

Cluster Preserving Embedding

Yubin Zhan Jianping Yin

Computer School, National University of Defense Technology, Changsha, 410073, P.R. China
YubinZhan@nudt.edu.cn

Abstract

Most of existing dimensionality reduction methods obtain the low-dimensional embedding via preserving a certain property of the data, such as locality, neighborhood relationship. However, the intrinsic cluster structure of data, which plays a key role in analyzing and utilizing the data, has been ignored by the state-of-the-art dimensionality reduction methods. Hence, in this paper we propose a novel dimensionality reduction method called Cluster Preserving Embedding(CPE), in which the cluster structure of original data is preserved via preserving the robust path-based similarity between pairwise points. We present two different methods to preserve this similarity. One is the Multidimensional Scaling(MDS) way, which tries to preserve similarity matrix accurately, the other one is a Laplacian-style way, which preserves the topological partial order of the similarity rather than similarity itself. Encouraging experimental results on a toy data set and handwritten digits from MNIST database demonstrate the effectiveness of our Cluster Preserving Embedding method.

1. Introduction

Many applications in pattern recognition involve in dealing with the high dimensional data, which will result in the so-called “curse of dimensionality”. Dimensionality Reduction(DR) is an effective and widely used approach to deal with such high dimensional data. Up to now, researchers have developed a variety of dimensionality reduction methods(DRs) under supervised, unsupervised and semi-supervised scenarios. The supervised DRs mainly include Linear Discriminant Analysis(LDA)[4] and Maximum Margin Criterion(MMC)[6], while the unsupervised DRs include Principal Component Analysis(PCA)[5], Multidimensional Scaling(MDS)[3], Isomap[8], Local Linear Embedding(LLE)[7] and Laplacian Eigenmap[1]. In this paper we only focus on unsupervised sce-

nario, however, the proposed method can be easily and straightforwardly extended to include the supervised information.

In methodology, the existing unsupervised approaches can be divided into two categories: global and local. Global DRs mainly include PCA, MDS and its variant Isomap, the embedding is obtained by preserving a certain global property of the original data. For example, the property that PCA preserves is the global variance, metric MDS tries to preserve distance(dissimilarity) between pairwise points and its variant Isomap aims to preserve the geometric distance on the intrinsic manifold. The local approaches try to preserve a certain local property of the data, such as locality, neighborhood relationship. Local methods mainly include LLE, Laplacian Eigenmap and Local Tangent Space Alignment(LTSA)[11].

In summary, although the aforementioned methods have different motivations and concrete implementations, they all try to preserve a certain property of the original data whether global or local. However, cluster structure as a key property of data, which reflects the intrinsic distribution of data and plays a crucial role in further analyzing and utilizing the data[9], has been ignored by the these approaches. Hence, in this paper we propose a cluster preserving DR method called Cluster Preserving Embedding(CPE).

In CPE, the cluster structure of the original high dimensional data is preserved via preserving the robust path-based similarity, which is widely used in clustering algorithm[2, 10]. After obtaining the robust path-based similarity matrix, two ways to compute the embedding while preserving this similarity are presented. One is to employ MDS which will try to preserve the similarity matrix exactly. The other one is Laplacian-style way, this method prefer to preserving topological partial order of the similarity matrix. Finally, we conduct experiments on a toy data set and handwritten digits from MNIST database to evaluate the performance of our CPE method. Encouraging experimental results have validated the effectiveness of our CPE method.

2. Cluster Preserving Embedding

Let us begin with briefly reviewing the robust path-based similarity proposed by Chang[2].

2.1. Robust Path-Based Similarity

Given n data points $X = (x_1, x_2, \dots, x_n)$ where each column vector $x_i \in R^D$ represents a sample, we can construct an undirected full connected graph $G = (X, W)$, where each vertex represents a sample in X and the element w_{ij} in matrix W is the weight of edge between node x_i and x_j . w_{ij} can reflect the similarity between x_i and x_j . The original similarity between points x_i and x_j can be assigned as follows:

$$w_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) & i \neq j \\ 0 & i = j \end{cases} \quad (1)$$

The scaling parameter σ controls how fast w_{ij} decreases with the distance between x_i and x_j .

Since the original pairwise similarity w_{ij} defined above is only determined by the Euclidean distance in the original high dimensional space, it can not reflect the real geometric structure of the data, consequently it can not reveal whether the two points belong to the same cluster. To capture this information, Ref.[2] defines the path-based similarity by exploiting the underlying manifold structure of the whole data as described below.

Let \mathcal{P}_{ij} be the collection of all the paths between points x_i and x_j . For each path $p \in \mathcal{P}_{ij}$, the *effective similarity* s_{ij}^p is the minimum edge weight along path p , then the path-based similarity s_{ij} between x_i and x_j is defined as the maximum effective similarity among all the path in \mathcal{P}_{ij} , it can be formally expressed as:

$$s_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq k < \#p} w_{p[k]p[k+1]} \right\} \quad (2)$$

where $\#p$ is the number of points that p goes through and $p[k]$ is the global index of the k th vertex in path p .

As pointed out in Ref.[2], the above defined path-based similarity is sensitive to noise and outliers, hence they introduce a weight α_i for each point x_i by robust M-estimation as:

$$\alpha_i = \sum_{x_j \in \mathcal{N}(x_i)} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (3)$$

where $\mathcal{N}(x_i)$ denotes the neighborhood of x_i . To further make α_i insensitive to parameter σ , normalized weights are computed as $\alpha'_i = \alpha_i / \max_{1 \leq i \leq n} \alpha_i$. Then finally, the robust path-based similarity is computed as:

$$s_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq k < \#p} \alpha'_{p[k]} \alpha'_{p[k+1]} w_{p[k]p[k+1]} \right\} \quad (4)$$

2.2. Cluster Preserving Embedding

Now let us take a close look at the robust path-based similarity matrix $S = (s_{ij})$. A large s_{ij} implies that there exists a path between samples x_i and x_j that goes through the high density region, then according to cluster assumption in machine learning, points x_i and x_j should be in the same cluster. On the contrary, small s_{ij} means that any path between x_i and x_j will pass through the low density region, hence they should be in different clusters. Therefore, the robust path-based similarity s_{ij} measures the likelihood that points x_i and x_j belong to the same cluster. To preserve the cluster structure of the original data, one should preserve the robust path-based similarity matrix S .

There are two ways for us to preserve the similarity matrix S . One is the classic MDS method, the other one is Laplacian-style way.

Given a distance(dissimilarity) matrix between n samples, classical metric multidimensional scaling(MDS) can give a configuration of this n samples in R^d while preserving the distance matrix as accurately as possible. Recall that in Isomap, the geometric distance matrix is preserved by MDS, hence we can also employ MDS to preserve our similarity matrix S . First we should covert our similarity matrix S to dissimilarity matrix $\hat{S} = (\hat{s}_{ij})$. For the robust path-based similarity matrix S obtained by Eq.(4), we set $s_{ii} = \max_{j \neq k} s_{jk}$ ($i = 1, 2, \dots, n$). Then we adopt the following standard conversion to obtain the dissimilarity matrix $\hat{S} = (\hat{s}_{ij})$:

$$\hat{s}_{ij} = s_{ii} - 2s_{ij} + s_{jj} \quad (5)$$

Then it is easy for us to call classic metric MDS to obtain the low dimensional embedding. We refer to our cluster preserving embedding with MDS solution as CPE-MDS.

The other way for us to preserve similarity between pairwise points is the Laplacian-style way. In Laplacian Eigenmap[1], the locality of data set is preserved by the following formula:

$$\min_{1 \leq i, j \leq n} w_{ij} \|y_i - y_j\|^2 \quad (6)$$

where y_i is the low dimensional coordinates of x_i and w_{ij} is the edge weight in the k -neighborhood graph on data set X . One can use the Gaussian function to define W as follows:

$$w_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) & x_i \in \mathcal{N}(x_j) \text{ or } x_j \in \mathcal{N}(x_i) \\ 0 & \text{otherwise} \end{cases}$$

In fact, the weight matrix W measures the locality relationship of data set X and formula (6) will make

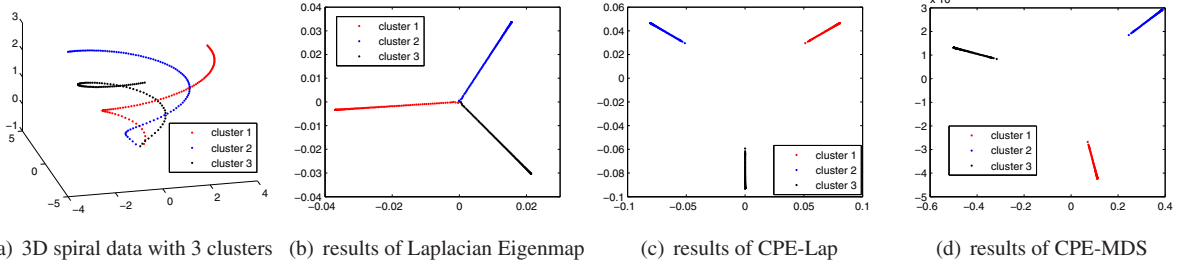


Figure 1. 3D spiral data and results of Laplacian Eigenmap, CPE-Lap and CPE-MDS on it.

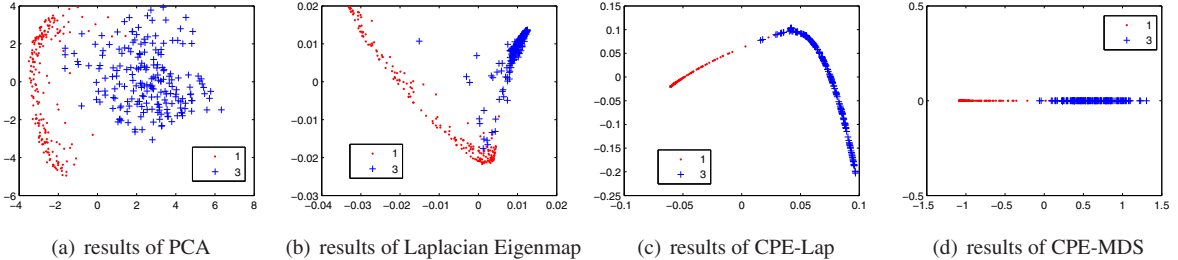


Figure 2. Results of PCA, Laplacian Eigenmap, CPE-Lap and CPE-MDS on digit subset containing “1” and “3”.

data points x_i and x_j close to each other if w_{ij} is large but far from each other if w_{ij} is small.

In our CPE method, large s_{ij} implies points x_i and x_j belong to the same cluster in original high dimensional space, they should be close to each other in the low dimensional space. While small s_{ij} implies x_i and x_j belong to the different clusters, they should be far from each other. Then similar to Laplacian Eigenmap, we can obtain the low dimensional embedding $Y = (y_1, y_2, \dots, y_n)$ while preserving the similarity matrix S via solving the following optimization problem:

$$\min_{1 \leq i, j \leq n} s_{ij} \|y_i - y_j\|^2 \quad (7)$$

Through trivial computations, the above optimization problem can be expressed as:

$$\min \text{tr}(Y^T L Y) \quad (8)$$

where $L = \hat{D} - S$, \hat{D} is diagonal matrix with $d_{ii} = \sum_{j=1}^n s_{ij}$. To remove the freedom of Y , we additionally require that $Y^T \hat{D} Y = I$, where I is the identity matrix. Then the optimal Y are composed of the eigenvectors corresponding to the first d -th small nonzero eigenvalues of the following generalized eigenvalue problem:

$$L y = \lambda \hat{D} y \quad (9)$$

Let v_i denote the eigenvector corresponding to the i -th small eigenvalue λ_i of the eigenvalue problem (9)

($0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$), then the optimal $Y = (v_2, v_3, \dots, v_{d+1})$.

We refer to our CPE with solution obtained by this method as CPE-Lap. Note that the local relationship matrix W in Laplacian Eigenmap can be seen as a special similarity matrix, therefore Laplacian Eigenmap can be regarded as a special case of our CPE-Lap.

3. Experimental results

We first conduct experiments on a toy data set, which contains 300 data points sampled from three 3D spiral curves (100 points for each curve). So it naturally forms 3 clusters in the 3-dimensional space. The original data points are plotted in fig.1(a). We then perform Laplacian Eigenmap, CPE-Lap and CPE-MDS to obtain the 2-dimensional embeddings of this data set, which are shown in fig.1(b)(c)(d). From the results, one can see that although Laplacian Eigenmap can preserve the manifold structure well, the cluster structure are not preserved so well as our CPE-Lap and CPE-MDS do. Our CPE can obtain 3 clusters mutually far apart from each other.

To further evaluate the performance of our CPE method, we perform experiments on the real world data set i.e., handwritten digits from the well-known MNIST database¹. Unlike the synthetic data, this data set is of much higher dimensionality. Each image has been size

¹<http://yann.lecun.com/exdb/mnist/>

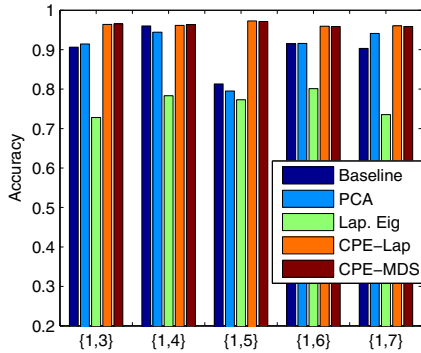


Figure 3. Clustering accuracy of different methods on different digits subsets.

normalized and centered to 28×28 gray-level images, so the dimensionality of the digit space is 784. In our experiments, we randomly select 200 images for each digit to obtain their low dimensional embedding.

Fig.2 shows the 2 dimensional embeddings of digit subset containing digits “1” and “3” obtained by PCA, Laplacian Eigenmap, CPE-Lap and CPE-MDS². It can be seen that our method can preserve the cluster structure better than others.

We conduct more experiments on different digit subsets under the same experimental settings. First the low dimensional embeddings are obtained by different methods, then the k -means clustering is performed on the low dimensional embedding, finally Rand Index are exploited to quantify the clustering accuracy. Since the images are randomly selected, we repeat 20 times to obtain the average clustering accuracy. Fig.3 shows the accuracy of different methods on different subsets. The accuracy of baseline is obtained by k -means clustering directly performed on the original 784-dimensional space without any preprocessing. From the results, one can see that k -means clustering can yield higher accuracy on the embedding obtained by our methods than those on others, hence we can draw a conclusion that our CPE can preserve the clustering structure well, the low embedding obtained by our CPE is more suitable for clustering.

4. Conclusion

Existing unsupervised dimensionality reduction methods obtain the low representations of high dimensional data points via preserving a certain property of

²For CPE-MDS, in the experiments, there is only one nonzero eigenvector corresponding to the positive eigenvalue.

the original data. However, clustering structure, which is a key property of the data and plays a key role in utilizing the data, has been ignored by state-of-the-art DR methods. This paper proposes a novel dimensionality reduction method called Clustering Preserving Embedding(CPE), which obtains the embedding via preserving the clustering structure. The encouraging experimental results on a toy data set and handwritten digits from MNIST database demonstrate the effectiveness of the proposed CPE method.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful and constructive comments. This work is supported by National Natural Science Foundation of China (Grant No. 60970034, 60603015), and the Foundation for Author of National Excellent Doctoral Dissertation(Grant No. 2007B4).

References

- [1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
- [2] H. Chang and D.-Y. Yeung. Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191–203, 1 2008.
- [3] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.
- [4] R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley & Sons, 2001.
- [5] I. Jolliffe. *Principal component analysis*. Springer verlag, 2002.
- [6] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 17(1):157–165, 2006.
- [7] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
- [8] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [9] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17:395–416, 2007.
- [10] Y. Zhang and D.-Y. Yeung. Semi-supervised discriminant analysis using robust path-based similarity. In *CVPR 2008*, pages 1–8, 2008.
- [11] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1):313–338, 2004.