Applying Nonlinear Manifold Learning to Hyperspectral Data for Land Cover Classification

Yangchi Chen and Melba M. Crawford
Center for Space Research
3925 W. Braker Lane, Austin, TX 78759
The University of Texas at Austin
Email: yanji@csr.utexas.edu and crawford@csr.utexas.edu

Joydeep Ghosh
Department of Electrical and Computer Engineering
The University of Texas at Austin
Email: ghosh@ece.utexas.edu

***Abstract*— The shortest path k-nearest neighbor classifier (SkNN), that utilizes nonlinear manifold learning, is proposed for analysis of hyperspectral data. In contrast to classifiers that deal with the high dimensional feature space directly, this approach uses the pairwise distance matrix over a nonlinear manifold to classify novel observations. Because manifold learning preserves the local pairwise distances and updates distances of a sample to samples beyond the user-defined neighborhood along the shortest path on the manifold, similar samples are moved into closer proximity. High classification accuracies are achieved by using the simple k-nearest neighbor (kNN) classifier. SkNN was applied to hyperspectral data collected by the Hyperion sensor on the EO-1 satellite over the Okavango Delta of Botswana. Classification accuracies and generalization capability are compared to those achieved by the best basis binary hierarchical classifier, the hierarchical support vector machine classifier, and the k-nearest neighbor classifier on both the original data and a subset of its principal components.**

I. Introduction

Achieving both high classification accuracy and good generalization when sample sizes are small relative to the dimension of the input space continues to be a challenging problem, especially when the number of classes is also large. Previous studies of supervised methods show that a complex classifier tends to overtrain in such situations, while a weak classifier is often inadequate [1].

As classifiers become more complex, the generalization error eventually increases because of over-training [2]. Ensemble methods can alleviate this problem by reducing the model variance. In particular, the random forest classifier, which combines bagging and random subspace methods, achieves both high classification accuracies and good generalization, but is computationally costly due to the large (50-100) number of classifiers required in the ensemble [3]. Complex classifiers also do not typically perform well when characteristics of the training/test data acquired over the study site evolve in a new area. This is referred as the knowledge transfer problem [3]. It is an important problem in land cover classification because it is often difficult to obtain labeled samples from a new area. Seasonal changes, unknown land cover types or a different mixture of classes can cause changes in spectral signatures. It is important to develop a simple classifier that can adapt to such changes, as well as maintain good classification accuracies for the training/testing data.

Most supervised classification algorithms, such as maximum likelihood, Fisher linear discriminant, or decision trees rely heavily on feature selection or feature extraction to mitigate the effect of high-dimensionality. Support vector machines (SVM), which maximize margin instead of classification accuracies in the objective function, can handle high dimensional data while not overfitting the training data [4]. These classifiers build their models according to the behavior of labeled samples in the reduced or original feature space.

In contrast, nonlinear manifold learning algorithms focus on samples. These algorithms assume that the original high dimensional data actually lie on a low dimensional manifold defined by local geometric differences between samples. Isometric feature mapping (Isomap) [5] and local linear embedding (LLE) [6] are representatives of this approach. Although they were developed to embrace the idea of nonlinear dimension reduction and representation of high dimensional observations, they also provide an alternative direction for classification of hyperspectral data.

Manifold learning was recently applied to hyperspectral data by Bachmann *et al.* [7]. Results indicated that Isomap is more efficient than the maximum noise fraction (MNF) transform [8] in data compression, but is computationally intensive. Implementation via a tiling method reduced computational time when manifold learning was applied to large-scale images. In this paper, we apply Isomap to hyperspectral data to evaluate the effects of nonlinear dimension reduction on classification. A discussion of Isomap, that includes the shortest path algorithm and multidimensional scaling (MDS), is contained in Section II-A. The proposed shortest path k-nearest neighbor classifier (SkNN), which is closely related to the shortest path updating scheme, is also described. Results of dimension reduction for a test site in Botswana and comparisons of classification accuracies achieved by SkNN and other competitive classifiers are presented in Section III. Conclusions and a discussion of future research directions are contained in Section IV.

II. Methodology

A. Isometric Feature Mapping (Isomap)

Isomap nonlinear manifold learning is based on shortest path network updating and multidimensional scaling (MDS).

The original input $\mathbf{X} \in \Re^{d \times n}$, representing n samples and d dimensions, is first used to calculate the pairwise distances within a user-defined neighborhood. A shortest path algorithm is applied to update those pairwise distances beyond the neighborhood. The updated distance matrix is used by MDS to evaluate the true dimension of the manifold.

1) Shortest Path Network: Isomap uses a user-defined neighborhood and the shortest path algorithm to discover the manifold. It first defines K_i, the set of neighborhood nodes of node i, to create a distance matrix $\mathbf{D}'$. If $j \in K_i$, $d'_{ij} = d_{ij}$. If $j \notin K_i$, $d'_{ij} = \infty$. Isomap then accumulates the distance beyond the set K_i along the shortest path to obtain $\mathbf{D}_{stp}$.

The shortest path network is constructed from a directed graph $G = (N, E)$, where N represents the nodes, and E represents the edges of the graph. The value of d_{ij} represents the length (cost) of E_{ij}, while x_{ij} is the amount of flow from N_i to N_j. The shortest path algorithm finds the paths from a root node N_1 to all other nodes to minimize the sum of the individual path lengths. This problem is formulated as a network flow programming problem:

$$\min \quad z = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij}$$

$$s.t. \quad \sum_{j=2}^{n} x_{1j} = n - 1 \tag{1}$$

$$\sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} = -1, \quad i = 2, ..., n \tag{2}$$

$$x_{ij} \geq 0, i \neq j = 1, ..., n \tag{3}$$

In this optimization problem, Eq. (1) is the supply of the root node, (2) represents conservation of flow, and (3) is the non-negativity constraint. Because this is a pure network flow problem, it can be modeled as a linear programming problem and solved either via the simplex method or an interior point method, yielding an optimal integer solution, x^* [9]. Isomap solves the problem efficiently via a simple, computationally efficient algorithm developed by Dijkstra [10] [1].

2) Multidimensional Scaling: Multidimensional scaling (MDS) is a linear dimension reduction technique that places a set of samples in a meaningful dimensional space that explains the similarity between samples. Given a distance matrix $\mathbf{D}$, and assuming that a $\mathbf{Y} \in \Re^{l \times n}, l \ll d$ exists such that $\delta_{ij}^2 = ||\mathbf{y}_i - \mathbf{y}_j||^2 \approx d_{ij}^2$ and $\mathbf{Y}_i$ are orthogonal, it can be shown that $\mathbf{Y}$, calculated by classical MDS, is equivalent to a vector of the first l principal components of $\mathbf{X}$ if the Euclidean pairwise distance matrix is used [11]. Here, MDS is used to evaluate the true dimension of $\mathbf{D}_{stp}$.

Experiments in the [5] demonstrated that $\mathbf{D}_{stp}$ is able to define the nonlinear manifold, and that it can be represented globally by MDS in a lower dimensional space. For example, if the pairwise distances between a set of 100 cities of the US are represented by MDS, a three dimensional space is required to preserve the pairwise relationships between these cities globally. If the distance is updated locally and nonlinearly so that only distances between a city and cities of a defined neighborhood are considered, these cities lie on a two dimensional map.

[1]For more details, please see http://www.cs.utexas.edu/users/EWD/

B. Shortest Path k-Nearest Neighbor Classifier

If high dimensional data can be preserved in a low dimensional manifold, the updated distance matrix, which preserves the local information on a graph while increasing the distances between non-neighbor samples, should be useful for classification. A k-nearest neighbor classifier is applied to the new distance features to investigate the advantage of manifold learning.

In this study, $\mathbf{D}_{stp}$ is shown to be potentially useful for land cover classification. Figs. 1 and Fig. 2 show that Isomap moved similar samples closer to each other, while dissimilar points are more separated. If a set of samples can be

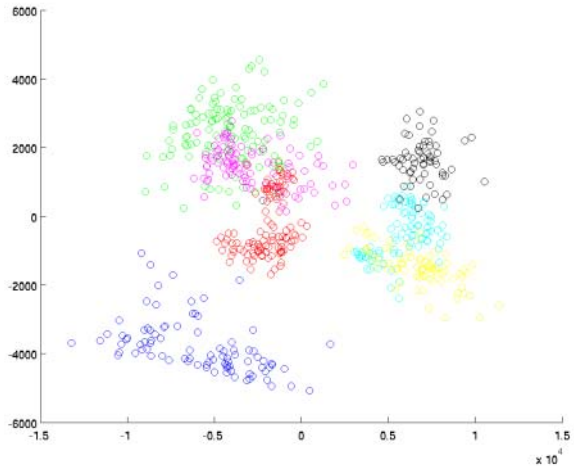

Fig. 1. Two dimensional PCA plot, 8 Classes (exclude water), Hyperion Data of Botswana

presented in a low dimensional space, the simple k-nearest neighbor classifier is often the most competitive algorithm. Given a novel observation, kNN classifies it according to the class label of its k nearest neighbors, in the distance sense. The kNN has several advantages. It is easy to implement, its classification accuracy is very good on low dimensional problems, and it provides nonlinear decision boundaries. It is also straightforward to extend for multi-class problems that are the norm in cover classification problems.

The shortest path k-nearest neighbor classifier (SkNN) is proposed to utilize the information learned from the low dimensional nonlinear manifold. Instead of using Euclidean distance, SkNN approximates each spectral signature as a probability distribution and uses

$$d_{ij} = \frac{1}{2} \sum_{\forall x} \left(f_i(x) \log \frac{f_i(x)}{f_j(x)} + f_j(x) \log \frac{f_j(x)}{f_i(x)} \right) \tag{4}$$

the average Kullback-Leibler divergence [12] between the spectral signatures of sample i and sample j. This KL-distance

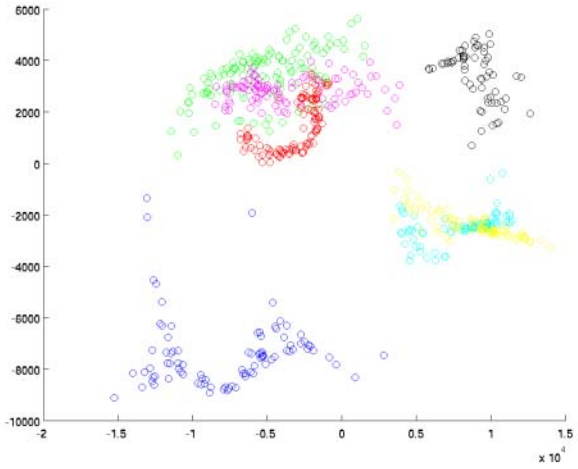

Fig. 2. Two dimensional Isomap plot, 8 Classes (exclude water), Hyperion Data of Botswana

matrix $\mathbf{D}$ is converted to $\mathbf{D}_{stp}$ as described in Section II-A.1. The k-nearest neighbor algorithm then classifies the unlabeled samples projected in the space of the new distance matrix $\mathbf{D}_{stp}$

III. Results

The benefits of applying nonlinear manifold learning to hyperspectral data were evaluated in terms of reduction and classification of the Hyperion data.

The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001-2003. The Hyperion sensor on EO-1 acquires data at $30m^2$ pixel resolution over a 7.7 km strip in 242 bands covering the 400-2500 nm portion of the spectrum in 10 nm windows. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220]. The data analyzed in this study, acquired May 31, 2001, consist of observations from 9 identified classes that include: water (158 total samples), primary floodplain (228), riparian (237), firescar (178), island interior (183), woodlands (199), savanna (162), short mopane (124) and exposed soils (111). These classes represent the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta.

A. Dimension Reduction

Although studies have shown that principal component analysis (PCA) can reasonably discover the true structure of the 150+ bands hyperspectral data on a linear subspace, the original high dimensional data may lie on a nonlinear manifold. To evaluate the true dimension of the manifold, SStress: $ss = \left[\frac{\sum\sum_{i<j}(d_{ij}^2-\delta_{ij}^2)^2}{\sum\sum_{i<j} d_{ij}^4}\right]^{\frac{1}{2}}$ [13] is used by MDS to evaluate the similarity of $\mathbf{D}_{stp}$ and Δ_l, where l is the dimension of $\mathbf{Y}$. The value of SStress is always between 0 and 1. Any value less than 0.1 is considered to indicate good representation in the given l dimensions. Values in Table I, indicate that SStress became less than 0.1 when $l \geq 2$. Thus, Isomap found the embedding manifold of this Hyperion data could be represented in a very low dimensional space with little loss of information.

TABLE I
BOTSWANA DATA: SSTRESS WITH l DIMENSIONS

l	1	2	3	4
SStress	0.167	0.046	0.042	0.046

B. Classification

Ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, these training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data. Because the training and test data are spatially collocated, a spatially disjoint test set was also acquired and used to evaluate the generalization of these classifiers to another area. Note that this extended data may have substantially different characteristics as it is collected from a geographically separate location. The goal here is to investigate the capability of the various methods for extending results obtained from one area to other areas where data are not so spatially correlated with the original training samples. Hereafter, these data are referred to as the test and spatially disjoint (SD) test data, respectively.

Experiments were performed using the best basis binary hierarchical classifier (BB-BHC) with weighted prior [14], hierarchical SVM (HSVM) [15], k-nearest neighbor (kNN) on the original space, and the proposed shortest path k-nearest neighbor (SkNN). Here, k = 5 was chosen by a cross-validation scheme. The cross-validation showed that SD test accuracies increased slightly, but test accuracies decreased when k was increased. Results were also obtained by kNN using $3 \sim 5$ PC bands, but the resulting accuracies were consistently lower than what the kNN classifier applied to the original space had achieved.

The average test data classification accuracies and their corresponding standard deviations for the 10 experiments conducted with each classifier are listed in Table II. The overall trend shows that classification accuracies of the test

TABLE II
BOTSWANA TEST DATA: ACCURACY (STD. DEV.)

Training %	BB-BHC	HSVM	kNN	SkNN
15%	92.6(2.16)	96.5(0.95)	83.2(3.17)	94.2(1.19)
30%	95.5(1.68)	97.3(1.14)	91.3(1.74)	96.4(1.33)
50%	97.6(0.74)	97.9(0.51)	95.0(1.27)	97.1(1.24)
75%	98.1(0.60)	97.7(0.51)	96.1(1.24)	97.5(0.81)

TABLE III
BOTSWANA SPATIALLY DISJOINT (SD) TEST DATA: ACCURACY (STD. DEV.)

Training %	BB-BHC	HSVM	kNN	SkNN
15%	84.1(2.70)	80.7(3.1)	77.3(2.22)	84.7(2.14)
30%	84.3(1.13)	84.1(1.57)	79.8(1.24)	86.1(2.60)
50%	84.5(1.24)	84.1(0.83)	81.3(0.86)	86.8(2.06)
75%	85.8(0.60)	84.6(0.61)	82.2(0.60)	87.5(1.06)

set increase as the size of training sample increases for all four classifiers, while HSVM achieves both the highest overall average accuracies, with the smallest standard deviations at small sample sizes. Besides kNN, the other three classifiers perform well at 15% sampling rate, which indicates that they can all handle small samples of test data.

Classification accuracies on the spatially disjoint (SD) test set are contained in Table III. HSVM performed consistently well on the test set, but not on the SD test set. Because the spectral characteristics of the train/test data are different from that those of the SD test set, this supported the notion that while SVM is a strong classifier, it might be robust to changes in data characteristics. Although BB-BHC was shown to be competitive to SkNN, as indicated by the relatively low variance of accuracies obtained on the SD test set, the average accuracy of the individual classes ranges from 70-100% for the BB-BHC, and 80-100% for the SkNN, while ranges of the standard deviations of the respective accuracies are 1.8-5.6 and 2.8-10, respectively. SkNN achieves higher accuracies accuracies but larger standard deviations because it is more sensitive to samples. Since $\mathbf{D}_{stp}$ evolves as new samples are included in the distance matrix, SkNN not only performed well on the test set but also produced the highest accuracies on the SD test set at all four sampling rates. Results from kNN are included to demonstrate that the method does not produce high accuracies using the original Euclidean pairwise distance matrix, but not the shortest path algorithm. For a Botswana experiment that has 790 samples, 9 classes and 145 feature spaces, using a 3GHz Pentium 4 CPU machine, kNN finished training and testing in 31 seconds, while HSVM required 40 seconds. BB-BHC required 65 seconds and the proposed SkNN required 149 seconds of CPU time.

IV. CONCLUSION

In this paper, we explored the concept of nonlinear manifold learning, which assumes that the original high dimensional data can be represented on a low dimensional manifold defined by pairwise distances between local samples. Evaluations of dimension reduction and representation of high dimensional observation by Isomap were conducted. This approach was also extended to the classification of hyperspectral data. The shortest path k-nearest neighbor classifier (SkNN), that utilizes nonlinear manifold learning, was proposed and compared to other competitive classifiers such as BB-BHC and HSVM.

Two conclusions were drawn from our experiments. First, high classification accuracies were achieved by using two simple algorithms collectively. Isomap found the nonlinear manifold of the 150+ bands hyperspectral data and represented it on a low dimensional space. Because of the availability of the low dimensional manifold, SkNN was competitive, as indicated by results shown in Fig. 2 and the two accuracy tables. Second, because of the shortest path updating scheme, SkNN evolved with changes of spectral characteristics from the train/test set to a new area, thus providing the highest overall average accuracies on the spatially disjoint test set.

Applying nonlinear manifold learning to hyperspectral data provided promising initial results. Future studies will involve investigation of alternative distance measures. Research is also being conducted to further increase the speed of SkNN when the number of samples is large. Additionally, ensemble methods will be investigated to reduce the variance, and approaches for incorporating neighborhood information will be explored.

ACKNOWLEDGMENT

This research was supported by NSF (Grant IIS-0312471). We thank Amy Neuenschwander of the UT Center for Space Research for help in pre-processing the Hyperion data and interpreting the overall classification results.

REFERENCES

[1] B. E. Boser, I. Guyon, and V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Computational Learning Theory*, 1992, pp. 144–152.

[2] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," in *Proc. 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 322–330.

[3] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. and Remote Sens.*, vol. 43, no. 3, pp. 492–501, March 2005.

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[7] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. and Remote Sens.*, vol. 43, no. 3, pp. 441–454, March 2005.

[8] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 26, no. 1, pp. 65–74, 1988.

[9] L. A. Wolsey, *Integer Programming*. John Wiley & Sons, 1998.

[10] E. W. Dijkstra, "Note on two problems in connection with graphs," *Numberische Mathematik*, vol. 1, pp. 269–271, 1959.

[11] G. A. F. Seber, *Multivariate Observation*. John Wiley & Sons, 1984.

[12] S. Kullback, *Information Theory and Statistics*. New York: John Wiley and Sons., 1959.

[13] Y. Takane, F. W. Young, and J. D. Leeuw, "Non-metric Individual Differences Multidimensional Scaling: Alternating Least Squares with Optimal Scaling Features," *Psycometrika*, vol. 42, pp. 7–67, 1977.

[14] S. Kumar, J. Ghosh, and M. M. Crawford, "A Hierarchical Multiclassifier System for Hyperspectral Data Analysis," *Lecture Notes in Computer Science*, vol. 1857, pp. 270+, 2000.

[15] Y. Chen, M. M. Crawford, and J. Ghosh, "Integrating support vector machines in a hierarchical output decomposition framework," in *2004 International Geosci. and Remote Sens. Symposium*, Anchorage, Alaska, Sept. 20-24 2004, pp. 949–953.