even thousands) of spectral bands which are oftentimes highly correlated. Dimensionality reduction seeks to decrease the computational complexity and ameliorate statistical ill-conditioning by discarding redundant features that can potentially deteriorate classification performance [1].

Popular dimensionality reduction techniques include unsupervised approaches, such as Principal Components Analysis (PCA) and Independent Component Analysis (ICA), and supervised approaches, such as Fisher's Linear Discriminant Analysis (LDA) [2]. There are numerous variants of these techniques. However, these are sub-optimal at best for a general pattern classification problem. In [3], the authors describe the limitations and shortcomings of PCA for hyperspectral classification theoretically and experimentally. A key limitation with techniques such as PCA, LDA, Quadratic Maximum-likelihood classification (MLE) and their variants is that they assume that the classes are Gaussian distributed [4], [5]. However, in many practical situations, most of the observational data are non-Gaussian, and in extreme cases, even severely multi-modal. PCA and LDA are expected to fail as dimensionality reduction techniques under such conditions.

In this paper, we propose a classification paradigm that is designed to exploit the rich statistical structure of the (very high dimensional) hyperspectral imagery. It does not make the Gaussian assumption, and performs effective dimensionality reduction and classification of highly non-Gaussian, multi-modal HSI data. The framework employs Local Fisher's Discriminant Analysis (LFDA) [6] to reduce the dimensionality of the data while preserving its multi-modal structure. Unlike LDA, LFDA is designed to handle multi-modal, non-Gaussian class distributions. This is followed by a Gaussian Mixture Model (GMM) classifier [7] for effective classification of the reduced dimensional multi-modal data.

The outline of this paper is as follows. In section 2, we describe the LFDA algorithm. In section 3, we describe the GMM algorithm. In section 4, we describe the HSI data employed in this work to measure the efficacy of the proposed LFDA-GMM classification strategy. In section 5, we present results comparing the proposed strategy with currently popular HSI classification techniques and provide concluding remarks.

2. LOCAL FISHER'S DISCRIMINANT ANALYSIS

Local Fisher's Discriminant Analysis (LFDA) [6] has been recently proposed as an extension of LDA which does not restrict the class distributions to uni-modal Gaussian. LFDA combines the properties of LDA and Locality Preserving

978-1-4577-2201-1/11/$26.00 ©2011 IEEE

Projection (LPP). LPP is an unsupervised classification technique that seeks to find a transformation preserving the neighborhood relationships of the data in the embedding space. In other words, nearby points in the original space are kept close in the LPP embedded space. In [6], the authors provide a detailed description of LPP and LFDA. LFDA is a transformation that exploits the advantages of both LDA and LPP, obtaining good between-class separation while preserving the within-class local structure at the same time. In doing so, it is able to perform well even when the data is not distributed normally. It is hence expected that LFDA will surpass LDA and LPP as a dimensionality reduction projection when the data is multi-modal, or significantly non-Gaussian.

As in LPP, LFDA preserves neighborhood relationships in the embedding by employing an "affinity" matrix that is described below. Consider a dataset with training samples $\{x_i\}_{i=1}^{n}$ and class labels $y_i \in \{1,2,\cdots,c\}$, where c is the number of classes, n is the total number of training samples. Let n_l be the number of available training samples for l'th class, and $\sum_{l=1}^{c} n_l = n$. Define $A_{i,j}(\in[0,1])$ as the "affinity" between x_i and x_j

$$A_{i,j} = \exp(-\frac{\left\|x_i - x_j\right\|^2}{\sigma_i \sigma_j}) \quad , \tag{1}$$

where $\sigma_i = \left\|x_i - x_i^{(m)}\right\|$ denotes the local scaling of data x_i, and $x_i^{(m)}$ is the m'th nearest neighbor of x_i. $A_{i,j}$ is then a symmetric matrix (referred to as the affinity matrix) of size $n \times n$, which measures the distance among data samples in a given region. Similar to LDA, the "local" between-class $S^{(lb)}$ and within-class $S^{(lw)}$ scatter matrices are defined as

$$S^{(lb)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(lb)}(x_i - x_j)(x_i - x_j)^T \ , \tag{2}$$

$$S^{(lw)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(lw)}(x_i - x_j)(x_i - x_j)^T \ , \tag{3}$$

where $W^{(lb)}$ and $W^{(lw)}$ are $n \times n$ matrices defined as

$$W_{i,j}^{(lb)} = \begin{cases} A_{i,j}(1/n - 1/n_l) & if\ y_i = y_j = l, \\ 1/n & if\ y_i \neq y_j \end{cases} \tag{4}$$

$$W_{i,j}^{(lw)} = \begin{cases} A_{i,j}/n_l & if\ y_i = y_j = l \\ 0 & if\ y_i \neq y_j \end{cases} \quad . \tag{5}$$

The solution that maximizes the Fisher's ratio within the context of the local scatter matrices is given by

$$S^{(lb)}\varphi = \lambda S^{(lw)}\varphi \, , \tag{6}$$

where λ is the diagonal eigenvalue matrix and φ is the eigenvector matrix.

LFDA can be viewed as a localized variant of LDA because it does not impose far-apart data pairs of the same class to be close in the projection. Due to the weaker constraint, LFDA outperforms LDA in maximizing separability between different classes. It can be easily verified that when $A_{i,j} = 1$ for all *i* and *j*, LFDA degenerates to LDA.

In this paper, we propose that LFDA based dimensionality reduction is an ideal preprocessing for GMM based classification of extremely high dimensional HSI data. The resulting classification is expected to accurately capture the class conditional statistics in a reduced dimensional subspace, especially when classes are multi-modal. It is important to note that owing to the immensely high dimensionality of HSI data, GMMs are not employed for such tasks. The key contribution of this paper is to demonstrate the efficacy and suitability of LFDA as a dimensionality reduction technique for high dimensional classification tasks using complex classifiers such as GMMs – we know of no other work that combines these two approaches.

3. GAUSSIAN MIXTURE MODEL

A Gaussian Mixture Model (GMM) [7] can be viewed as a combination of two or more normal Gaussian distributions. In a typical GMM representation strategy, a probability density function for $x \in \Re^d$ is written as the sum of K Gaussian components (modes)

$$p(x) = \sum_{k=1}^{K} \alpha_k N(x, \mu_k, \Sigma_k) \, , \tag{7}$$

where,

$$N(x, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_k\right|^{1/2}} \exp\{-\tfrac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\} \tag{8}$$

Here, K is the number of mixture components, and α_k, μ_k, Σ_k are the mixing weight, mean and covariance of the *k*'th component, which are expressed by the parameter vector $\Theta = \{\alpha_k, \mu_k, \Sigma_k\}$. Once the 'optimal' number of components, *K*, per GMM has been determined (using the Bayesian Information Criteria, BIC [8] in this work), the parameters for each class can be estimated by using the Expectation Maximization (EM) algorithm [7] – an iterative optimization strategy. Specifically, given a sample dataset $x_o = \{x_j\}_{j=1}^{n_l}$, n_l is the number of samples in *l*'th class, the resulting complete data log-likelihood is

$$L(\Theta, x_o) = \sum_{k=1}^{K}\sum_{j=1}^{n_l} p(k\left|x_j, \Theta\right.) \log[\hat{\alpha}_k N(x_j, \hat{\mu}_k, \hat{\Sigma}_k)] \ , \tag{9}$$

where $p(k\left|x_j, \Theta\right.)$ is the posterior probability for the *k*'th component of the GMM, and can be written as

$$p(k\left|x_j,\Theta\right.)=\frac{\alpha_k N(x_j,\mu_k,\Sigma_k)}{\sum\limits_{k=1}^{K}\alpha_k N(x_j,\mu_k,\Sigma_k)}. \tag{10}$$

At each iteration, the parameter set Θ is obtained by maximizing the likelihood function $L(\Theta,x_o)$ described in (9) (the *M-step*). The parameter set is then updated with an expected value of these parameters for the next iteration (the *E-step*)

$$\hat{\alpha}_k=\frac{1}{n_l}\sum_{j=1}^{n_l}p(k\left|x_j,\Theta\right.)\ , \tag{11}$$

$$\hat{\mu}_k=\frac{\sum\limits_{j=1}^{n_l}p(k\left|x_j,\Theta\right.)x_j}{\sum\limits_{j=1}^{n_l}p(k\left|x_j,\Theta\right.)}\ , \tag{12}$$

$$\hat{\Sigma}_k=\frac{\sum\limits_{j=1}^{n_l}p(k\left|x_j,\Theta\right.)(x_j-\mu_k)(x_j-\mu_k)^T}{\sum\limits_{j=1}^{n_l}p(k\left|x_j,\Theta\right.)}\ . \tag{13}$$

The above procedure is iterated until the relative difference between successive values of the complete data log-likelihood provided by (9) reaches a pre-determined convergence threshold.

The fundamental hesitation most researchers have when employing a technique such as GMMs for hyperspectral image analysis tasks is the impractical size of the resulting parameter space. Employing conventional linear dimensionality reduction techniques such as LDA as a pre-processing often destroys the underlying multi-modal structure of the data, rendering GMMs ineffective. Given these constraints, we propose and demonstrate that LFDA based dimensionality reduction is much better suited for GMMs.

4. EXPERMENTAL HYPERSPECTRAL DATA

The experimental hyperspectral imagery employed in this work to quantify the efficacy of the proposed approach covers the Washington DC Mall area [9]. The image represents an urban scenario with 307 × 1280 pixels and 191 bands (hence the dimensionality of the input feature space is 191) in the 0.4 to 2.4 μm region of the visible and infrared spectra acquired by the HYDICE sensor. In this work, five classes are extracted (Roofs, Streets and Paths, Grass, Trees, and Water). Figure 1 depicts the spectral signatures for these five classes extracted from the imagery. This dataset is partitioned into approximately 8000 training pixels and 14000 test pixels to validate and quantify the efficacy of the proposed system. Further, to avoid spatial biases, our training and test data is collected from spatially separate regions.

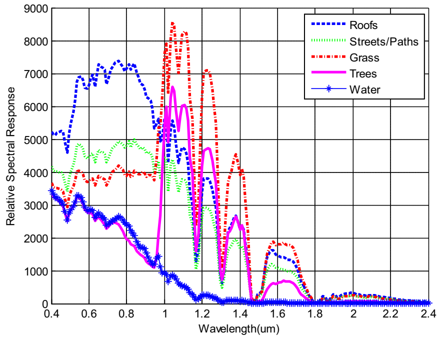

Fig. 1. Spectral signatures of the five classes that form the hyperspectral classification task employed in this work.

5. EXPERIMENTAL RESULTS AND CONCLUSIONS

In this section, we present experimental results demonstrating the benefits of LFDA based dimensionality reduction, followed by GMM based classification. We report the performance of classification systems as measured by the overall classification accuracy with test data, along with the 95% confidence intervals in those accuracies. To create very challenging operating conditions, we provide results for a wide range of pixel mixing conditions. In many real life situations, the spatial resolution may not be fine enough to resolve the object of interest, and inadvertent mixing between multiple classes may occur [10]. In this paper, we linearly mix signatures (pixels) from background classes with the signatures (pixels) of the class being classified. We report results over a wide range of target abundance values (i.e., the fraction of target abundance to background abundance). Target class simply refers to any class that is being considered for mixing. The background signatures used for mixing the target class are gathered (with uniform weights) from across all the other classes.

Fig. 2 (a) shows the overall accuracy for three techniques at different mixing rates. An abundance of 100% implies that the pure DC-mall dataset is employed without any pixel mixing. Results are provided using three parametric classification strategies: (1) LDA based dimensionality reduction, followed by a quadratic maximum-likelihood classification (LDA+MLE) – a very common classification technique for hyperspectral image analysis, (2) LDA based dimensionality reduction followed by a GMM classifier (LDA+GMM), and, (3) The proposed method - LFDA based dimensionality reduction followed by a GMM classifier (LFDA+GMM). Since the LFDA projection is not rank-limited to *c*-1 [6], we varied the dimension of the LFDA projected subspace and measured the corresponding training performance. We found that a dimensionality of 10 was sufficient for the hyperspectral dataset. The resulting GMM parameter space (in the 10-dimensional LFDA projected subspace) has an approximate dimensionality of 500 – a much more manageable number (by many orders of magnitude) compared to the dimensionality of the GMM parameter space when the input (191-dimensional) features are used. We can make the following important observations from these results. The method proposed in this paper – LFDA followed by GMM is indeed a very powerful strategy, outperforming conventional approaches significantly, even under adverse pixel mixing conditions. The performance of LDA+GMM is poorer than LFDA+GMM. Hence, LDA is not a very effective dimensionality reduction strategy for GMMs.

We also conducted an experiment where we varied the amount of training data and studied the sensitivity of the proposed method relative to conventional methods over a range of training data abundance. In many practical situations, the amount of training samples available is insufficient to estimate models (particularly complex ones, such as GMMs). Due to the effective dimensionality reduction provided by LFDA, the LFDA+GMM classifier is hence expected to perform well even when very little training data is used. Fig. 2 (b) depicts results from this experiment. We report the overall accuracy of the three classification systems as a function of the relative training data abundance. This abundance (on the *x*-axis) is expressed as a fraction of the dimensionality of the data. Hence, an abundance of 4N implies that the amount of training data used is four times the dimensionality of the feature space.

Note that the proposed LFDA+GMM system significantly outperforms other methods, even when the training data abundance is a mere 4N. Also note that with decreasing training data abundance, the overall accuracy for all systems decreases. However, the rate of decay in accuracy of the LFDA+ GMM is much slower compared to others. Even at very low training data abundances (e.g., 4N) the classification accuracy of the proposed method is impressive.

To conclude, we proposed a multi-modal parametric classification technique for robust classification of hyperspectral data. We tested the performance of the proposed system under two real-life adverse operating scenarios – under very limited training data, and under severe pixel mixing, and showed that the proposed approach outperforms conventional approaches to HSI classification significantly under both these conditions. Although we have not included results of GMM classification without any dimensionality reduction in this paper, we performed such classification as well and found that the overall classification performance of basic GMMs with HSI data is dismally low (for example, when all of the training data was used, the accuracy was a low 64%, and when the amount of training data was dropped to 5N, this dropped to 20% indicating that GMM training effectively broke down due to an immensely high dimensional parameter space). This further strongly demonstrates the benefits of the classification approach proposed and demonstrated in this paper.

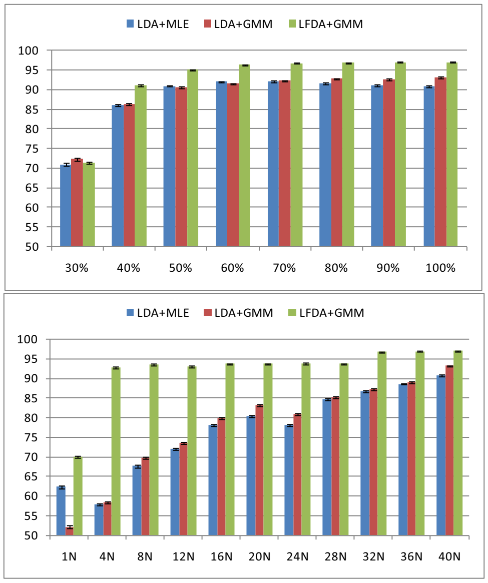

Fig. 2. (a) Overall Accuracy versus pixel mixing abundance, both expressed in percentage, for three different classification methods (*top*); (b) Accuracy versus different amount of training data (*bottom*).

6. REFERENCES

[1] S. Prasad and L. M. Bruce, “Decision fusion with confidence based weight assignment for hyperspectral target recognition,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448-1456, May 2008.

[2] M. D. Farrel and R. M. Mersereau, “On the impact of PCA dimension reduction for hyperspectral detection of difficult targets,” *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 192-195, Apr. 2005.

[3] S. Prasad and L. M. Bruce, “Limitations of principal components analysis for hyperspectral target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625-629, Oct. 2008.

[4] S. D. Zenzo, S. D. Degloria, R. Bernstein and H. C. Kolsky, “Gaussian maximum likelihood and contextual classification algorithm for multicrop classification,” *IEEE Trans. Geosci Remote Sens.*, vol. 25, pp. 805-814, Nov. 1987.

[5] A. M. Martinez and A. C. Kak, “PCA versus LDA,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228-233, Feb. 2001.

[6] Sugiyama, M., “Local fisher discriminant analysis for supervised dimensionality reduction,” *Proc. of the 23rd International Conf. on Machine Learning*, Pittsburgh, PA, 2006.

[7] A. Berge and A. H. S. Solberg, “Structured Gaussian components for hyperspectral image classification,” *IEEE Trans. Geosci Remote Sens.*, vol. 44, no. 11, pp. 3386-3396, Nov. 2006.

[8] G. Scharz, “Estimating the dimension of a model,” *the Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[9] D. Landgrebe, “Hyperspectral image data analysis as a high dimensional signal processing problem,” *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17-28, Jan. 2002.

[10] N. Dobigeon, S. Moussaoui, M. Coulon, J. Y. Tourneret and A. O. Hero, “Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery,” *IEEE Trans. Signal Processing*, vol. 57, no. 11, Nov. 2009.