A Classification-Based Linear Projection of Labeled Hyperspectral Data

Lior Weizman
School of Engineering, Bar-Ilan University
Ramat-Gan, Israel 52900
weizmanl@gmail.com

Jacob Goldberger
School of Engineering, Bar-Ilan University
Ramat-Gan, Israel 52900
goldbej@eng.biu.ac.il

Abstract— **In this study we apply a variant of a recently proposed linear subspace method, the Neighbourhood Component Analysis (NCA), to the task of hyperspectral classification. The NCA algorithm explicitly utilizes the classification performance criterion to obtain the optimal linear projection. NCA assumes nothing about the form of the each class and the shape of the separating surfaces. Experimental studies were conducted on the basis of hyperspectral images acquired by two sensors: the Airborne Visible/Infrared Imaging Spectroradiometer (AVIRIS) and AISA-EAGLE. Experimental results confirm the significant superiority of the NCA classifier in the context of hyperspectral data classification over methodologies that were previously suggested.**

Index Terms - Classification, hyperspectral images, remote sensing, linear projection, NCA.

I. Introduction

A remarkable increase in spectral resolution has led to imaging sensors that can gather data in hundreds of contiguous narrow spectral bands to generate hyperspectral images. This ability of imaging sensors to acquire the reflectance spectrum of a pixel in significant detail, leads to substantial differences in the reflectance values of the pixels belonging to disparate materials on the Earth's surface [7]. The automatic analysis of hyperspectral data, however, is not a minor task. One of the major difficulties in hyperspectral classification is the curse of dimensionality. The enormous number of features in a hyperspectral image is often a major drawback for several reasons: 1) When using supervised classification, the number of training samples required for reasonable classification results depends on a large number of feature vectors [9] which is not always available; 2) On increasing the number of features given as input to the classifier over a given threshold, the classification accuracy decreases (this behavior is known as the Hughes phenomenon [15]); 3) A hyperspectral image generally consists of thousands of pixels over hundreds of spectral bands. Classification of this tremendous amount of data, is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications. Therefore, the traditional but still common approach for classification of a hyperspectral image is consists of a feature reduction/selection procedure and a conventional classifier. There are, however publications on classification that do not follow this approach, such as [17], [6].

Reducing the data dimensionality can be done by a selection of only several suitable bands for classification, with respect to the specific application, such as presented in [13]. More complicated tools, such as Principal Component Analysis (PCA)[8] and the minimum noise fraction transform [12], [16] have been developed specifically to address the efficient extraction of spectral features from hyperspectral data sets.

As for an the appropriate classification algorithm many supervised methods have been developed to tackle the multi- and hyperspectral data classification problem. Basically, several approaches exist for supervised classification algorithms of hyperspectral data. The Bayes classifier is very common when classifying a hyperspectral image, but it suffers from improper modeling versus the real world [10]. Kernel-based methods are another approach for supervised classification of hyperspectral data [5]. Neural networks [4], standard support vector machines (SVMs) [17], kernel Fisher discriminant (KFD) analysis [11], and regularized AdaBoost [18] are the most common kernel-based methods used for hyperspectral classification today.

In this paper we apply a recently proposed linear subspace method, the Neighbourhood Component Analysis (NCA) [14] to the task of hyperspectral classification. The NCA algorithm explicitly utilizes the classification performance criterion to obtain the optimal linear projection. The fact that the optimization criterion of previously proposed subspace methods is not explicitly related to the classification target results in a need for an additional learning procedure that should find a suitable distance function in the transformed subspace. In the proposed method the distance measure used in the transformed subspace is explicitly stated in the optimization cost function. The optimal transformation is selected such that using the Euclidean distance in the transformed space yields optimal classification results.

The experimental studies were carried out on the basis of hyperspectral images acquired by two sensors: the Airborne Visible/Infrared Imaging Spectroradiometer (AVIRIS) sensor [1], and the AISA-EAGLE sensor [2]. Experimental results confirm the significant superiority of the NCA classifier in the context of hyperspectral data classification over the conventional classification methodologies, whatever multi-class strategy is utilized to face the multi-class dilemma.

II. Methods

In this section we review the NCA algorithm [14]. We begin with a labeled data set consisting of n real-valued input vectors $x_1, \ldots, x_n$ in $\mathcal{R}^D$ and corresponding class labels $c_1, ..., c_n$. In the case of hyperspectral images, the vectors are the spectral signatures of the pixels, and the labels are the land-cover classes. We want to find a low-dimensional linear transformation $\mathbf{A} : \mathcal{R}^D \rightarrow \mathcal{R}^d$ that maximizes the performance of nearest neighbour classification in the reduced space. Ideally, we would like to optimize performance on future test data, but as we do not know the true data distribution we instead attempt to optimize leave-one-out (LOO) performance on the training data. Given a finite set of linear transformations to choose from, we can easily select the best one, namely the one that minimizes the number of classification errors. The nearest-neighbour classification error, however, is quite a discontinuous function of the transformation $\mathbf{A}$, given that an infinitesimal change in $\mathbf{A}$ may change the neighbour graph and thus affect LOO classification performance by a finite amount. Hence we can not use this optimization criterion in our case where there is a continuously parameterized family of linear transformations which must be searched. Instead, we adopt a more well-behaved measure of nearest-neighbour performance, by introducing a differentiable cost function based on stochastic ("soft") neighbour assignments in the transformed space. In particular, each point i selects another point j as its neighbour with some probability p_{ij}, and inherits its class label from the point it selects. We define the p_{ij} using a softmax over Euclidean distances in the transformed space:

$$p_{ij}(\mathbf{A}) = \frac{\exp(-\frac{1}{2}\|\mathbf{A}x_i - \mathbf{A}x_j\|^2)}{\sum_{k \neq i} \exp(-\frac{1}{2}\|\mathbf{A}x_i - \mathbf{A}x_k\|^2)} \quad , \quad p_{ii} = 0 \tag{1}$$

Denote the set of points in the same class as i by $C_i = \{j | c_i = c_j\}$. Under the stochastic selection rule (1), we can compute the probability p_i that a point i will be correctly classified:

$$p_i = \sum_{j \in C_i} p_{ij} \tag{2}$$

The objective function we maximize is the following:

$$C(\mathbf{A}) = \sum_i \log(\sum_{j \in C_i} p_{ij}) = \sum_i \log(p_i) \tag{3}$$

Maximizing this objective would correspond to maximizing the probability of obtaining a *perfect (error free) classification of the entire training set*. Maximizing the objective function $C(A)$ is also equivalent to minimizing the Kullback-Leibler divergence between the true class distribution (having probability one on the true class) and the stochastic class distribution induced by p_{ij} via $\mathbf{A}$. Note that since $\|\mathbf{A}x_i - \mathbf{A}x_j\|^2 = (x_i - x_j)^\top \mathbf{A}^\top \mathbf{A}(x_i - x_j)$, the optimization criterion depends only on $\mathbf{A}^\top \mathbf{A}$. Hence, every orthogonal matrix $R_{d \times d}$ yields a solution $R \cdot A$ that is completely equivalent to $\mathbf{A}$. To keep the representation parsimonious we can use the Choleski decomposition representation by forcing the entries of $\mathbf{A}$ below the main diagonal to be zero and the entries on the diagonal to be non-negative. This makes the representation of $\mathbf{A}$ unique.

Differentiating C with respect to the transformation matrix $\mathbf{A}$ yields an expression which can viewed as the difference between the overall variability and the intra-class variability defined by the probabilistic model (1) induced from $\mathbf{A}$ [14]. The learning algorithm therefore is: maximize the above objective (3) using a gradient-based optimizer such as conjugate gradients. Of course, as the cost function above is not convex, some care must be taken to avoid local maxima during training. We have experimentally observed that the linear transformation obtained by the RCA method [3] can serve as a good and easily computed starting point for the conjugate gradient algorithm.

III. EXPERIMENTAL RESULTS

A. Dataset Description and Experiment Design

The following experimental results are based on two hyperspectral datasets. The first hyperspectral dataset used in our experiments is a section of a scene taken over northwest Indianas Indian Pines by the AVIRIS sensor in 1992 [1]. From the 220 spectral channels acquired by the AVIRIS sensor, 20 channels were discarded as they were affected by atmospheric problems. From the 16 different land-cover classes available in the original ground truth, seven were discarded, since only few training samples were available for them (this makes the experimental analysis more significant from the statistical viewpoint). The remaining nine land-cover classes were used to randomly generate a set of 4757 training samples (used for learning the classifiers). The remaining 4588 samples were used as test samples (exploited for assessing their accuracies). This dataset was also used in [17] to examine the classification of hyperspectral images with support vector machines. In experiment 1, we use this dataset to compare the performances of NCA with those of the SVM and four other nonparametric classifiers. These nonparametric classifiers are the radial basis function neural network, which is another kernel-based classification method (like SVMs) that uses a different classification strategy based on a statistical rather than a geometrical criterion, the K-nearest neighbors classifier, which is widely used in pattern recognition as a reference classification method. We have also compared the NCA with linear dimensionality reduction methods, the RCA and the LDA that were mentioned in the previous section. The second hyperspectral dataset is subset of a scene acquired by the AISA EAGLE sensor over southern Israel in 2004. This dataset has spatial dimensions of 294×850 pixels, and 59 spectral bands. To reduce the computational time, only 20 bands were used in the experiments. The ground truth of this image consists of 8 major land-cover classes. Ten percent of the pixels in this dataset were used as training samples, while the other pixels in this set were used as test samples. Experiment 2 is applied over the second dataset to examine the performances of the NCA while the number of test samples is relatively large compared to the number of training samples.

B. Experiment 1: Results of DATASET I

Melgani and Bruzzone ([17] examined the SVM classifier when applied to hyperspectral data. Their results imply that

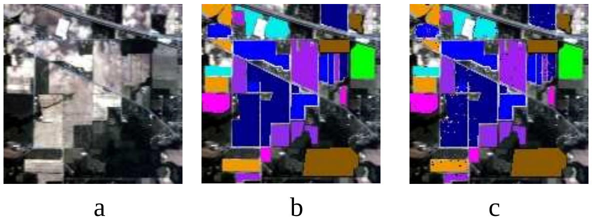

Fig. 1. Classification results of experiment 1. (a) The hyperspectral image in true color, (b) The ground truth, (c) The NCA classification results.

nonlinear SVM gives the best overall accuracy of the selected dataset. We decided to repeat the same setup using the NCA as a classifier, while the results in terms of classification accuracy and computational time provided by the different classifiers are summarized in Table I (The results of SVM-Lineae, SVM-RBF, K-nn classifier and RBF classifier are taken from [17]). Fig. 1 shows a visual impression of the NCA results on database I. The data was projected on an 8-dimensional space as part of the NCA algorithm.It can be seen from Table I that the NCA exhibited the best Overall Accuracy (OA), i.e., the best percentage of correctly classified pixels among all the test pixels considered. It is worth noting that, the small number of training samples (4757) is insufficient to properly fill the emptiness of the hyperdimensional feature space. This fact leads to a relatively poor classification accuracies of the K-nn classifier. However, the NCA computes the linear transformation matrix which brings the data into a lower feature space, and then also uses the K-nn classifier. Therefore, it can be seen that the same training points which were not sufficient for the K-nn in the original feature space, now lead to the best results in the reduced features space, computed by the NCA. This fact shows the ability of the NCA transformation matrix to maximize the results given the training samples set.

C. Experiment 2:Results of DATASET II

The second experiment examined the NCA performance on large database, with the number of test samples being relatively large compared to the number of training samples. Only 10% of the pixels were used as training pixels, while the remaining 90% of the pixels were used as testing samples. The data was reduced to dimension of 4 as part of the NCA algorithm. The results of this experiment show the NCA results compared to five other classifiers: the linear SVM, non-linear SVM, Mahalanobis Distance, LDA and the RCA classifiers. The linear and non-linear SVM parameters were chosen to be the parameters that led to the best results in [17]. Detailed description about the Mahalanobis distance classifier can be found in [19]. The results of this experiments are shown in Table II, and in Figure 2. We can see from the results that the NCA overcomes all the other classifiers in term of overall accuracy. It can be noted that the Mahalanobis distance classifiers fail to classify classes with high number of training and test samples. This is attributes for the fact is that classes with high number of samples vary highly over the feature space, and the single Gaussian model can not model this variability.

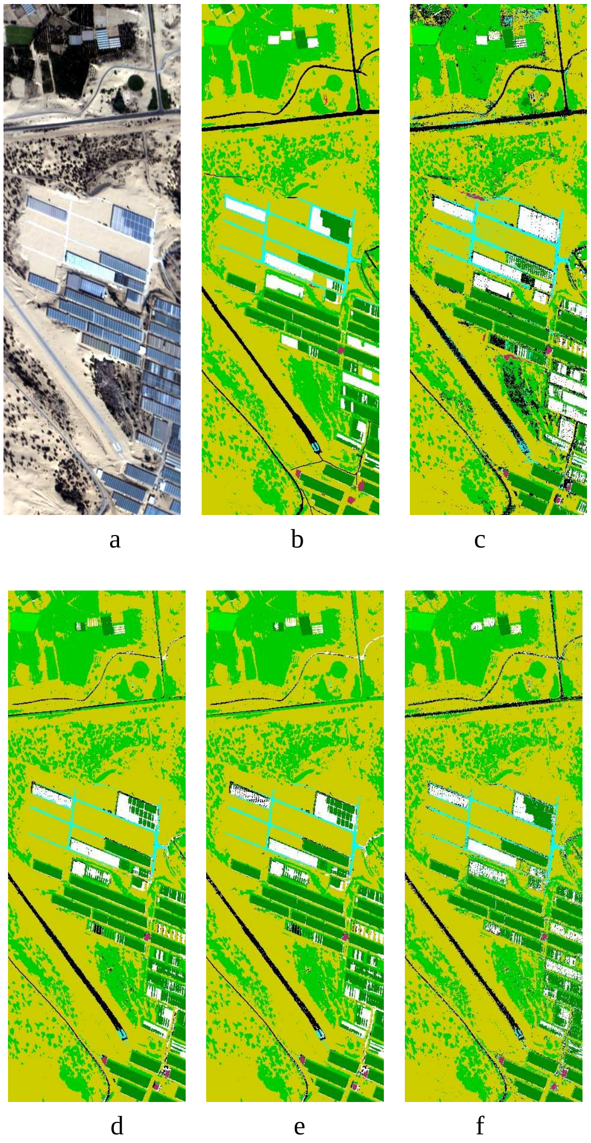

Fig. 2. Classification results of experiment 2. (a) The hyperspectral image in true color, (b) The manually obtained, ground truth, (c) The Mahalanobis Distance classification results, (d) The Linear-SVM results, (e) The SVM-RBF classification results, and (f) The NCA classification results.

IV. DISCUSSION AND CONCLUSION

In this paper, we addressed the problem of classifying hyperspectral remote sensing data using Neighbourhood Components Analysis. This method is now in wide use, particularly in areas of image classification. The NCA exhibited the best results compared to the state-of-the-art classifiers currently being used. This paper also contains a modification of the method that can be used to weight more important and less important classes during the classification. The modified method is useful in cases where a high accuracy is required only for some of the classes in the scene. This work was carried out on images that were acquired using hyperspectral sensors. The use of multi-spectral sensor is much more financially attractive and

TABLE I

BEST OVERALL AND CLASS-BY-CLASS ACCURACIES, AND COMPUTATIONAL TIMES ACHIEVED ON THE TEST SET BY THE DIFFERENT CLASSIFIERS, (DATABASE I)

Method	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	OA	TIME (s)
NCA	90.61	94.39	97.89	99.44	99.21	92.31	92.97	92.23	100	94.7	1770
SVM-Linear	89.02	69.13	94.51	98.60	100	75.47	83.48	83.17	99.22	87.10	40342
SVM-RBF	91.47	87.76	94.94	98.88	100	88.57	91.25	95.79	99.38	93.42	2702
K-nn classifier	96.73	61.16	86.59	80.46	99.60	98.88	90.72	65.82	74.42	83.94	2618
RBF classifier	98.44	74.11	88.47	79.83	99.21	98.04	91.98	73.72	80.06	86.99	4743
LDA	71.97	77.81	97.47	98.32	99.21	81.50	80.95	67.96	98.13	84.13	15
RCA	80.06	81.12	98.73	99.16	99.21	81.70	86.51	88.07	99.53	88.82	17

TABLE II

BEST OVERALL AND CLASS-BY-CLASS ACCURACIES, AND COMPUTATIONAL TIMES ACHIEVED ON THE TEST SET BY THE DIFFERENT CLASSIFIERS, (DATABASE II)

Method	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	OA	TIME (s)
NCA	64.06	86.74	91.16	89.73	75.92	93.97	57.00	60.87	89.37	4582
SVM-Linear	43.59	86.16	97.28	85.31	62.96	95.36	45.48	47.82	88.22	1309
SVM-RBF	36.76	86.70	95.00	86.12	63.53	95.29	43.13	44.02	88.17	1293
Mahalanobis Distance	83.37	80.94	99.62	66.64	79.26	89.27	64.80	49.76	83.90	67
LDA	54.88	84.33	92.86	88.64	68.47	93.32	52.24	51.99	87.74	59
RCA	58.73	85.21	99.07	90.22	72.84	93.53	55.07	51.99	88.48	63

much less time consuming than use of hyper-spectral sensors. The ability of the NCA to obtain high-accuracy classification in a low-dimension space, can lead to a specially designed multi-spectral sensor for a specific scene or a mission. This way, much money and time can be saved if the type of scene or mission is well specified before acquiring the image. Note that neither spatial nor morphological information was used in our experiments. It is obvious, however, that using such information will improve the results. Therefore, our future work will focus on improving the results while embedding spatial and morphological information. Another future research direction is reducing the computational complexity of the NCA algorithm.

References

[1] Aviris nw indianas indian pines 1992 data set [online]. Available:ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C(original files)and ftp://ftp.ecn.purdue.edu/biehl/PC_MultiSpec /ThyFiles.zip (ground truth).

[2] Spcim web-site. Available:http://www.specim.fi/products-aisa-eagle.html.

[3] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis distance from equivalence constarints. *Journal of Machine Learning Research*, pages 937–965, 2005.

[4] L. Bruzzone and D. Fernndez-Prieto. A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images. *IEEE Trans. Geosci. Remote Sens.*, 37(2):1179–1184, 1999.

[5] G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral images classification. *IEEE Trans. Geosci. Remote Sens.*, 43(6):1351–1362, 2005.

[6] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *Geosci. and Remote Sens. Letters, IEEE*, 3:93–97, 2006.

[7] C.-I. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Kluwer, 2003.

[8] C.-I Chang and Q. Du. Interference and noise adjusted principal components analysis. *IEEE Trans. Geosci. Remote Sensing*, 37:2387–2396, 1999.

[9] M. P. Derde and D. L. Massart. Evaluation of the required sample size in some supervised pattern recognition techniques. *Analytica Chimica Acta*, 223(1):19–44, 1989.

[10] M. M. Dundar and D. A. Landgrebe. Toward an optimal supervised classifier for the analysis of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 42(1):271–277, 2004.

[11] M. Murat Dundar and A. Landgrebe. A cost-effective semisupervised classifier approach with kernels. *IEEE Trans. Geosci. Remote Sens.*, 42(8):1778–1796, 2004.

[12] A. A. Green et al. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. Geosci. Remote Sensing*, 26:65–74, 1998.

[13] M. E. Martin et al. Determining forest species composition using high spectral resolution remote sensing data. *Remote Sens. Environ.*, 65:249–254, 1998.

[14] J. Goldberger, S. Roweis, G. Hinton, and R. Salkhutdinov. Neighbourhood components analysis. *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[15] G. F. Hughes. On the mean accuracy of statistical pattern recognition. *IEEE Trans. Inform. Theory*, IT-14:55–63, 1968.

[16] J. B. Lee, A. S.Woodyatt, and M. Berman. Enhancement of high-spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Trans. Geosci. Remote Sensing*, 28:295–304, 1990.

[17] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.*, 42(8):1778–1796, 2004.

[18] R. Nishii and S. Eguchi. Supervised image classification by contextual adaboost based on posteriors in neighborhoods. *IEEE Trans. on Geosci. and Remote Sens.*, 43:2547–2554, 2005.

[19] A. J. Richards and X. Jia. *Remote Sensing Digital Image Analysis*. Springer-Verlag, 1999.