

Locality-Preserving Discriminant Analysis in Kernel-Induced Feature Spaces for Hyperspectral Image Classification

Wei Li, Saurabh Prasad, James E. Fowler, and Lori Mann Bruce

Abstract—Linear discriminant analysis (LDA) has been widely applied for hyperspectral image (HSI) analysis as a popular method for feature extraction and dimensionality reduction. Linear methods such as LDA work well for unimodal Gaussian class-conditional distributions. However, when data samples between classes are nonlinearly separated in the input space, linear methods such as LDA are expected to fail. The kernel discriminant analysis (KDA) attempts to address this issue by mapping data in the input space onto a subspace such that Fisher's ratio in an intermediate (higher-dimensional) kernel-induced space is maximized. In recent studies with HSI data, KDA has been shown to outperform LDA, particularly when the data distributions are non-Gaussian and multimodal, such as when pixels represent target classes severely mixed with background classes. In this letter, a modified KDA algorithm, i.e., kernel local Fisher discriminant analysis (KLFDA), is studied for HSI analysis. Unlike KDA, KLFDA imposes an additional constraint on the mapping—it ensures that neighboring points in the input space stay close-by in the projected subspace and vice versa. Classification experiments with a challenging HSI task demonstrate that this approach outperforms current state-of-the-art HSI-classification methods.

Index Terms—Dimensionality reduction, feature space, hyperspectral imagery (HSI), kernel methods.

I. INTRODUCTION

HYPERSPECTRAL imagery (HSI) has hundreds (even thousands) of spectral bands that are oftentimes highly correlated. Dimensionality-reduction algorithms [1], [2] are typically designed to reduce the dimensionality of the feature space without losing desirable information. Conventional dimensionality-reduction techniques include unsupervised approaches such as principal component analysis (PCA) and independent component analysis, as well as supervised approaches, such as Fisher's linear discriminant analysis (LDA) [3], [4]. Numerous variants of these techniques exist.

A common characteristic of the approaches listed above is that, being linear methods, they are expected to be suboptimal

(and even entirely fail) for nonlinear classification tasks (i.e., when the data distributions are such that the resulting decision boundaries are highly nonlinear). Kernel methods attempt to address this problem. The central idea behind kernel-based methods is to map the input data onto an intermediate feature-induced space (potentially possessing a much higher dimensionality), such that complex nonlinear decision boundaries in the input space become simpler linear decision boundaries in the kernel-induced space. A variety of kernel-projection techniques, such as kernel discriminant analysis (KDA) and kernel PCA, have been studied for various classification tasks. In [5], the authors show that KDA can provide a significantly superior classification performance over LDA if the data in the input space possesses nonlinear class separation. KDA has been successfully employed for hyperspectral-data classification in [6]. In [7], Prasad and Bruce incorporated KDA within a multiclassifier and decision-fusion framework for HSI target recognition.

In this letter, we employ a kernel-based dimensionality reduction for HSI classification, i.e., kernel local Fisher discriminant analysis (KLFDA), which was recently proposed by Sugiyama in [8]. The algorithm combines locality-preserving projection (LPP) [9], which helps preserve local-neighborhood information, with discriminant analysis in a kernel-induced space. LPP preserves neighborhood relationships and forces neighboring points in the input space to remain close in the projected space. It is hence expected that incorporating LPP within the conventional KDA will be significantly beneficial, particularly when the input space has complicated class-conditional distributions.

In this letter, we propose an approach for the classification of HSI data, which employs KLFDA for dimensionality reduction, followed by a quadratic maximum-likelihood-estimation (MLE) classifier that assumes Gaussian class-conditional distributions [10] in a KLFDA-projected subspace. Our choice of the classifier is motivated by the observation [7] that class-conditional distributions after KDA projections are expected to be Gaussian. Classification experiments with a challenging HSI task demonstrate that this approach outperforms currently popular HSI-classification methods.

This letter is organized as follows: In Section II, we discuss conventional discriminant-analysis-based dimensionality-reduction techniques and provide a motivation for a KLFDA-based dimensionality reduction for hyperspectral classification. We also provide a description of the KLFDA algorithm and provide empirical evidence of its benefits with a synthetic data set. In Section III, we provide a description of the experimental

Manuscript received November 16, 2010; revised January 28, 2011 and February 21, 2011; accepted February 23, 2011. Date of publication May 10, 2011; date of current version August 26, 2011. This work was supported in part by the National Science Foundation under Grant CCF0915307 and in part by the National Geospatial-Intelligence Agency under Grant HM1582-10-1-0001.

The authors are with the Geosystems Research Institute and the Department of Electrical and Computer Engineering, Mississippi State University, MS 39762 USA (e-mail: saurabh.prasad@ieee.org).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/LGRS.2011.2128854

hyperspectral data set used to validate the proposed approach. In Section IV, we describe the experimental setup and show how to optimize the proposed system. We also compare the performance of the proposed system with conventional parametric supervised HSI-classification techniques. We demonstrate benefits and improvements with the proposed technique for a variety of real-life operating scenarios, such as severe pixel mixing and reduced training sample size. We summarize our results and provide concluding remarks in Section V.

II. DISCRIMINANT ANALYSIS

A. Linear Fisher Discriminant Analysis

LDA seeks to find a linear transformation φ such that the within-class scatter is minimized and the between-class scatter is maximized. The LDA solution is obtained by maximizing Fisher's ratio [1], i.e.,

$$J_1(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \quad (1)$$

where S_b is the between-class scatter matrix, and S_w is the within-class scatter matrix. The maximizing solution is obtained by solving the following generalized eigenvalue problem:

$$S_b \varphi = \Lambda S_w \varphi \quad (2)$$

where Λ is the diagonal eigenvalue matrix. The dimensionality of the projected subspace after an LDA transformation is upper bounded to $c - 1$ by design (c is the number of classes in the classification task). A detailed analysis of LDA and its variants can be found in [1].

B. KDA

The following is intended to be a brief overview of KDA for continuity. KDA seeks to find a projection w of vectors in a (higher-dimensional) kernel-induced space such that it maximizes Fisher's ratio in that space. For a given nonlinear mapping function Φ , the Fisher's ratio in the resulting kernel-induced space can be expressed as

$$J_2(w) = \frac{w^T S_b^\Phi w}{w^T S_w^\Phi w} \quad (3)$$

where S_b^Φ is the between-class scatter matrix, and S_w^Φ is the within-class scatter matrix in the space induced by the mapping function Φ .

The kernel-induced space is typically of a much higher dimension than the input space (it can potentially be infinite-dimensional), but Baudat and Anouar [11], by reformulating the discriminant-analysis problem in terms of inner products of vectors in the input space and by exploiting the kernel trick, proposed a computationally tractable algorithm for KDA. The "kernel trick" [12] allows for the computation of algorithms in a kernel-induced space without explicitly evaluating the mapping, as long as the algorithm can be expressed in terms of dot products of vectors in the input space.

C. KLFDA

Before introducing KLFDA, we first introduce local Fisher discriminant analysis (LFDA) [8], which combines the properties of the LDA and LPP [9]. LPP is a linear manifold-learning technique that seeks to find a linear map that preserves the local structure of neighboring samples in the input space. In other words, nearby points in the original space are kept close in the LPP-embedded space. LFDA preserves neighborhood relationships in the embedding by employing an "affinity" matrix that is defined below.

Consider a data set with training samples $\{x_i\}_{i=1}^n$ and class labels $\{y_i\}_{i=1}^n$, $y_i \in \{1, 2, \dots, c\}$, where c is the number of classes, and n is the total number of training samples. Let n_l be the number of training samples available for the 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 given by

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\gamma_i \gamma_j}\right) \quad (4)$$

where $\gamma_i = \|x_i - x_i^{(k)}\|$ denotes the local scaling of data x_i , and $x_i^{(k)}$ is the k 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 local distance between the data samples in the input space. Similar to LDA, the "local" between-class $S^{(\text{lb})}$ and within-class $S^{(\text{lw})}$ scatter matrices are defined as

$$S^{(\text{lb})} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(\text{lb})} (x_i - x_j)(x_i - x_j)^T \quad (5)$$

$$S^{(\text{lw})} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(\text{lw})} (x_i - x_j)(x_i - x_j)^T \quad (6)$$

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

$$W_{i,j}^{(\text{lb})} = \begin{cases} A_{i,j}(1/n - 1/n_l), & \text{if } y_i = y_j = l \\ 1/n, & \text{if } y_i \neq y_j \end{cases} \quad (7)$$

$$W_{i,j}^{(\text{lw})} = \begin{cases} A_{i,j}/n_l, & \text{if } y_i = y_j = l \\ 0, & \text{if } y_i \neq y_j. \end{cases} \quad (8)$$

When these modified scatter matrices are employed in (1), optimizing the modified Fisher ratio results in an LFDA formulation. The weights defined in (7) and (8) give LFDA its neighborhood-preserving properties. The KLFDA algorithm can be viewed as a kernel extension of LFDA via the kernel trick. In this letter, the kernel function employed is the radial basis function (RBF) kernel [6], defined as

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (9)$$

where $\sigma > 0$ is a user-defined parameter of the kernel. In [8], Sugiyama invokes the kernel trick and reformulates the LFDA algorithm in kernel-induced spaces. In other words, the local within- and between-class scatter matrices are defined in the kernel-induced space. Projection \tilde{w} in the kernel-induced space that maximizes the modified Fisher ratio is given by the solution of the generalized eigenvalue problem, i.e.,

$$KL^{(\text{lb})} K \tilde{w} = \tilde{\Lambda} (KL^{(\text{lw})} K + \varepsilon I_n) \tilde{w} \quad (10)$$

where $\tilde{\Lambda}$ is the diagonal eigenvalue matrix; ε is a small (regularization) constant; \tilde{w} is the eigenvector matrix; K is the kernel matrix defined in (9); $L^{(lw)} = D^{(lw)} - W^{(lw)}$, where $D^{(lw)}$ is a diagonal matrix with the i th diagonal element being $D_{ii}^{(lw)} = \sum_{j=1}^n W_{i,j}^{(lw)}$; and $L^{(lb)} = L^{(m)} - L^{(lw)}$, where $L^{(m)}$ is the local mixture matrix defined as $L^{(m)} = D^{(m)} - W^{(m)}$, and $D^{(m)}$ is a diagonal matrix with the i th diagonal element being $D_{ii}^{(m)} = \sum_{j=1}^n W_{i,j}^{(m)}$.

KLFDA borrows the key idea of preserving the local structure of class-conditional distributions from the LPP. The central tenet behind the LPP formulation is to find a linear embedding that is optimal in preserving local structures. In other words, the objective function that is minimized in LPP is designed such that it incurs a heavy penalty [as shown in (4), (7), and (8)] if neighboring points that are close in the input space are mapped far apart in the projection. An adjacency (affinity) matrix identical to that in (4) is used in this formulation to preserve local neighborhood relationships. In [9], the authors perform such an optimization using spectral-graph theory. The efficacy of this approach to preserve local structures of data distributions in the projection is demonstrated in [8]. This benefit is expected to be more pronounced when the statistical structure of the data in the input space is complex, such as when it is multimodal. In KLFDA, the affinity matrix is employed to weight the scatter matrices in the kernel-induced space for a similar effect.

In order to compare the performance of LDA, KDA, and KLFDA, a simple two-class 2-D synthetic data with a significant overlap in the two classes is employed, as illustrated in Fig. 1(a). The RBF kernel was employed for KDA and KLFDA. Fig. 1(b) depicts the probability density function (pdf) of the data in these projected subspaces. It has been previously observed that class-conditional distributions in kernel-projected subspaces tend to be Gaussian [12]. As expected, for this challenging data set, LDA fails to discriminate the data, as shown in Fig. 1(a). It can also be observed from this figure that, for the same input samples, KLFDA (by imposing locality-preserving constraints) provides a tighter distribution (cf., the variance of each class) in the projected space as compared with KDA. The resulting Bayes error is also further reduced with the KLFDA projection. This simple experiment highlights two key points that motivated us to study the benefits of the KLFDA as a dimensionality-reduction method for high-dimensional HSI data. The first point is that LDA becomes ineffective when the data/features in the input space are nonlinearly distributed, which motivates us to use kernel projections under such situations. Several previous studies [6], [7] have demonstrated this benefit for HSI data. The second and more important point is that, by imposing an additional locality-preserving constraint, the class separation in the resulting KLFDA-projected subspace is expected to be higher than that when KDA is employed alone. This is the key motivation behind studying the benefits of this method for HSI classification.

Below, we study the benefits of KLFDA for HSI classification and compare it with other traditional dimensionality-reduction methods, including LDA, regularized LDA (RLDA) [14], subspace KDA, and KDA. As in subspace LDA [15], an intermediate PCA projection is employed in subspace KDA to project the data onto a reduced-dimensional subspace, and

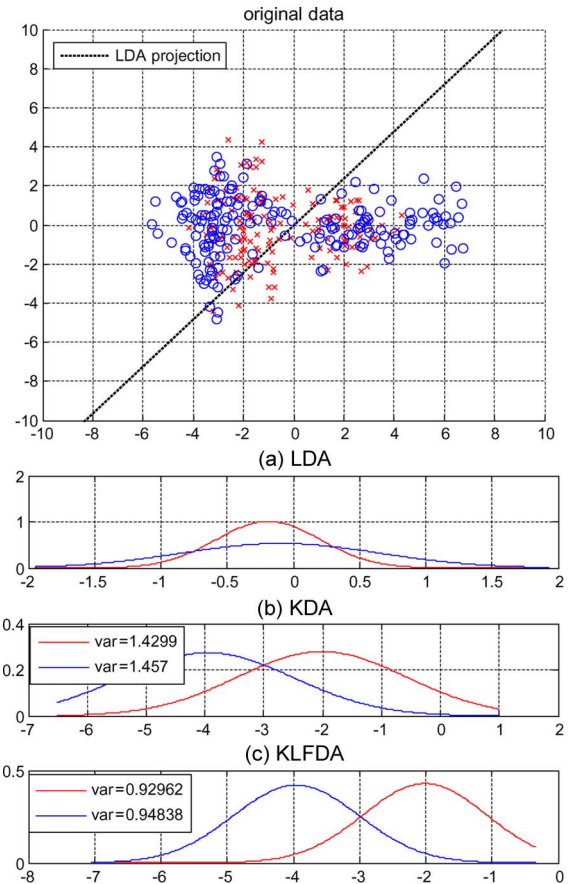


Fig. 1. (a) (Top) Original 2-D data and dimensionality-reduction by the LDA. (b) (Bottom) Pdf of the synthetic data in the left figure projected onto a 1-D subspace using the LDA, the KDA, and the KLFDA (with estimated variances).

KDA is employed on this subspace. This would serve as a good comparison to KLFDA, which essentially combines LPP with KDA to highlight the benefits of employing LPP within the KDA setup. Data distributions in kernel-projected spaces tend to be Gaussian [12]. Hence, we employ a quadratic Gaussian MLE classifier [10] for these experiments. In addition to the baseline experiments above, we also compare the performance of these methods to a standard support vector machine (SVM) [12] classifier. In this approach, a standard recursive feature-elimination (RFE) algorithm [13] is employed to discard features that do not contribute to good discrimination, and an SVM classifier is employed for the final classification. This system (RFE-SVM) [13] is becoming increasingly popular in the remote-sensing and pattern-classification community and serves as a powerful baseline with which to compare the proposed approach.

III. EXPERIMENTAL HYPERSPECTRAL DATA

In this letter, the experimental HSI employed was acquired using the National Aeronautics and Space Administration's airborne visible/infrared imaging spectrometer sensor and was collected over Northwest Indiana's Indian Pine test site in June 1992 [16]. The image represents a vegetation-classification scenario with 145×145 pixels and 220 bands in the 0.4–2.45- μm region from visible to midinfrared. Fig. 2 depicts the spectral signatures for the eight classes extracted from this imagery.

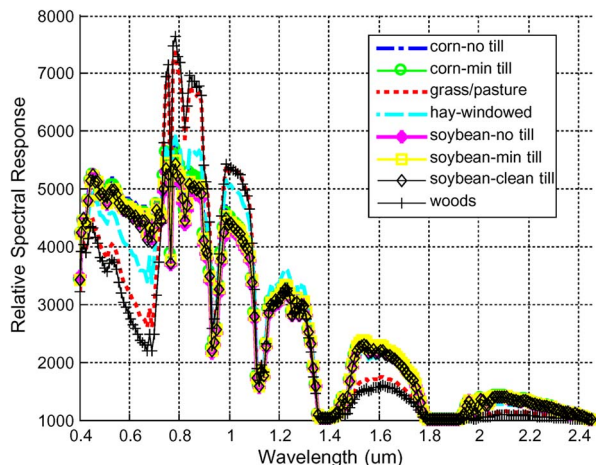


Fig. 2. Spectral signatures of the eight classes that form the hyperspectral classification employed. The number of training samples for every class is 187; the numbers of testing samples are 1247, 647, 310, 302, 781, 2281, 427, and 1107, respectively.

Approximately 8600 labeled pixels are employed to train and validate/quantify the efficacy of the proposed system. This data set is partitioned into approximately 1500 training pixels and 7100 test pixels (the ratio of the number of testing to training samples is approximately 5 : 1). The data used for training and testing the classification models are normalized to have a range of [0,1].

IV. EXPERIMENTAL RESULTS

Here, we provide a quantitative assessment of the efficacy of KLFDA as a dimensionality-reduction technique for HSI data. The performance, as measured by the overall classification accuracy, is reported. The proposed KLFDA-MLE approach is compared with conventional algorithms, including LDA-MLE, RLDA-MLE, KDA-MLE, subspace KDA-MLE, and RFE-SVM. To create challenging real-life operating conditions, we provide results over 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. In these results, we use the data set previously described and linearly mix signatures from background classes with the signature of the class being classified. We report results over a range of percentage target-abundance (TA) values. For example, a TA of 70% indicates that 30% of background signatures were linearly mixed with 70% of the target class. Here, the target class simply refers to any class that is being considered for mixing. The background signatures used for mixing the target class are uniformly gathered from across all the other classes.

A. Optimizing the KLFDA

A variety of kernel functions can be employed for the kernel projection, including linear, polynomial, and RBF kernels. We employ the RBF kernel in our letter since it has been shown to work well for the HSI analysis [6], [7]. For this kernel, parameter σ [cf., (9)] is important because it impacts the generalization ability of the classifier in the resulting kernel-induced space. Fig. 3 shows the overall development data accuracy as

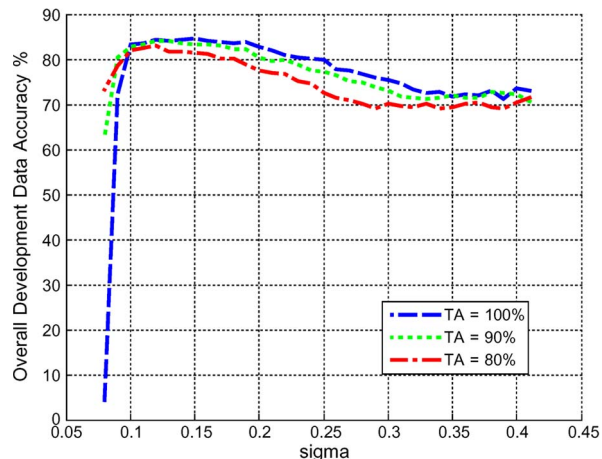


Fig. 3. KDA: Overall development data accuracy versus kernel parameter (σ) for different TAs.

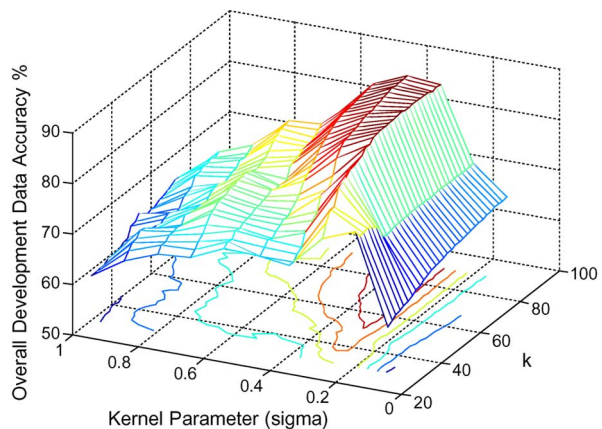


Fig. 4. KLFDA: Overall development data accuracy versus kernel parameter (σ) and parameter k for the TA = 100%.

a function of σ for the KDA. We partition the training samples into further training and testing (development) sets for tuning the parameters. Results with three different TAs are reported. From these figures, it can be inferred that the optimal value of σ is 0.1 for KDA at all TAs.

It is expected that parameter k used to estimated γ_i in (4) will affect the affinity term (and, hence, the locality-preserving properties) in KLFDA. Fig. 4 illustrates the overall development accuracy as a function of k , as well as σ at TA = 100%. We obtained similar performance curves at other TAs. We observe that the accuracy increases when k increases, and it is practically constant when k is higher than 70. We hence infer the optimal value of k to be 70 and that of σ to be 0.3 for KLFDA.

B. Comparison Against Current State of the Art

Fig. 5 illustrates the overall accuracy as a function of the TA. An abundance of 100% implies that pure pixels are employed without any mixing. Results are provided using the proposed method (KLFDA-MLE) and using established baselines (LDA-MLE, RLDA-MLE, KDA-MLE, subspace KDA-MLE, and RFE-SVM). An intermediate PCA dimensionality of 15 was employed in the subspace KDA algorithm (determined by

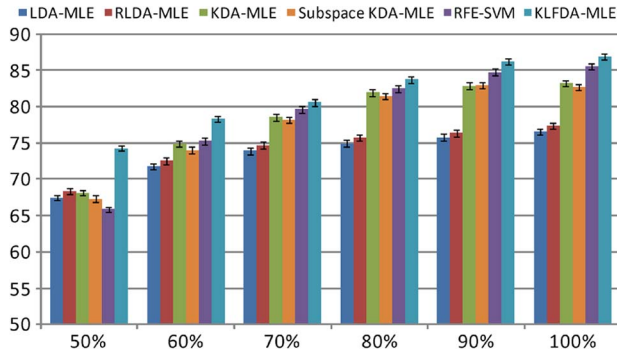


Fig. 5. Overall accuracy versus pixel-mixing abundance, both expressed in percentage, for several different classification methods.

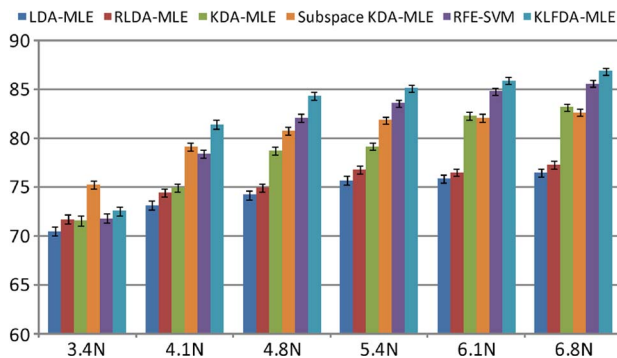


Fig. 6. Overall accuracy (expressed in percentage) versus the training-data set size. Error bars indicate the 95% confidence interval in the accuracy estimates.

tuning the system with the development data). By design, LDA, RLDA, KDA, subspace KDA, and KLFDA result in a 7-D feature subspace after the dimensionality-reduction projection. For the RFE-SVM baseline, the optimal dimensionality for the RFE algorithm was found to be 80, and an RBF kernel with $\sigma = 0.2$ was employed (both values determined by tuning the system with the development data). It can be seen that KLFDA outperforms all other baselines over a wide range of pixel mixing. Furthermore, the drop in the performance as a function of the severity of pixel mixing is much slower with KLFDA when compared with other methods.

We also conducted an experiment wherein we varied the amount of the training data and studied the sensitivity of the proposed method relative to conventional methods over a range of training-data set sizes. In practical situations, oftentimes the number of available training samples is insufficient to estimate models for each class. We report the overall accuracy of the classification systems enlisted above as a function of the relative training sample size. This sample size (on the x -axis) is expressed as a fraction of the dimensionality of the data. Hence, an abundance of $5N$ implies that the amount of the training data used is five times the dimensionality of the feature space. Fig. 6 illustrates results from this experiment. Note that the performance of KLFDA-MLE is significantly higher than that of all the baseline methods even when the training-data set size is small. Subspace KDA fares better than other methods when the amount of training data is very small ($3.4N$)—this is expected because the intermediate PCA projection is discarding the null space of the covariance matrix due to the small sample size.

V. CONCLUSION

In this letter, we have demonstrated the benefits of KLFDA for effective dimensionality reduction and classification of hyperspectral data. It is shown that, when employed together with a simple quadratic Gaussian maximum-likelihood classifier, the resulting classification performance is better than that of most conventional approaches, including those employing powerful and computationally complex classifiers such as RFE-SVM. We have also shown that the proposed approach outperforms others over a wide range of operating conditions, e.g., when the TA in the pixels is poor or when the training-data abundance is poor. The classification task chosen in this letter is a challenging vegetation-classification scenario, wherein the spectral signatures are very similar across the eight classes. By studying the development accuracy as a function of the system parameters (as shown in Figs. 3 and 4), one can tune this system and employ it for any HSI-classification task.

REFERENCES

- [1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY: John-Wiley and Sons, 2001.
- [2] L. Zhang and Y. Zhong, "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.
- [3] 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.
- [4] 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.
- [5] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Neural Netw. Signal Process. Workshop*, 1999, pp. 41–48.
- [6] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [7] S. Prasad and L. M. Bruce, "Information fusion in kernel-induced spaces for robust subpixel hyperspectral ATR," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 572–576, Jul. 2009.
- [8] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. of Mach. Learn. Res.*, vol. 8, pp. 1027–1061, 2007.
- [9] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2003.
- [10] 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, no. 6, pp. 805–814, Nov. 1987.
- [11] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, Oct. 2000.
- [12] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, Dec. 2001.
- [13] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.
- [14] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.
- [15] 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.
- [16] AVIRIS NW Indiana's Indian Pines 1992 Data Set. [Online]. Available: <ftp://ftp.ecn.purdue.edu/biehl/MultiSpec>