Sparse Manifold Learning based on Laplacian Matrix

Wenjing Li
School of informatics and computer science
Indiana University Bloomington
Bloomington, USA
wenjing__li@126.com

ABSTRACT
In machine learning, a group of high-dimensional data point set which represents an image set can be looked upon as the point set distributing on a nonlinear manifold. Typically, the manifold dimension is much lower than the dimension of data points. Therefore, it is important to explore the real dimension and the real geometry of high-dimensional data point set. Based on this objective, researchers put forward the concept of manifold learning. Traditional manifold learning can achieve dramatic dimensional reduction on high-dimension points, but there are still some problems of it. In the aspect of processing a large amount of data currently, traditional manifold learning algorithm reveals many weaknesses mainly in time consumption. To improve the deficiency, the paper puts forward a new algorithm, aiming at simplifying the time process of manifold learning algorithm which is also abbreviated as SLEP. In the part of the experiment, the study makes a comparative experiment between the synthetic data set and the real data set. The result shows that the proposed algorithm improves time efficiency.

1. INTRODUCTION
Machine learning is a subset of artificial intelligence based on statistics to train the computers making decisions wisely. The primary methods in implementing data-driven include prediction and predictive analytics. And the essential function of machine learning is finding the pattern by extracting the intrinsic pattern of data from large amount of unstructured data. The major methods, including PCA [6], and Manifold learning [3, 5], and others are divided to supervised methods and unsupervised methods.

In supervised method, the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.(wiki) Comparably, unsupervised method are not given with the specific features. As one of the representing method of unsupervised method, Laplacian [1] is local-preserving and has simple core algorithm. And it solved the problem of representing low-dimensional data when data arise from sampling a probability distribution on a manifold.

The one of the deficiency of the current algorithms of manifold is selecting limited neighbors within unqualified high-density data set. In this letter, we explore an approach that solves the problem of constructing adjacency graph by constructing connected set from the constructed adjacency graph. The implementation of SLEP is simple and local-preserving. And the solution reflects the geometrical-intrinsic of the manifold.

2. RELATED WORK
Manifold learning is an important branch of machine learning, which has a great development since the 21st century. In 2000, there were two papers released in Science, Isomap [3] and LLE [5], firstly put
forward the concept of manifold learning. The objectives of manifold learning are to reduce the redundant information about high-dimensional data set, and realize dimensional reduction of the data set. Presently, manifold learning algorithm is divided into two types, one of which is global dimensional reduction algorithm like Isomap [3]. The other one is local dimensional reduction algorithm such as LLE [5], LEP [1], LPP [4], LTSA [7], HLLE [2] and so on. The global dimensional reduction algorithm aims at maintaining the global structure during the process of manifold dimension reduction. For example, Isomap [3] algorithm is to maintain the geodesic distance between two arbitrary points of manifold. However, local dimensional reduction algorithm is to maintain local geometric structure of manifold. For example, LLE [5] algorithm is to maintain the linear relationship between local neighborhood point sets of a point while LEP [1] is to maintain the distant structure between local neighborhood points which can keep close relationship between two closer points after dimensional reduction.

The important step of traditional manifold learning algorithm is to divide a data point set into different neighbor sets and the number of neighbor sets should be equal to the number of data points. Then according to different objectives, various geometric structures between data sets can be explored. However, with the social development, there are a lot of large data sets with high dimension and massive data volumes to be solved at present. If researchers still adopt the traditional manifold learning algorithm to reduce dimension of this type of massive data set, it will lead to a complex learning process. Aiming at this circumstance, a new algorithm should be designed to reduce time complexity of the algorithm to a great extent without affecting the result of dimensional reduction. By using the proposed algorithm, the efficiency of processing massive data sets will be improved.

3. SOLUTION OF THE ALGORITHM

Firstly, it needs to explain the symbols used in the algorithm. If the high-dimensional input data point set that we process shows as \( \{x_1, x_2, \ldots, x_N\}, x_i \in \mathbb{R}^D \), where \( N \) refers to the number of data point, and \( D \) refers to dimension of data point, the corresponding lower dimension after dimensional reduction shows as \( \{y_1, y_2, \ldots, y_N\}, y_i \in \mathbb{R}^d \), where \( d \) refers to dimension of lower dimensional space. In the following, we first give the method of selecting local minimal neighborhood set, then we describe our proposed sparse manifold learning algorithm (SLEP for short) and in the last we analyze the comparison between traditional LEP algorithm and our SLEP.

3.1 Selection of Local Minimal Neighborhood Set

Firstly, K-nearest neighborhood can be used to solve the local K-nearest neighbor points of \( x_i \) at each point, which aims at finding a group of neighborhood subset from the neighborhood set to minimize the number of neighborhood subset, and completely cover all sample points. Moreover, adjacent neighborhood sets are in the relationship of mutual coverage. Therefore, the key of the problem is to find the neighborhood subcovering with finding an algorithm to search the minimal covering subset. The key of the algorithm is to define a set of rules to depict the structural relationship between any two neighborhoods. The existing method is using Greedy Algorithm to search, but it’s best to find an algorithm which is more suited to our needs, or improve the existing algorithm based on our objectives. First of all, we can directly construct a k-neighborhood set at any point. However, obviously, there are many redundant sets in this group of neighborhood set. Our objective is to select a subset from this group of neighborhood set. The main idea of this method is referred to [8]. The specific algorithm is shown as following:

Firstly, it needs to construct a K-neighborhood set \( \{U_i\}, i = 1, \ldots, N \) at any point \( x_i \), and randomly select a K-neighborhood set \( U_i \) as the initial set. Then a neighborhood subset should be selected from the rest neighborhood sets to make this group of the neighborhood subset fully cover all sample point sets and each K-neighborhood is in the relationship of mutual coverage. To make the neighborhoods cover each other and the coverage rate not be too high, we should make an index to measure the coverage rate between two neighborhood sets:

\[ |S \cap U| \leq (1 - \alpha)|S| \]
Where $\alpha$ refers to the coverage rate between two sets.

3.2 Algorithm Process
The thesis aims at improving the LEP algorithm. The improved algorithm is called SLEP, and the main process is shown as following:

1. Use K-nearest neighborhood algorithm to solve the K-neighborhood set $U_i$ at $x_i$ of each point.
2. Use the minimal sub-neighborhood set selection algorithm given in the section 3.1 to select a group of sub-neighborhood set shown as $\{M_j\}_{j=1}^m$ from the set $\{U_i\}_{i=1}^n$.
3. Use the reconstructed neighborhood set $\{M_j\}_{j=1}^m$ between data points to construct the weight between data points. The constructing algorithm is shown as following:

$$W_{ik} = \begin{cases} e^{-\frac{\|x_i - x_k\|^2}{2\sigma^2}}, & \text{if } x_i, x_k \in M_j \\ 0, & \text{if } x_i, x_k \notin M_j \end{cases}$$

4. Use LEP to construct the corresponding Laplacian matrix, and then solve the corresponding lower dimension to show $\{y_1, y_2, \cdots, y_N\}$

3.3 The comparison between SLEP and LEP
Different from LEP, the Laplacian matrix constructed by the proposed algorithm is sparse. Because we construct a group of sparse local neighborhood set at a dataset, its number of local neighborhoods is less than the number of data points. Compared with the weight matrix $W$ constructed by LEP, the weight matrix $W_s$ among data points constructed by the minimal neighborhood sets is sparser. The proposed algorithm mainly aims at the circumstances of a huge number of data points, which can lower the time complexity of lower dimension algorithm.

4. EXPERIMENT
In this chapter, a comparative experiment between the proposed algorithm and traditional manifold learning algorithm will be made to illustrate that the proposed algorithm is better than other algorithms from the perspective of time efficiency. In the experiment, the comparative experiment is between synthetic data sets and real data sets. Specifically, synthetic data sets are Swiss Roll data set and Puncture Sphere data set which are generated from the document mani.m while the real data set is USPS data set. During the experiment, five traditional manifold learning algorithms, Isomap [3], LLE [5], LTSA [7], HLLE [2] and PCA [6] are compared with the proposed algorithm, SLEP.

4.1 Synthetic data set
The experiment respectively produces different quantities of Swiss Roll data sets and Puncture Sphere data sets from the document mani.m. These two groups of data sets are three-dimensional data sets and the two-dimensional manifold structure embedding into the three-dimensional space. The objective of manifold learning is to reduce the dimension of these three-dimensional data sets in the two-dimensional space. And the objective of the proposed algorithm is to deduce the time complexity of traditional manifold learning algorithm. In order to verify that the proposed algorithm has an advantage in time consumption, the experiment is based on different quantities of data sets.

During the experiment, researchers respectively select 2000 and 5000 data points to make experiment for each data set. The experiment also takes the advantage of traditional manifold learning algorithm and SLEP to reduce the dimension of these two groups of data sets and then calculate time consumption of each algorithm. The corresponding time consumption is shown in table 1. It can be seen that compared with traditional manifold learning algorithm, time consumption of the proposed algorithm is lower. During the process of processing 5000 data points, the advantage of SLEP in time consumption is more obvious. In results of the experiment, time comparison between SLEP and LEP is much concerned. Table 1 shows that the proposed algorithm SLEP has a great advantage.
In addition to the advantage in time consumption, researchers hope that SLEP can also have an advantage in dimensional reduction, so the experiment is also made in dimensional reduction to verify the hypothesis. The result is shown in graph 1 and 2 which reflect that there is not much difference between the effect of dimensional reduction of the proposed algorithm and LEP.

Table 1. The comparison of time consumption of manifold learning algorithms

|                | Swiss Roll/5000 | Sphere/5000 | Swiss Roll/10000 | Sphere/10000 |
|----------------|-----------------|-------------|------------------|--------------|
| PCA [6]        | 0.6858          | 0.5886      | 2.4815           | 2.3984       |
| Isomap [3]     | 1071.2383       | 969.1432    | 0.6467           | 0.9261       |
| LLE [5]        | 1.8241          | 1.6205      | 6.5013           | 4.8359       |
| LEP [1]        | 2.1818          | 1.8506      | 8.6720           | 7.6406       |
| HLLE [2]       | 38.3258         | 34.5965     | 257.5510         | 254.9848     |
| LTSA [7]       | 3.2619          | 3.6690      | 18.2465          | 16.9970      |
| SLEP           | 0.8965          | 0.7536      | 3.1584           | 2.8156       |

Fig.1. The dimensional reduction of data sets in different manifold learning algorithms, from left to right and from top to bottom: Swiss Roll data set, PCA, Isomap, LLE, HLLE, LTSA, LEP, SLEP
4.2 Real data set

USPS data set is a group of electronic digital data set, showing the image data set of the script written from 0 to 9, including 9298 images in total. For the pixel of each digital image is 16*16, the dimension of the processed data point is 256 in the experiment. In the aspect of design, there are two steps, including using manifold learning algorithm to reduce dimension of USPS data set and using K-nearest neighbor algorithm to classify and identify data sets in the low-dimensional space. The objective is to compare the image recognition accuracy rate of traditional manifold learning algorithm and SLEP in the real data set and the time consumption during the process of recognition. The experiment selects 400, 500, and 600 data points from different types of data sets as sample sets and other data points are testing data sets. The recognition accuracy rate of different algorithms to USPS data set is shown in table 2. It shows that the recognition accuracy rate of the proposed algorithm is almost the same as traditional manifold learning algorithm. The result indicates that the proposed algorithm does not reduce the accuracy rate of dimensional reduction of data sets.

The final objective of the experiment is to analyze the time consumption of the proposed algorithm, so the experiment compares the time consumption between traditional manifold learning algorithm and the proposed algorithm in different K values. The result is shown in table 3. By the comparison of the experimental result, it can be seen that the time consumption of the proposed algorithm is less than traditional manifold learning algorithm in different K values.

Table 2. Recognition accuracy rate of USPS data set by different algorithms

|                | USPS-train400 | USPS-train500 | USPS-train600 |
|----------------|---------------|---------------|---------------|
| PCA [6]        | 83.95 ± 1.13  | 84.01 ± 1.26  | 85.27 ± 1.47  |
| LLE [5]        | 88.14 ± 1.32  | 89.54 ± 1.37  | 90.46 ± 1.25  |
| LEP [1]        | 90.58 ± 1.25  | 91.14 ± 1.05  | 91.69 ± 1.14  |
| SLEP           | 89.76 ± 1.08  | 90.81 ± 1.16  | 91.15 ± 1.02  |

Table 3. Comparison of time consumption of different dimensional reduction algorithms

|                | USPS-train400 | USPS-train500 | USPS-train600 |
|----------------|---------------|---------------|---------------|
| PCA [6]        | 313.15 s      | 298.53 s      | 265.36 s      |
| LLE [5]        | 388.14 s      | 425.36 s      | 509.14 s      |
| LEP [1]        | 526.31 s      | 579.12 s      | 617.43 s      |
| SLEP           | 324.36 s      | 368.52 s      | 413.17 s      |

5. CONCLUSION

Traditional manifold learning algorithm exists a big problem in time consumption during the process of processing massive data sets. To improve this circumstance, SLEP algorithm is put forward to reduce the time consumption of an algorithm as far as possible without affecting the accuracy of dimensional reduction. A group of the comparative experiment shows that the proposed algorithm has a great improvement in time efficiency.

References

[1] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” in NIPS, vol. 14, 2001, pp. 585-591.
[2] D. L. Donoho and C. E. Grimes, “Hessian Eigenmaps: Locally linear embedding techniques for high-dimensional data,” in Proceedings of the National Academy of Sciences of the United States of America, vol. 100, 2003, pp. 5591-5596.

[3] J. B. Tenenbaum, V. d. Silva and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” Science, vol. 290, 2000, pp. 2319-2323.

[4] X. He and P. Niyogi, “Locality preserving projections,” in NIPS, vol. 16, 2003, pp. 153-160.

[5] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality by locally linear embedding,” Science, vol. 290, 2000, pp. 2323-2326.

[6] H. Abdi and L. J. Williams, “Principal component analysis,” inc. WIREs Comp Stat, vol. 2, 2010, pp. 433-459.

[7] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, SIAM Journal of Scientific Computing 26 (2004) 313–338.

[8] Li Yang, “Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction,” IEEE Tran. On Pattern Analysis and Machine Intelligence, vol. 30, 2008, pp. 438-450.