Similarity Consideration for Visualization and Manifold Geometry Preservation

Shashwati Mishra
School of Computer Engineering
KIIT University
Bhubaneswar, India

Chittaranjan Pradhan
School of Computer Engineering
KIIT University
Bhubaneswar, India

ABSTRACT

Manifold learning techniques are used to preserve the original geometry of dataset after reduction by preserving the distance among data points. MDS (Multidimensional Scaling), ISOMAP (Isometric Feature Mapping), LLE (Locally Linear Embedding) are some of the geometrical structure preserving dimension reduction methods. In this paper, we have compared MDS and ISOMAP and considered similarity as an approach to find the reduced representation of original data using ISOMAP.

General Terms

Manifold preservation in data visualization

Keywords

Manifold learning technique; mds; isomap; geodesic distance; euclidean distance

1. INTRODUCTION

To overcome the problems of mining and data analysis on huge amount of data different data reduction techniques are used. These techniques obtain a reduced representation of the data that maintains the integrity of the original data. Dimensionality reduction is one such data reduction and feature selection technique which reduce the original data keeping as much original information as possible and is used both for data reduction and visualization process. Main goal of such techniques is to discover the compact representation of data with reduced computational time and to solve the problem of curse of dimensionality of high-dimensional spaces [9].

Visualization refers to the graphical representation of the information present in the data. But visualizing high dimensional data after conversion to lower dimension is one of the important problems in data mining. For highly twisted and folded manifold, preserving the manifold structure creates problem. Due to explicit control of information during reduction process information loss also occurs.

The main problems are:

- Unknown intrinsic dimensionality, which means no effective way to determine the number of independent variables that satisfactorily represent the phenomena.
- Nonlinear relationships among data make the process complicated.
- Unknown relevance of information. Lossless dimension reduction is the ideal one, but often it is not possible to reduce the dimension without loss of information.

A large number of methods are developed to handle such complex problems.

1.1. Manifold Learning Techniques

Manifold learning techniques are used to convert the higher dimensional data set to lower dimension, preserving the local geometry on the manifold as much as possible. MDS(Multi-Dimensional Scaling), PCA(Principal Component Analysis)[9], ISOMAP(Isometric Feature Mapping), LLE, Hessian LLE, Laplacian Eigenmap, Diffusion maps are some of the manifold learning techniques. [2, 8].

PCA (Principal Component Analysis), MDS (Multi-Dimensional Scaling) basically works well when the data is linear in nature. Other techniques are unsupervised nonlinear techniques which try to preserve the manifold and are known as non-linear manifold learning techniques [2].

The basic idea behind nonlinear manifold learning algorithms is that the original high dimensional data actually lie on a low dimensional manifold which is the difference of local geometry between the samples. This is the cause of the development of nonlinear dimension reduction methods and representation of high dimensional observations through nonlinear mapping [1]. Some of the non-linear techniques of dimensionality reduction concentrate on preserving the geodetic distances.

1.2. Geodetic Distances Preserving Methods

Most commonly used distance measure for continuous data is the Euclidean distance. This distance depends only on the value of the point coordinates and equals to the straight line segment joining the two points. But this distance does not focus on data manifold. Distances along the straight line path between two points may not be equal to the distances along a geodesic in the manifold. So considering Euclidean distance as an input to MDS gives incorrect result. To obtain the correct result geodetic distance is used instead of Euclidean distance. The geodetic distance between two points of a manifold is the minimum of the length of a path joining both points that is contained in the manifold. These paths having minimum lengths are called geodesics [3].

2. MDS

MDS is an information visualization technique which considers proximity among data values for obtaining the reduced representation of original data. Proximity refers to the similarity and dissimilarity found among the examined objects. Matrix which contains such information is called proximity matrix. MDS takes proximity matrix as input which are in the form of distances and finds the corresponding coordinate values embedded in the low dimensional space [7].

Steps involved in classical MDS algorithm are:-

- Set up the matrix of squared proximities $P^{(2)}=[p^2]$.
- Apply the double centering: $B=-\frac{1}{2} J P^{(2)} J$ using the matrix $J= I - n^{-1}11'$, where n is the number of objects.
- Extract the m largest positive eigenvalues $\lambda_1 \ldots\ldots \lambda_m$ of B and the corresponding m eigenvectors $e_1 \ldots. e_m$.
- A m-dimensional spatial configuration of the n objects is derived from the coordinate matrix $X= E_m \Lambda_m^{(1/2)}$,where E_m is the matrix of m eigenvectors and Λ_m is the diagonal matrix of m eigenvalues of B, respectively[7].

3. ISOMAP

Tenenbaum proposed a method called ISOMAP which was a combination of topology-preserving network algorithm and multidimensional scaling for manifold modeling. Later in 2000 he presented a variation of the previous ISOMAP. In the later version of ISOMAP, for constructing topology preserving network every data point is linked to its K-nearest neighbouring data points or to points within Ɛ distance [3].

ISOMAP (Isometric Feature Mapping) is a nonlinear generalization of Classical MDS, which works well both for real world and artificial data. From statistics point of view ISOMAP is one of widely used low dimensional embedding methods where geodesic distances imposed on a weighted graph are incorporated with classical MDS.

The algorithm given by Tenenbaum involves three steps, which takes distance matrix measured, either in the standard Euclidean metric or in some domain-specific metric as input and gives those coordinate vectors as output that best represents the geometry of the data [4].

- *Construction of neighborhood graph:* Two points i and j are connected if they are closer than Ɛ or i is one of the k-nearest neighbors of j. This process is repeated for all points and a graph called neighborhood graph is constructed. Edge lengths are set equal to the distance between the two end points. If i and j are two points then $d_x(i, j)$ represents the distance between i and j and set as the length of the edge connecting points i and j.

- *Shortest path computation:* $d_G(i, j)$ is initialized as $d_x(i, j)$ if i and j are connected by an edge otherwise $d_G(i, j)$ is set to ∞. Then shortest path algorithm is applied to find shortest path distances between all pairs of points in G.

- *Construction of d-dimensional embedding:* Classical MDS is used for constructing lower dimensional embedding [4,5,6,8].

4. COMPARISON OF OUTPUTS OF MDS AND ISOMAP

For experimental analysis we have applied MDS and ISOMAP algorithms on Swiss-roll data set.

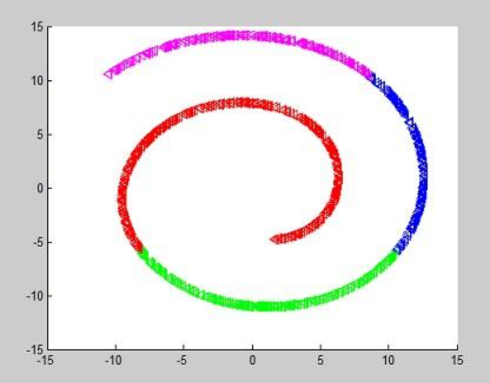

Figure 1. 1000 data points from Swiss-roll data set

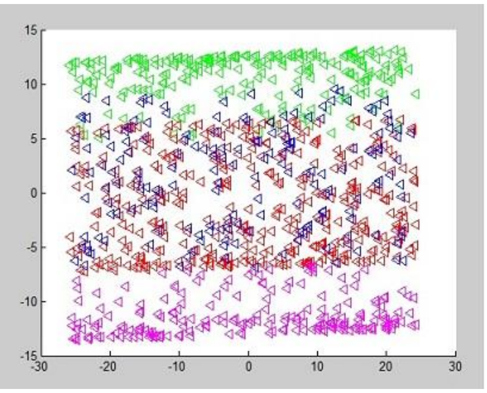

Figure 2. Output obtained by applying MDS algorithm on 1000 data points of Swiss-roll data set

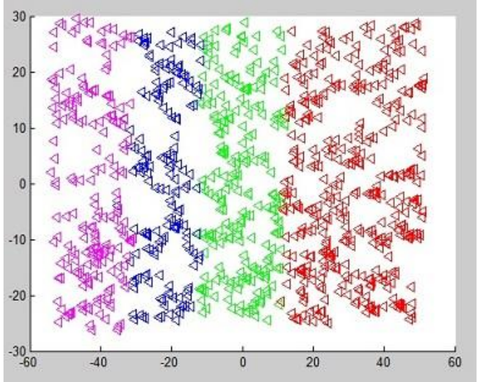

Figure 3. Output obtained by applying ISOMAP algorithm on 1000 data points of Swiss-roll data set

MDS concentrates on Euclidean distance where as ISOMAP focuses on geodetic distance. In case of a curved, twisted manifold MDS does not preserve the original geometry and reduced representation shows overlapping of data points. But ISOMAP takes geodetic distances into consideration and preserve the original manifold structure after reduction. This can be verified from above outputs.

5. RESULTS OBTAINED BY CHANGING NEIGHBORHOOD SIZE (K)

With variation of neighborhood size output of ISOMAP also varies. So, selecting a suitable neighborhood size is a very difficult task.

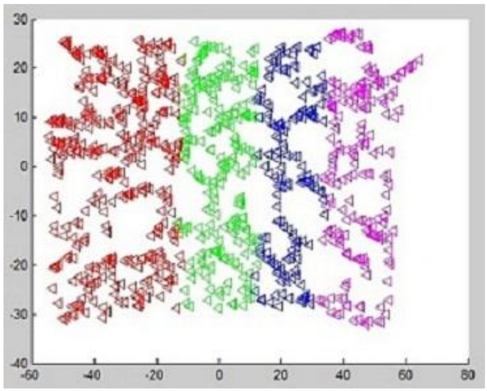

Figure 4. Output obtained by applying ISOMAP algorithm on 1000 data points of Swiss-roll data set with neighborhood size 5

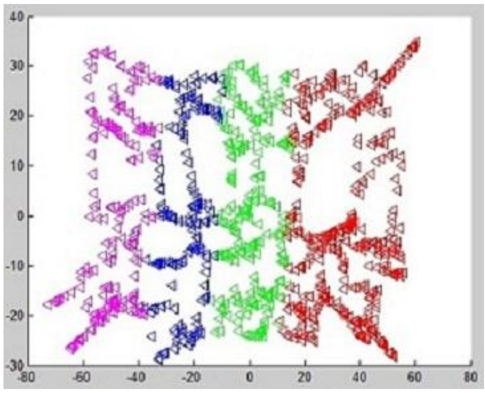

Figure 5. Output obtained by applying ISOMAP algorithm on 1000 data points of Swiss-roll data set with neighborhood size 4

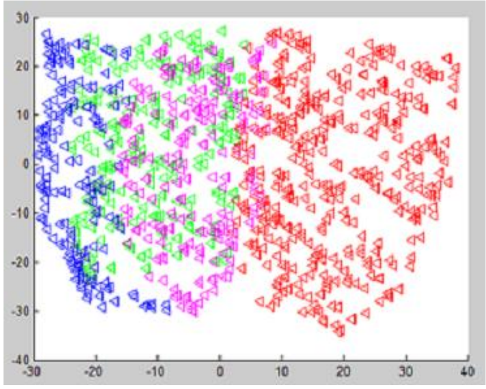

Figure 6. Output obtained by applying ISOMAP algorithm on 1000 data points of Swiss-roll data set with neighborhood size 9

From the above outputs we can conclude that with the change in number of nearest neighbors output of ISOMAP also changes. With decrease in neighborhood size information loss occurs and with increase in neighborhood size overlapping of information takes place.

6. SIMILARITY APPROACH

Though ISOMAP works well on high-dimensional data sets, still it has some drawbacks.

These are:-

- Since it uses MDS for low dimensional embedding and MDS is slow so ISOMAP is also slow.

- For different values of K we get different outputs. Determining the optimal value for K is very difficult.

- It assumes the data set as convex and does not handle non-convex data sets.

- ISOMAP removes outliers in preprocessing, so it is extra sensitive to noise.

Though it has so many negative points, it is popularly used for high dimensional data sets due to its quality of output as compared to other methods [2].

In this paper we have considered the similarity among data values for constructing the neighborhood graph, instead of using the concept of k-nearest neighbor. For constructing the similarity matrix we have used the distance matrix as input. The values of the similarity matrix will always be with in the range 0 and 1. Here we have used the average of the similarity and the difference between maximum and average similarity to decide the neighbors of a particular data point.

Our modified method is:

- *Construction of neighborhood graph:* From the distance matrix we calculate the similarity matrix by using the formula

$$\text{Similarity}=1/(1+\text{Distance}) \qquad (1)$$

Thus we obtain the similarity matrix which contains values with in 0 and 1 having all diagonal elements as 1. Then we find the average of all values except the diagonal elements. Suppose this value is stored in variable avg. Difference of maximum similarity and the calculated average is x.

$$x=\text{maximum similarity}-\text{avg} \qquad (2)$$

To reduce the value after the decimal point we divide the difference by n.

$$x=x/n \qquad (3)$$

For obtaining better output we have to vary the value of n.

Then set the limit d to find the neighbors as the sum of average and x.

$$d=\text{avg}+x \qquad (4)$$

- *Shortest path computation:* d_G (i, j) is initialized as d_x(i,j) (distance between i and j) if the similarity value is greater than d otherwise d_G (i, j) is set to ∞. Then shortest path algorithm is applied to find shortest path distances between all pairs of points in G.

- *Construction of d-dimensional embedding:* Classical MDS is used for constructing lower dimensional embedding.

7. EXPERIMENTAL RESULT OF CONSIDERING SIMILARITY

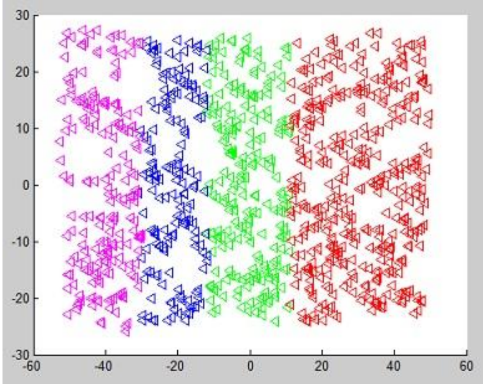

Figure 7. Output obtained by considering similarity in ISOMAP algorithm on 1000 data points of of Swiss-roll data set

For obtaining the above output we have considered n as 8. This value of n gives best output for a variation of data set size from 600 to 1000. The result obtained by considering similarity (figure7) is similar to that obtained by considering distance among data points (figure 3).

8. CONCLUSION

In this paper, we have compared MDS and ISOMAP algorithms and considered similarity among data points as a measure of deciding the neighbors of a data point instead of considering neighborhood size. This, some way, helps to reduce the size of variance, because the size of similarity matrix is limited between 0 and 1. Time complexity of our algorithm is nearly equal to that of the original version of ISOMAP. The proposed process can be considered as another approach of creating neighborhood graph for the purpose of solving visualization problem using ISOMAP.

Our future work is to extend the approach of obtaining reduced representation and manifold geometry preservation for the manifold where a gap is present between two parts.

9. REFERENCES

[1] Chen, Y., Crawford, M. M., Ghosh, J. 2006, “Improved nonlinear manifold learning for land cover classification via intelligent landmark selection”, Geoscience and Remote Sensing Symposium. IEEE,545-548.

[2] Wittman, T. 2005, “Manifold LearningTechniques: So which is the best?”.

[3] Carreira-Perpiñán, M. Á. 2001, “Contious latent variable models for dimensionality reduction and sequential data reconstruction”, Submitted to the University of Sheffield for the degree of Doctor of Philosophy.

[4] Tenenbaum, J. B., Silva, V. de, Langford, J. C. 2000, “A Global Geometric Framework for Non-linear Dimensionality Reduction” Science, 290(5500):2319-2323.

[5] Wu, Y., Chan, K. L. 2004, “An Extended Isomap Algorithm for Learning Multi-Class Manifold”, Machine Learning and Cybernetics, vol. 6, 3429-3433, IEEE.

[6] Yang, M. 2002, “Extended Isomap for Pattern Classification”, Eighteenth National Conference on Artificial Intelligence.

[7] Wickelmaier, F. 2003, “An Introduction to MDS”, Sound Quality Research Unit,Aalborg University,Denmark.

[8] Samko, O., Marshall, A. D., Rosin, P. L. 2006, “Selection of the optimal parameter value for the Isomap algorithm”, Pattern Recognition Letters,Vol. 27, Issue-9, Pages 968-979.

[9] Fodor, I. K.. 2002, “A survey of dimension reduction techniques”, Center for Applied Scientific Computing, Lawrence Livermore National Scientific Computing, Citeseerx.