http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 8651-8656, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.8651.8656

Improved Locally Linear Embedding Method Suitable for Multi-dimensional Visualization of Economic Statistics

[1]Rui Xiao-Ping, [2]Dong Cheng-Wei and [1]Song Xian-feng
[1]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Beijing Institute of Surveying and Mapping, Beijing, 100038, China

Abstract: Manifold learning is a nonlinear dimension reduction method that is increasingly capturing researchers' interests. It can be used to discover the intrinsic structure of sample data. This study discusses the application of the Locally Linear Embedding (LLE) method on dimensionality reduction for sparse and non-uniform economic statistics data. Aiming at the sparse and uneven distribution of the characteristics of the statistical sample data, this study develops an Adaptive LLE (ALLE) method and adopts a new distance metric and neighbor value selection method for calculating the neighbor points. The improved methods are employed in an experiment that analyzes the 2007 Sichuan statistical data from China. By visualizing the dimension reduction data, the experiment shows that the improved method can retain more information and reveal the distribution of regional economic situation more accurately when a suitable nearest neighbor value is selected. By comparing the results from ALLE when different nearest neighbor values are used, it is shown that the classifications still depend on the selected neighbors. Finally, this paper explains the results calculated using different target dimensions and shows that the results can reflect the economic status of the Sichuan district better when the intrinsic dimensionality is selected optimally.

Key words: Nonlinear dimension reduction, locally linear embedding, multi-dimensional visualization, intrinsic dimensionality

INTRODUCTION

With the rapid development of the information era, the amount of accumulated information in various fields is exploding. In the field of economic statistics, it is often necessary to deal with large amounts of high-dimensional data. Given the inherent limitations of human cognitive abilities, the current challenges that are associated with finding and extracting information and knowledge from massive multi-dimensional databases are unprecedented. Therefore, multidimensional information visualization techniques are an effective tool for expressing abstract information during the process of knowledge discovery, information cognizing and decision-making. These techniques are very popular and can be used to support researchers during the process of understanding and analyzing the inner structure of multidimensional datasets (Parmanto *et al*., 2008). Humans can easily visualize the information in a three-dimensional space. The characteristics of high-dimensional data, on the other hand, are not easy to understand or perceive. This can easily lead to the problem of the "curse of dimensionality." In order to study the characteristics and inherent structures hidden in high-dimensional data, we will use dimensionality reduction methods to reduce the number of dimensions to a low value and then we will use visualization methods to study the features in multi-dimensional information.

Dimensionality reduction techniques use linear or nonlinear methods to transform the high-dimensional observational data into a low-dimensional feature space and thereby attempting to find hidden meaningful low-dimensional structures in the high-dimensional observation space. Linear dimensionality reduction can project high-dimensional data onto a low dimensional linear subspace. The principle is simple and is easy to calculate. However, for non-linear distribution data, it is difficult to find an internal structure. In recent years, the manifold learning method has attracted extensive attention from various researchers because of its advantages in nonlinear dimensionality reduction.

General manifold learning methods include linear methods, such as Principal Component Analysis (PCA) (Jolliffe, 1986) and Multidimensional Scaling (MDS)

Corresponding Author: Song Xian-feng, College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China

(Borg and Groenen, 1997) and other non-linear manifold learning methods, such as Isomap (Tenenbaum *et al*., 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Laplacian Eigenmaps (LE) (Belkin and Niyogi, 2003) and Local Tangent Space Alignment (LTSA) (Zhang and Zha, 2005). The manifold learning method can determine the intrinsic data structure using correlation matrix eigenvalue decomposition. Most manifold learning algorithms suffer from a convex optimization problem and can find the optimal solution within polynomial time. In recent years, many studies of manifold learning have been published. It is generally believed that the Isomap and LLE methods put forward a new research direction for dimensionality reduction in 2000. Since then, the LE and LTSA manifold learning algorithms have been proposed. Improved methods that are based on these algorithms have been developed and studied rapidly.

PRINCIPLE OF LLE METHOD

Roweis and Saul (2000) developed a manifold learning algorithm called Locally Linear Embedding (LLE) [4]. This is an unsupervised, non-linear technique that analyses high-dimensional data sets and reduces their dimensionalities while preserving local topology (i.e., the data points that are close in the high-dimensional space remain close in the low-dimensional space). This algorithm is able to find nonlinear structures in high-dimensional data and also has the features of translational and rotational invariance.

As a nonlinear dimensionality reduction method, LLE assumes that the topology of the local neighborhood points is preserved in the high-dimensional observation space and the low-dimensional embedding space. LLE requires the weight of each neighborhood in low-dimensional space to be unchanged and the reconstruction error becomes minimal if the data is linearized locally. Each sample point in the observation space can be embedded by weighted average reconstruction from its neighborhood points. The weighted value of all points in the high-dimensional space forms a weight matrix, based on which, it is possible to compute the samples' embedding coordinates in the low-dimensional space.

The LLE algorithm consists of the following three main steps.

- **Step 1:** Finding the neighborhood of each data point.
- **Step 2:** Assigning weights to pairs of neighboring points
- **Step 3:** Computing the low-dimensional embedding forms based on the weights and minimizing the errors

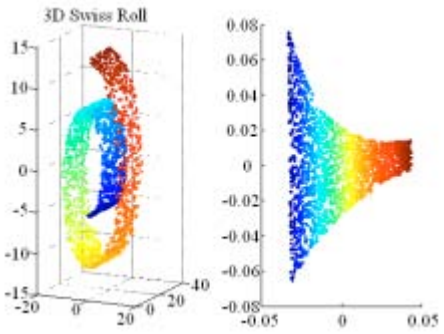

Fig. 1: Uniform sampling and its dimension reduction result

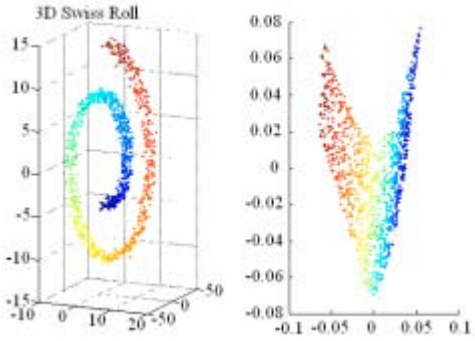

Fig. 2: Random sampling and its dimension reduction result

The advantages of LLE are as follows: (1) Only two parameters need to be set, (2) a single global coordinate system exists for the embedded space, (3) the embedded space preserves the local geometry of the high-dimensional data relatively well and (4) a non-iterative method exists for scaling to large, high-dimensional data sets (due to a sparse eigenvector problem) and avoiding the problem with a local minimum. Only the sample points of the neighborhood are involved in the reconstruction for the LLE algorithm. As a result, the dimensionality reduction results depend on the selection of the neighborhood. This usually requires a larger sample size of the sampling data. Moreover, the sampling data should be uniform and free of noise.

In Fig. 1 and 2, the images on the left represent the Swiss-roll ideal sampling and sparse non-uniform sampling of the three-dimensional renderings and the images on the right represent the two-dimensional results that are generated by the LLE algorithm.

Figures 1 and 2 show that, for the same neighborhood value k, if the sampled data is ideal, the neighborhood structure of the original data can be

preserved when using the LLE algorithm to map the neighborhood of the original data to the target space neighborhood. If the sampled data is not ideal, the embedded data structure may be distorted significantly. Therefore, we can see that the LLE algorithm has a higher level of sensitivity to non-ideal sampling Swiss-roll data.

IMPROVEMENT OF LLE

Improved distance calculation method for LLE: In the LLE algorithm, it is necessary for the original sample set to be continuous and uniformly distributed in the manifold. The selection of the neighbor number k is very important. If k is too small, it is difficult to ensure the geometry feature of the data. If k is too large, distant data points may be taken as neighbors and this may cause distortions during the dimensionality reduction. For sparse and uneven economic statistics data sets, the number that is selected for the neighborhood selection is particularly important.

For the sample data, in the region where the sample points are distributed sparsely, the neighborhood that consists of the k neighboring points is clearly larger than the sample point distribution-intensive areas. In order to reduce LLE's dependence on the distribution of sample points, this paper calculates the neighbor point for each sample point in step 1 using the following distance formula:

$$d_{ij} = \frac{|x_i - x_j|}{\sqrt{M(i)M(j)}}$$

where, $|x_i-x_j|$ represents the Euclidean distance between x_i and x_j and M(i) represents the average distance between point x_i, i = 1, 2, 3, …, n and its neighbors. The distance d_{ij} increases between sample points in intensive distributed areas and decreases in the sparse distributed areas. This process can reduce the impact caused by uneven distributions of sample points on the dimensionality reduction results.

Adaptive LLE (ALLE) method: The neighbor value k obtained by the LLE algorithm is generally greater than the intrinsic dimension d of high-dimensional data. In order to restore the global data structure, there must be a certain amount of overlap between the neighborhoods. If k is large, the reconstruction value will be non-unique and the dimension reduction process will pull non-adjacent points into the neighborhood and as a result, will not guarantee a local linear structure. The LLE algorithm is sensitive to the selection of k in cases where the manifold structure of the data (e.g., economic statistics data) is unknown.

Different parts of the manifold may have different properties. The data may be more or less sparse in different areas of the manifold and the extent of bending may vary throughout the manifold as well. In cases like this, selecting the same neighbor values for the entire manifold will cause distortions in the nature of the manifold structure. In order to fit the manifold structure, the k value for neighbors should be variable. Assume that the k neighbor points of x_i are x_{ij}, j = 1, 2, 3, …, k and that the average distance between x_i and neighbor points is:

$$d_i = \frac{1}{k}\sum_{j=1}^{k} x_i - x_{ij}$$

i = 1, 2, … n. As a result, the average distance from a given point to its neighbor points is:

$$d = \frac{1}{n}\sum_{i=1}^{n} d_i$$

In order to fit the structure of manifold, the neighbor points of the sample point x_i should be $k_i = k*d/d_i$. This value may not be an integer, but it may be set up to round to an integer.

Since $k_i = k*d/d_i$, the average distance within the neighborhood is large when the neighborhood is sparsely populated. Therefore, the algorithm can let k be small so as to avoid judging the non-neighbor points as neighbor points. When the neighborhood is densely populated, the algorithm increases the value of k automatically in order to avoid weak manifold associations that are due to lack of data and are able to distort the overall structure of the manifold.

After this, the steps of the ALLE algorithm can be concluded as follows. Step 1 involves calculating the Euclidean distance between each sample point x_i, i = 1, 2, 3, …, n and its neighboring points, calculating the average distance d_i for each sample point and its neighbor points, calculating the average distance for the whole set of sample points d, setting the number of neighboring points to be k and selecting the new neighboring points based on the value of k. Step 2 and 3 remain the same as they are in the traditional method described above.

DESCRIPTION OF STUDY AREA AND DATA

Based on statistics data from the 2007 Sichuan Statistical Yearbook from China, this paper selects 18 available properties that can be used to represent the economic development, including the value of domestic production (first industry, secondary industry, tertiary industry, industrial production and GDP), private

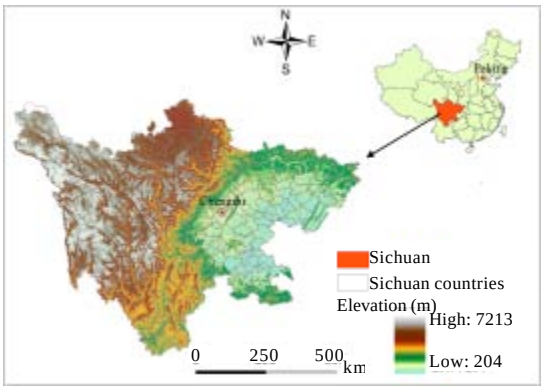

Fig. 3: The location and terrain of sichuan province

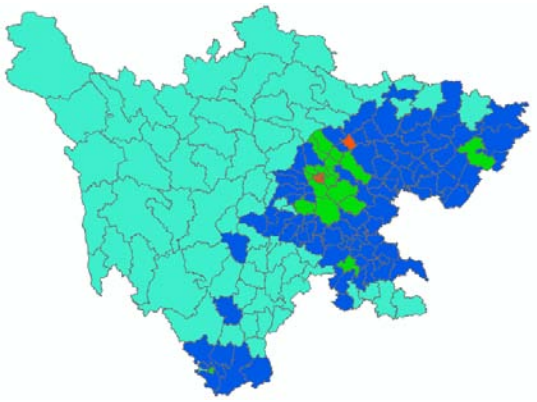

Fig. 4: LLE classification result when k = 8

economic production (first industry, secondary industry, tertiary industry, industrial production and the private economic value added per capita), employees (employees, number of employees, average wages), local governments (revenue and expenses), animal husbandry and fishery output, retail sales, fixed asset investment and the analysis of these categories.

The Sichuan Province is located in southwest China on the upper reaches of the Yangtze River. The east longitude and north latitude are 97°21'-108°31' and 26°03'-34°19', respectively. Chengdu city is the political, economic and financial center of the Sichuan Province and is developing very well because of economic development policies, business investments and other stimuli. The terrain of the eastern Sichuan Province is mainly level and flat. It has a highly developed transportation system, the most densely populated towns in Sichuan and the highest level of economic development in Southwest China. The terrain in the southern part of Sichuan is mountainous, complex and diverse, but the resources are very rich (Fig. 3). In recent years, the government has put a lot of effort into developing the tourism industry in this area. As a result, the economy there is more developed than it used to be. The terrain of the western and northern parts of Sichuan is a mountainous plateau. Infrastructure is undeveloped and the industrial foundation is weak. As a result, the socio-economic development gap between the eastern and the western parts of Sichuan is very noticeable.

RESULTS AND DISCUSSION

Comparison of classification results obtained by different algorithms: In order to reveal the merits of the different methods of analysis of the multi-dimensional data, this paper compares the results from the LLE method, the LLE improved distance calculation method and the ALLE method using the statistical data from the 2007 Sichuan Statistical Yearbook. The results are analyzed in order to illustrate the differences between the various methods based on the status quo of economic development and economic planning in Sichuan.

Based on the algorithms of LLE, distance improved LLE and ALLE, we reduce the total 18 economic factors to three primary ones. In order to evaluate the dimension reduction effect of each algorithm, we determine the parameter k using K-Nearest Neighbor (KNN) method and use the Self-Organizing Feature Map (SOFM) algorithm to classify the dimension reduction results with k = 8. The following figures show the clusters.

The classification result in Fig. 4 reflects the overall geographical distribution of Sichuan's economic development status. It assigns Chengdu and some of the surrounding counties to the first level. Chengdu, as the provincial capital, has high-quality infrastructure, rich tourism, cultural resources and supporting policies. These advantages contribute to its high level of economic development. The eastern and northeastern counties are assigned to the second level. The economic development in these counties is high because of convenient transportation and the positive effects from nearby Chengdu. The western and southwestern areas of Sichuan, on the other hand, are relatively rugged and inaccessible. The poor infrastructure and below average GDP in this region make it reasonable to assign this geographical area to the fourth level. However, there are only a few counties assigned to the third level by the LLE method, which is contrary to the actual situation.

The ALLE improved distance calculation method reveals the economic development situation in Sichuan more accurately than the LLE algorithm. Figure 5 shows that the first level includes the counties surrounding Chengdu and a small number of northeastern counties. For this method, there are more counties in the first level.

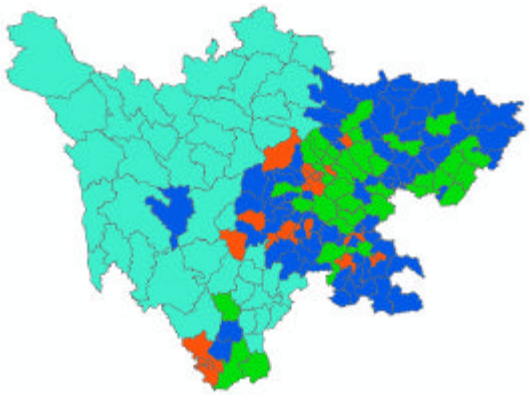

Fig. 5: ALLE classification result when k = 8

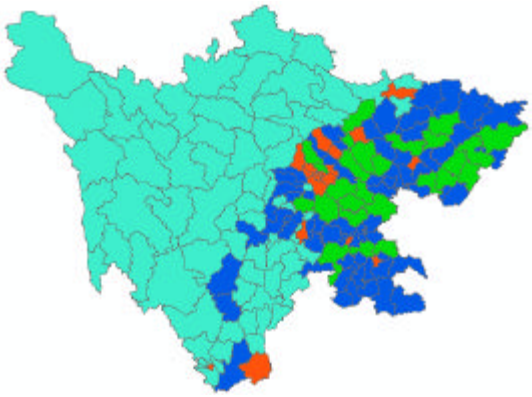

Fig. 7: ALLE k = 10

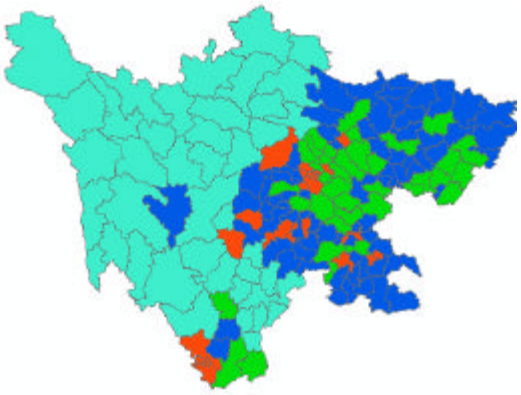

Fig. 6: ALLE k = 8

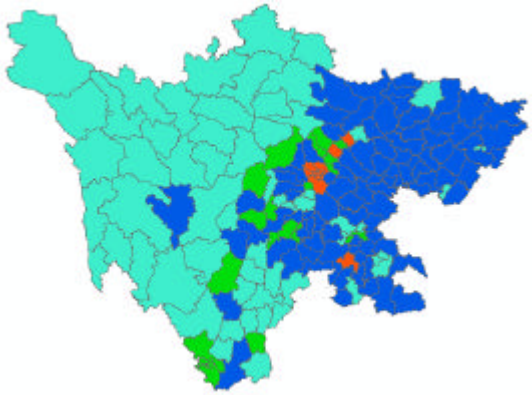

Fig. 8: ALLE k = 12

This reflects the positive impact that the capital city Chengdu has on the economies of the surrounding areas. This method also shows there are more counties in the second level. This matches the actual economic situation in Sichuan more closely. However, the distribution of these second level counties is dispersed and this is somewhat contrary to the actual situation.

The ALLE method assigns more counties to the first level than the LLE improved distance calculation algorithm. Figure 5 shows that more northeastern counties are added. Fewer counties are assigned to the second level and their distributions are more dispersed.

Comparison of classification results obtained by ALLE for different k values: We chose 8, 10 and 12 as the neighbor value k in order to determine whether the ALLE algorithm is less dependent on the neighbor value than the other LLE algorithms. We reduced the data to three dimensions and classified the data using the SOFM method. Figure 6-8 show the results. From the comparison of the results, we can see that each of the classifications is able to reveal the basic features of the economic development in Sichuan, but the different k values lead to slight diversity in the results. After multiple experiments, we concluded that the results were more accurate near a certain neighbor value and that the self-adaptive LLE algorithm does not completely eliminate the dependency on the k number of neighbors. Therefore, we suggest that the range of the neighbor value should be determined first, that the elementary results should be analyzed next and that the neighbor value should be adapted last of all in order to achieve more accurate results.

CONCLUSIONS

Because of the nonlinear distribution of the economic statistical data and the features of dispersion and uneven distribution, this paper uses a local linear embedded method to perform the nonlinear dimension reduction process and improves the methodology for calculating the

distance of neighbor points and choosing the number of neighbors. The visualized results show that the improved algorithm can retain much more information during the dimensional-reduction process and as a result, can obtain more accurate results. However, the local linear embedded method cannot eliminate the dependency on the neighbor value completely and this can affect the results from the dimension reduction process more or less depending on the properties of the data.

REFERENCES

Belkin, M. and P. Niyogi, 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput., 15: 1373-1396.

Borg, I. and P. Groenen, 1997. Modern Multidimensional Scaling: Theory and Applications. Springler-Verlag, New York, USA.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, Berlin, Germany.

Parmanto, B., M. Paramita, W. Sugiantara, G. Pramana, M. Scotch and D. Burke, 2008. Spatial and multidimensional visualization of Indonesia's village health statistics. Int. J. Health Geographics, Vol. 7 10.1186/1476-072X-7-30

Roweis, S.T. and L.K. Saul, 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 290: 2323-2326.

Tenenbaum, J.B., V. de Silva and J.C. Langford, 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319-2323.

Zhang, Z. and H. Zha, 2005. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. Sci. Comput., 26: 313-338.