Improved nonlinear manifold learning for land cover classification via intelligent landmark selection

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

Melba M. Crawford
Laboratory for Applications of Remote Sensing
Lilly Hall of Life Sciences
Purdue University
Email: mcrawford@purdue.edu

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

***Abstract*— Nonlinear manifold learning algorithms, mainly isometric feature mapping (Isomap) and local linear embedding (LLE), determine the low-dimensional embedding of the original high dimensional data by finding the geometric distances between samples. Researchers in the remote sensing community have successfully applied Isomap to hyperspectral data to extract useful information. Although results are promising, computational requirements of the local search process are exhorbitant. Landmark-Isomap, which utilizes randomly selected sample points to perform the search, mitigates these problems, but samples of some classes are located in spatially disjointed clusters in the embedded space. We propose an alternative approach to selecting landmark points which focuses on the boundaries of the clusters, rather than randomly selected points or cluster centers. The unique Isomap is evaluated by SStress, a good-of-fit measure, and reconstructed with reduced computation, which makes implementation with other classifiers plausible for large data sets. The new method is implemented and applied to Hyperion hyperspectral data collected over the Okavango Delta of Botswana.**

I. Introduction

Nonlinear manifold learning algorithms assume that the original high dimensional data actually lie on a low dimensional manifold defined by local geometric differences between samples. Recent research has demonstrated the potential of this approach for nonlinear dimension reduction and for representation of high dimensional observations through nonlinear mapping [1], [3]. Isometric feature mapping (Isomap) [12] and local linear embedding (LLE) [9] are examples of this approach. Although these methods were developed to represent high dimensional nonlinear phenomena in lower dimensional spaces, the imbedded features are potentially useful for classification of hyperspectral data.

The Isomap method was recently applied to hyperspectral data by Bachmann et al. [1] and Chen et al. [3]. While their results indicate that it can represent nonlinear information and characterize hyperspectral data in low dimensional spaces, the shortest path updating scheme creates a bottleneck in the local search process as the complexity of the inter-pixel network increases exponentially with image size. Remote sensing images typically contain a very large number of pixels, so divide-and-conquer approaches were utilized in these two papers. A tiling method that adjusted the eigenvalue/vector pairs of each image tile to form a unique map reduced computation time when Isomap was applied to large-scale images [1]. In our own previous research [3], we avoided building the global Isomap by segmenting the image and using the k-nearest neighbor (kNN) method as the base classifier. Because kNN classifies samples only according to the updated geometric distance between labeled and novel samples, a unified re-projected map is not required. A segmented map dramatically reduces the number of connected edges and prevents searches of certain nonessential shortest paths. Although the overall mean classification accuracy was high for large areas, we found that kNN is sensitive to outliers, which resulted in high standard deviations of the classification accuracies.

In order to develop a more robust classifier that still exploits the advantage of nonlinear dimension reduction and is computationally still competitive, we adopt Landmark Isomap (L-Isomap) [4] and enhance it via an intelligent landmark point selection scheme. L-Isomap is identical to Isomap, except that it uses a randomly selected subset of points to build the map. Samples that are not selected as landmarks are then placed on the map by utilizing the derived embedding vectors. Experiments indicated that if samples are equally distributed on a smooth manifold, L-Isomap is capable of reaching the same level of data compression as Isomap without losing too much information. Because spectral signatures of pixels of different land cover types are often not located in spectrally contiguous clusters, L-Isomap fails to reconstruct the low dimensional manifold when it is applied to hyperspectral data for dimension reduction. Results of our experiments show that if we assume samples lie on a manifold that has k facets and can be represented by k clusters, extreme points perform better as landmark points in preserving the manifold than random points or centers of clusters (facets).

In this paper, we construct the minimum spanning tree (MST) by using the pairwise distances between samples. We then determine the samples which define the boundaries of the clusters by recursively cutting the heaviest edge of the MST. After obtaining nonlinear manifold defined by these boundary samples, novel samples are placed on the map according to each sample's distances to its k-nearest landmark points in the original space. Details of the proposed process are presented in Section II. To show that more information is preserved by our new approach, the proportions of the manifold represented by landmark samples are measured by SStress, a goodness-of-fit measure. The unique Isomap is reconstructed with reduced computation, which makes implementation with other classifiers plausible for large data sets. The new method is implemented in conjunction with a variety of classifiers and applied to Hyperion hyperspectral data collected over the Okavango Delta of Botswana. Classification accuracies, SStress results and processing times achieved by the proposed algorithm are compared to those obtained by clustering centers and random landmark selection in Section III.

II. Methodology

Landmark-Isomap and the proposed MST-cut landmark point selection process are presented with further details in this section.

A. Landmark Isometric Feature Mapping (L-Isomap)

1) 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.

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. Isomap solves the problem efficiently via a simple, computationally efficient algorithm developed by Dijkstra [5] [1]. The process is repeated for each sample, which in turn becomes the root node, to create a shortest path network $\mathbf{D}_{stp}$.

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 [10]. Here, MDS is used to evaluate the true dimension of $\mathbf{D}_{stp}$.

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

2) L-Isomap: Experiments in [12] 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. However, although Dijkstras algorithm is efficient for finding the shortest path from a root node to the rest of the nodes, building the whole shortest path network, with an order of $O(KN^2 \log N)$, is problematic when the total number of samples, N, is large. To eliminate unnecessary calculations and speed up the shortest path search process, L-Isomap randomly selects n landmark points from the original data to construct its manifold (ref). Instead of building a $N \times N$ shortest path network, L-Isomap uses a much smaller $n \times N$ network, which requires less iterations. MDS operations are also reduced on this reduced network. Samples that are not selected for landmarks are placed on the manifold via the derived embedding vectors and their updated distances to n landmark points.

While L-Isomap is achieves little loss of information on examples that have smooth manifolds, it does not perform well on data with more complex manifolds, such as remote sensing images. This is because spectral signatures of different land cover types often belong to clusters that are not smoothly connected. We propose that samples closer to cluster boundaries are better choices for landmark points than either randomly chosen samples or cluster centers.

B. Minimum Spanning Tree and Landmark Point Selection

Let us assume that sample points lie on a manifold that has k facets and these facets can be represented by k clusters. To preserve the manifold by using a subset of the samples (landmark points), extreme points, rather than centers of clusters (facets), should be selected. For example, in order to reconstruct a human's face (a manifold), points that are close to the boundaries of each facet, such as the tip of the nose, chin and the dip between two eye bows, should be selected as landmark points instead of cluster centers. There are many ways to find such extreme points. In this paper, minimum spanning tree cut(MST-cut) is used to locate these landmark points.

Assume a connected, undirected graph $G = (N, E)$. A spanning tree is a graph that connects all the N and has no cycles. The minimal spanning tree is the tree(s) that has the lowest total cost. There are two advantages in the use of MST for finding landmark points for nonlinear imbedding: 1) Many existing algorithms can solve MST in polynomial time and provide the optimal solution; 2) MST is unique for a given graph G, unlike the shortest path tree that is unique for each root node.

Since MST is closely related to single linkage hierarchical clustering [7], partitioning by cutting the heaviest edge in the MST provides two boundary samples that come from two clusters (facets). We propose to use the set of boundary samples as landmark points to reconstruct the manifold.

Besides using landmark points on the boundaries to create the manifold, we consider the distances from a novel sample to its k-nearest landmark points. We then use that information

TABLE I

BOTSWANA DATA: AVE. SSTRESS (STD. DEV.)

Sampling Rates	Random	K-means	MST
50%	0.42(0.05)	0.39(0.04)	0.20(0)
25%	0.49(0.03)	0.47(0.03)	0.21(0)
12.5%	0.55(0.01)	0.54(0.01)	0.30(0)
10%	0.57(0.01)	0.55(0.01)	0.32(0)

to locate its position on the manifold. Such an embedding approach is widely used in the global positioning system (GPS). It not only speeds up the re-projecting process, but is also able to accommodate a manifold that has many facets and samples that are disjoint. Results of SStress and classification accuracy from multiple experiments are presented in the next section to support our finding.

III. RESULTS

The benefits of applying the new L-Isomap to hyperspectral data were evaluated in terms of manifold reconstruction 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. Manifold Reconstruction

Experiments are used to evaluate the ability of different landmark point selection methods by comparing their updated distance matrices. SStress: $ss = \left[\frac{\sum\sum_{i<j}(d_{ij}^2-\delta_{ij}^2)^2}{\sum\sum_{i<j}d_{ij}^4}\right]^{\frac{1}{2}}$ [11] is used to evaluate the similarity of $\mathbf{D}_{stp}$ and $\mathbf{D}_l$, where l includes random selection, k-means clustering centers and the proposed MST-cut approach for selecting landmarks. Results are obtained for 4 sampling rates (50%, 25%, 12.5% and 10%), each with 10 runs to compute the estimated standard deviation. The value of SStress is always between 0 and 1. Any value less than 0.2 is considered to indicate good representation. Table I shows that MST is consistently the best among these three methods in all 4 different sampling rates. Because landmark points collected by MST-cut are the same for all ten experiments, it has zero standard deviation. These results support our contention; points on the edges are better representatives than cluster centers or randomly selected points when used to reconstruct a manifold created by the whole dataset.

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. Only 50% training data are used in our experiments and 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.

These data are processed by Isomap and L-Isomap (which uses 25% of the total samples) to perform feature extraction. The original 145 bands are re-projected to its first 5 MDS bands. These data set are trained and classified by a set of classifiers to investigate how landmark points impact the overall classification accuracy.

Experiments are performed using the k-nearest neighbor, C4.5 decision tree [8], logistic regression [6] and linear kernel SVM on the low-dimensional input space determined by Isomap and L-Isomap. 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 stronger classifiers such as SVM and Logistic regression achieve higher classification accuracies than kNN and the C4.5 decision tree. The low-dimensional input space defined by Isomap gives consistently better results than that obtained by L-Isomap. Because the spatial signatures of training and testing samples came from the same distribution, it is clear that the new input space defined by landmark points (L-Isomap) can only achieve equal or lower classification accuracy than Isomap.

Classification accuracies on the spatially disjoint (SD) test set are contained in Table III. Because the spectral characteristics of the train/test data are different from that those of the SD test set, the embedding approach, which only

TABLE II

BOTSWANA TEST DATA: ACCURACY (STD. DEV.)

Training: 50%	kNN	C4.5	Logistic Reg.	SVM
Isomap	91.4(2.51)	90.7(1.79)	91.6(1.98)	92.2(1.19)
L-Isomap	88.6(1.54)	87.4(1.48)	87.8(2.59)	89.1(1.31)

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

Training: 50%	kNN	C4.5	Logistic Reg.	SVM
Isomap	77.5(2.13)	76.9(3.32)	79.0(2.39)	80.7(2.89)
L-Isomap	76.9(1.79)	76.4(1.67)	79.5(2.75)	78.2(1.79)

considers distances from points to their k-nearest landmark points, helps narrow the accuracy gap between Isomap and the proposed mst-cut L-Isomap. The proposed L-Isomap even provides higher accuracy than Isomap achieves when trained and classified by logistic regression classifier. Similar to results from our test set, stronger classifiers do well in these experiments. This contradicts our previous finding in [2] that SVM is better able to adapt to signature changes in this data set. The reason is that the nonlinear manifold reshaped the input space to adapt to changes. It takes over the updating part and gives SVM a chances to perform. Since the processing time is reduced, it becomes plausible to apply Isomap as a preprocessor for dimension reduction and then apply classifiers other than the k-NN to the new data set. For a Botswana experiment that has 790 samples, 9 classes and 145 feature spaces, using a 3GHz Pentium 4 CPU machine, Isomap feature extraction requires 139 seconds of CPU time while L-Isomap reduces that to 40 seconds. Time spent on training and testing are the same for both inputs with kNN requiring 9 seconds, C4.5 needing 19 seconds, logistic regression taking 45 seconds and linear-SVM requiring 31 seconds.

IV. CONCLUSION

In this paper, we investigated the concept of L-Isomap and its advantages and weakness when applied to hyperspectral data. Evaluations of dimension reduction and representation of high dimensional observation by Isomap and L-Isomap are conducted. This study also included an investigation of L-Isomap in conjunction with classification of hyperspectral data. The proposed MST-cut landmark selection approach was compared to random selection and k-means cluster centers.

Two conclusions emerge from our experiments. First, samples on the cluster boundaries are better representatives that preserve the manifold created by the whole sample set. Thus, these landmark points can be used initially to construct a similar manifold with significantly less processing time. Second, the new embedding process makes applying nonlinear manifold learning on large scale data possible. Stronger classifiers such as SVM can be applied to the modified data set to achieve both high classification accuracy and robustness.

Applying nonlinear manifold learning to hyperspectral data provided promising initial results. Future studies will involve investigation of alternative distance measures and alternative methods to identify boundary samples. Approaches for incorporating neighborhood information will be explored to take advantage of closeness in both spectral signatures and spatial information.

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] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Exploiting manifold geometry in hyperspectral imagery. *IEEE Trans. Geosci. and Remote Sens.*, 43(3):441–454, Mar 2005.
[2] 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*, pages 949–953, Anchorage, Alaska, Sept. 20-24 2004.
[3] Y. Chen, M. M. Crawford, and J. Ghosh. Applying nonlinear manifold learning to hyperspectral data for land cover classification. In *2005 International Geosci. and Remote Sens. Symposium*, Seoul, South Korea, Jul. 24-29 2005.
[4] V. de Silva and J. B. Tenebaum. *Advances in Neural Information Processing System*, chapter Global versus local methods in nonlinear dimensionality reduction, pages 705–712. MIT Press, 2002.
[5] E. W. Dijkstra. Note on two problems in connection with graphs. *Numberische Mathematik*, 1:269–271, 1959.
[6] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University, 1998.
[7] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.
[8] J. R. Quinlan. Induction of Decision Trees. 1990.
[9] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by local linear embedding. *Science*, 290(5500):2323–2326, 2000.
[10] G. A. F. Seber. *Multivariate Observation*. John Wiley & Sons, 1984.
[11] Y. Takane, F. W. Young, and J. De Leeuw. Non-metric Individual Differences Multidimensional Scaling: Alternating Least Squares with Optimal Scaling Features. *Psycometrika*, 42:7–67, 1977.
[12] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.