Non-linear Dimensionality Reduction by Locally Linear Isomaps

Ashutosh Saxena[1], Abhinav Gupta[2], and Amitabha Mukerjee[2]

[1] Department of Electrical Engineering,
Indian Institute of Technology Kanpur, Kanpur 208016, India
Ashutosh.Saxena@ieee.org
[2] Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur, Kanpur 208016, India
{Abhigupt,Amit}@cse.iitk.ac.in

Abstract. Algorithms for nonlinear dimensionality reduction (NLDR) find meaningful hidden low-dimensional structures in a high-dimensional space. Current algorithms for NLDR are Isomaps, Local Linear Embedding and Laplacian Eigenmaps. Isomaps are able to reliably recover low-dimensional nonlinear structures in high-dimensional data sets, but suffer from the problem of short-circuiting, which occurs when the neighborhood distance is larger than the distance between the folds in the manifolds. We propose a new variant of Isomap algorithm based on local linear properties of manifolds to increase its robustness to short-circuiting. We demonstrate that the proposed algorithm works better than Isomap algorithm for normal, noisy and sparse data sets.

1 Introduction

Nonlinear dimensionality reduction involves finding low-dimensional structures in high-dimensional space. This problem arises when analyzing high-dimensional data like human faces, speech waveforms, handwritten characters and natural language. Previous algorithms like Principal Component Analysis, Multidimensional scaling and Independent Component Analysis fail to capture the hidden non-linear representation of the data [1, 2]. These algorithms are designed to operate when the manifold is embedded almost linearly in the high-dimensional space. There are two approaches to solve this problem: Global (Isomaps [3, 4]) and Local (Local Linear Embedding [5] and Laplacian Eigenmaps [6]).

Tenenbaum [3] describes an approach that uses easily measured local metric information to learn the underlying global geometry of a data set based on isomaps. It attempts to preserve geometry at all scales, by mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points. Since, the algorithm aims to find correct geodesic distances by approximating them with a series of euclidean distances between neighborhood points, it gives correct representation of the data's global structure.

Local approaches (LLE [5] and Laplacian Eigenmaps [6]) try to preserve the local geometry of the data. By approximating each point on the manifold with a

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 1038–1043, 2004.

linear combination of its neighbors, and then using the same weights to compute a low-dimensional embedding. LLE tries to map nearby points on the manifold to nearby points in the low-dimensional representation. In general, local approaches are computationally efficient, but Landmark Isomaps [4] achieve computational efficiency equal to or in excess of existing local approaches. Local approaches have good representational capacity, for a broader range of manifolds, whose local geometry is close to Euclidean, but whose global geometry may not be. Conformal Isomaps [4], an extension of Isomaps, are capable of learning the structure of certain curved manifolds. However, Isomap's performance exceeds the performance of LLE, specially when the data is sparse.

In presence of noise or when the data is sparsely sampled, short-circuit edges pose a threat to Isomaps and LLE algorithms [7]. Short-circuit edges occur when the folds in the manifolds come close, such that the distance between the folds of the manifolds is less than the distance from the neighbors. In this paper, we propose an algorithm which increases the robustness of the Isomaps. Locally Linear Isomaps (LL-Isomaps), a variant of Isomaps are proposed which use the local linearity properties of the manifold to choose neighborhood for each data point. We demonstrate that the proposed algorithm works better than Isomap algorithm for normal, noisy and sparse data sets.

In Section 2 we discuss Tenenbaum's approach using Isomaps and the Roweis approach using Local Linearity to solve the problem. The proposed algorithm has been described in Section 3. In Section 4, results are discussed, followed by conclusion in Section 5.

2 Current Approaches

2.1 Isometric Feature Mapping (Isomaps)

NLDR algorithm reduces the dimensionality of high-dimensional data and hence only the local structure is preserved. This implies that the euclidean distance is meaningful between the nearby points only. Tenenbaum et. al [3] proposed an algorithm that measures the distance between two far-away points on the manifold (called the geodesic distance) and tries to obtain a low-dimensional embedding using these distances.

Isomap algorithm can be described in three steps:

1. Neighbors of each point are determined. The neighbors are chosen as points which are within the ϵ distance or using K-nearest neighbor approach. These neighborhood relations are represented as a weighted graph G over the data points, with edges of weight $d_X(i, j)$ between neighboring points.
2. Isomap estimates the geodesic distances $d_M(i, j)$ between all pairs of points on the manifold M by computing their shortest path distances $d_G(i, j)$ in the graph G. The shortest path can be found by using Floyd-Warshall's algorithm or Dijasktra algorithm.
3. Reduce the dimensionality of the data by using MDS algorithm on the computed shortest path distance matrix.

The residual error of the MDS algorithm determines the performance of the Isomap algorithm. A zero error implies that the computation of the geodesic distance was correct. The dimensionality of a manifold is determined by decrease in the error as dimensionality of low-dimensional embedding vectors Y is increased. The correct low-dimensional embedding is obtained when the error goes below a certain threshold.

2.2 Local Linear Embedding

The LLE algorithm proposed by Roweis et. al [5] uses the fact that the data point and its neighbors lie on a linear patch whose local geometry is characterized by linear coefficients that construct the point. This characterization is also valid in lower dimensions. Suppose the data consist of N real-valued vectors X_i, each of dimensionality n, sampled from some underlying manifold and let Y_i represent global internal coordinates on the manifold (coordinates in low-dimensional space). The algorithm can described in three steps below

1. Assign neighbors to each data point $\boldsymbol{X_i}$ using K-nn approach.
2. Compute the weights W_{ij} that best linearly reconstruct $\boldsymbol{X_i}$ from its neighbors.
3. Compute the low-dimensional embedding vectors $\boldsymbol{Y_i}$ best reconstructed by W_{ij}.

3 Proposed Algorithm ($\boldsymbol{K_{LL}}$ Isomaps)

There is a serious problem in the Isomap algorithm which is referred to as Short Circuiting [7]. When the distance between the folds is very less or there is noise such that a point from a different fold is chosen to be a neighbor of the point, the distance computed does not represent geodesic distance and hence the algorithm fails (Fig. 2(a)).

We propose an algorithm which uses local linearity property of the manifolds to determine the neighborhood of a data point as opposed to using a K-nearest neighbor or ϵ-map. This results in a better neighborhood of the point, which in turn gives lower residual variance and robustness. The problem with the previous algorithms is that they consider only the distance for determining the neighborhood and they fail when the folds of manifold come close to each other. This approach not only overcomes the problem of short-circuiting but also produces better estimates of geodesic distances and hence the residual error is less than the Tenenbaum's algorithm.

The proposed algorithm first finds a candidate neighborhood using K-nearest neighbor (K-nn) approach. The linear combination of the candidate neighbors are used to reconstruct the data-point. The weight for each neighbor can be estimated by reducing the reconstruction error:

$$\epsilon(W) = \sum \left| X_i - \sum_{j \neq i} W_{ij} X_j \right|^2 \tag{1}$$

Now $K_{LL} \leq K$ neighbors are chosen based on the values of reconstruction weights. The neighbors whose Euclidean distance is less and those lying on the locally linear patch of the manifold get higher weights, and hence are selected preferably. These K_{LL} (same for every point) neighbors are used in the rest of the Isomap algorithm (Section 2.1) to calculate geodesic distances and the low-dimensinal embedding. The proposed algorithm has two parameters $\{K, K_{LL}\}$.

4 Results and Discussion

We compare the results of our algorithm for following classes of data:

1. Sparsely Sampled data
2. Noisy Data
3. Dense data without noise

Swiss-roll data set (n=3, d=2) and Synthetic face data set (n=4096, d=3) [3] were used. The quality metric for comparing the proposed algorithm with the Tenenbaum's algorithm is the residual variance at the expected Manifold dimension d. We show that the proposed algorithm not only overcomes the problem of short-circuiting but also gives less residual variance.

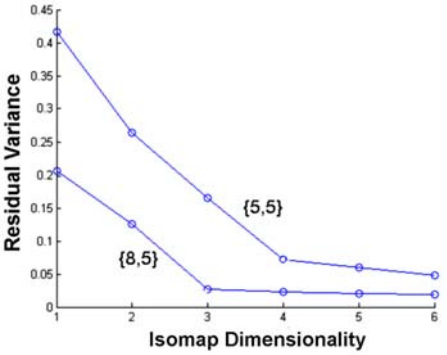
In sparsely sampled data sets, the Euclidean distance between points in neighborhood becomes larger as compared to the distance between different folds of the manifold. Tenenbaum's algorithm either faces the problem of short-circuiting (Fig. 1(a)), or has to choose a very low value of K, which gives a large residual variance. The proposed algorithm uses the same number of neighbors (=5) as Tenenbaum's algorithm, but is able to find the correct dimensionality (=3). Fig. 1(b) shows plot of residual variance of proposed K_{LL}-Isomaps and Tenenbaum's K-Isomaps, using best performances of both. The proposed algorithm works with higher value of K_{LL}, even with sparse data, hence gives a much lower value of residual variance. In worst case, when $K_{LL} = K$, the proposed algorithm performs as good as Tenenbaum's algorithm.

Additive-White-Gaussian-Noise (AWGN) with SNR of 10 dB is added to the original Swiss-roll data with 1000 sample points. Tenenbaum's algorithm with $K = 6$ (which works with noiseless data) fails due to short-circuit edges (Fig. 2(a)). Using the proposed method, the problem of short-circuiting is easily removed as shown in Fig. 2(b), and the correct low-dimensional embedding is found more robustly with $K_{LL} = 6$. For noisy data, Tenenbaum's algorithm has to choose a lower value of K to avoid short-circuiting. In Fig. 3(b), the best possible results with Tenenbaum ($K = 5$), and our algorithm ($K = \{7, 4\}$) are shown. It can be seen that our algorithm out-performs Tenenbaum's algorithm.

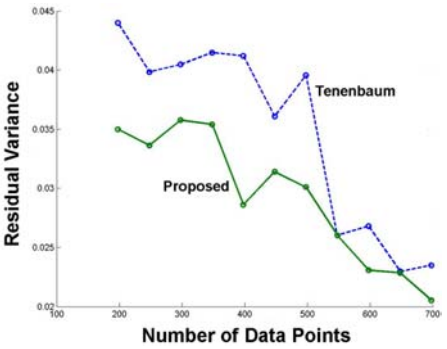
Even for dense Synthetic Face data (without noise), our algorithm gives better residual variance as compared to Tenenbaum's algorithm (Fig. 3(a)).

5 Conclusion

The Isomap algorithm, with its broad appeal, opened up new frontiers by its various applications; but was not robust to short-circuiting, resulting in drastically

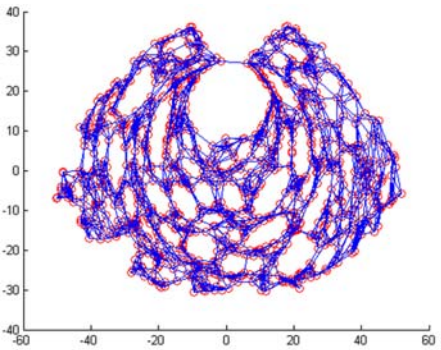


(a) Tenenbaum v/s Proposed (Sparse Face Dataset)

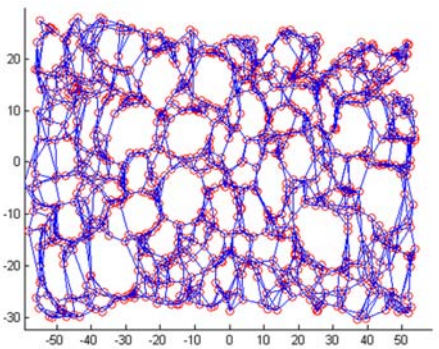


(b) Sparseness v/s Residual Variance (Face Dataset)

Fig. 1. Comparison of Tenenbaum's algorithm with the proposed algorithm. (a) For $N = 349$, Tenenbaum's algorithm is represented by {5,5}, and it predicts the manifold dimensionality to be 4 because of a short-circuit edge. This problem can be overcome by reducing K but this leads to a high residual variance. Proposed K_{LL}-Isomap gives smaller error for the *same* number of neighbors and the dimensionality is correctly predicted to be 3. (b) Comparison of Tenenbaum's Isomap with K_{LL}-Isomap for varying level of sparseness. The number of sample data-points was varied and the error in both the algorithms (with their best case) was computed. The K_{LL}-Isomap outperforms Isomap in all the cases, except two where the errors are same in both the algorithms.



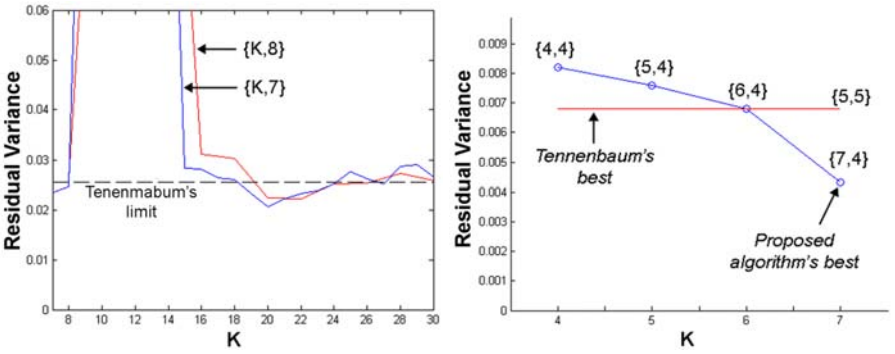
(a) Tenenbaum's Isomap



(b) Proposed K_{LL}-Isomap

Fig. 2. Noisy swiss roll embeddings in two dimensions as obtained by Tenenbaum's Isomap and K_{LL}-Isomaps. The swiss roll dataset consisted of 1000 points. The Isomap algorithm had a short-circuit edge and hence gave incorrect embedding.

different (and incorrect) low-dimensional embedding. We proposed a new variant of Isomaps based on local linearity properties of the manifolds to increase the its robustness to short-circuiting. We demonstrated that the proposed algorithm works better than Isomap algorithm for normal, noisy and sparse data sets.



(a) Dense Face data without noise (b) Noisy swiss roll data

Fig. 3. Accuracy as a function of parameter K. Increasing K gives more choice to the proposed algorithm to choose K_{LL} neighbors on the basis of weights and hence the performance improves (Residual Error decreases).

References

1. Murase, H., Nayar, S.: Visual learning and recognition of 3d objects from appearance. International Journal Computer Vision **14** (1995)
2. J.W. McClurkin, L.M. Optican, B.R., Gawne, T.: Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. Science **253** (1991) 675–657
3. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science **290** (2000) 2319–2323
4. Silva, V.d., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S.T., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15. MIT Press, Cambridge, MA (2003) 705–712
5. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science **290** (2000) 2323–2326
6. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002)
7. Balasubramanian, M., Schwartz, E.L., Tenenbaum, J.B., Silva, V.d., Langford, J.C.: The Isomap Algorithm and Topological Stability. Science **295** (2002) 7a