1 Paper Title

A Global Geometric Framework for Nonlinear Dimensionality Reduction

2 Summary

Real-world data from biological and physiological sources doesn't necessarily follow a linear structure. Dimensionality reduction techniques like PCA and MDS assume that the true structure of the data is inherently linear or lies on a linear subspace in a high dimensional space. They are able to learn the structure of the data if it is on a linear subspace of the higher dimension, but finds difficult to correctly represent data if some non-linear structures are involved. This paper builds on the classical techniques like PCA and MDS for dimensionality reduction. This technique not only is able to learn non-linear representations, but also get a globally optimal solution for the same.

3 Detailed Analysis

The major contribution of the paper[3] is proposing a technique called Isometric Feature Mapping or Isomap which aims to find a low dimensional representation of our data that is lying in a manifold embedded in a high-dimensional space. The representation is recovered using classical MDS algorithm, but with a twist. Since the data lies on a manifold, Euclidean distance cannot represent the distance between points correctly. From Fig 1, we see that the distance between the two circles is larger in reality than when calculated using Euclidean metrics.

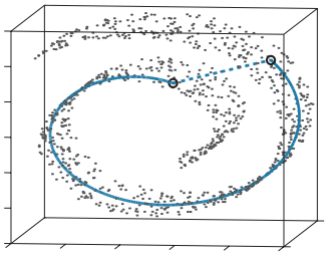

Figure 1: Points on a manifold

The distance in the high-dimensional space is actually the geodesic distance along the manifold on which the data lies. Isomap algorithm uses the geodesic distance to calculate the embeddings.To approximate the geodesic distance, a clever technique is used which will be discussed in the following points:
Steps of the algorithm:

1. Find the neighbourhood points and construct a grpah of all the points in the data.

   This can be done using either K-nearest neighbour method and using Euclidean distances for finding the nearest neighbours. Note that a manifold can be approximated as a Euclidean space in a small patch, hence this works. Or epsilon-ball method can be used to find points within a ball of epsilon radius around the point whose neighbours to calculate.

   After getting the distances, a weighted graph can be constructed over all the data points with the Euclidean distances as weights.

2. In the second step, this algorithm finds the geodesic distances between all the points by finding the shortest distance in the graph between said points. The result is a distance matrix with distances between all pairs of points.

3. In the third and last step, Isomap uses classical MDS to the distance matrix obtained in step 2 to calculate the embeddings in a k-dimensional space preserving the geometry of the manifold (because we used geodesic distances).

Since we want to find embeddings y_i in a Euclidean space Y, we wish to minimize E given as $E = ||\tau(D_G) - \tau(D_Y)||^2_L$ where, D_G is the geodesic distance matrix, D_Y is the Euclidean distance matrix, and τ is converting distances to inner products acting as an efficient optimization operator.

The authors mention that the Isomap algorithm is guaranteed to recover the true dimensionality of the data. The residual variance (or error) continues to decrease as dimension of Y is increased and stops after a certain dimension of Y. The property of convergence comes from a theorem that as the number of data points increases, the graph distances are increasingly better approximations of the actual geodesic distances.
Another major contribution is the myriad of real-world examples the athors demonstrate - capturing illumination and pose in images (2-d), hand gesture images, and patterns in handwritten digits, even data that has no clear manifold structure. Isomap is able to capture interpolations between data points that are distant (but visually make sense seen from a human's point of view), which is a feat in itself.

4 Your critiques

4.1 Assumptions made

The authors assume that the K in nearest neighbours algorithm is some value, but don't address it in the paper. The value of K will affect the distance matrix and in turn the geodesic distances. The algorithm converges when the number of data points is large, but real world data might not be large. Also, approximating geodesic distances by the shortest path euclidean distance is an assumtion which might break.

4.2 Pros and cons

Isomap works brilliantly is cases where the manifold is convex and has been applied to many real world scenarios[4]. It is capable of discovering global structures in the data. Also, this algorithm uses few free parameters and guarantees to converge to the real dimensionality given enough data points is something that's lacking in other non-linear dimensionality reduction algorithms.
But this algorithm has some cons too. First of all, it might be a little hard to understand given it needs shortest path finding algorithms in the second step. Also, the distance matrix has distances from all points to all points, and is non-sparse which makes the algorithm computationally expensive O(n^3) (dense matrix eigen-reduction). Also, Isomap suffers from topological instability and the choice of neighbourhood can make or break the algorithm [1]. Also, sometimes the neighbourhood approach may connect points that are far apart due to large K or epsilon, which leads to something known as short-circuiting. This can lead to erroneous representations and derail the entire algorithm. To conclude, if the manifold is non-convex, has holes or has a structure that can be easily broken by choosing large epsilon or K, this algorithm will give erroneous results.

4.3 Possible future extensions

Future extensions include algorithms that will overcome the short-circuiting problem by checking which neighbours are in a locally Euclidean space and which violate that assumption.
Also, algorithms that can approximate a non-convex manifold by combining local techniques with Isomap could be a possible extension to including non-convex manifolds because local algorithms like LLE can handle a certain amount of curvature which global approaches like Isomap can't [2].

References

[1] BALASUBRAMANIAN, M., AND SCHWARTZ, E. L. The isomap algorithm and topological stability. *Science 295*, 5552 (2002), 7–7.

[2] SILVA, V. D., AND TENENBAUM, J. B. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2003, pp. 721–728.

[3] TENENBAUM, J. B., SILVA, V., AND LANGFORD, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 5500 (2000), 2319–2323.

[4] VAN DER MAATEN, L., POSTMA, E., AND VAN DEN HERIK, J. Dimensionality reduction: a comparative review. *J Mach Learn Res 10* (2009), 66–71.

[4] [1] [2] [3]