LTPI: A Spectral Clustering Method Based on Local Topology Preserving Indexing and Its Application for Document Clustering

Jieqing Xing and Chuanyi Fu

Department of Information Technology, Qiongtai Teachers College, Haikou 571100, China
xingjieqing@gmail.com, fuchuanyi123@gmail.com

Abstract

In terms of machine learning theory, the intrinsic geometrical structure of the original data space is usually embedded in the low-dimensional manifold. The extraction of optimized manifold features could improve the performance of clustering. This paper presents a new spectral clustering method called local topology preserving indexing (LTPI). In this algorithm, the data are projected into a low-dimensional feature space in which the distances between the data points in the same local patches are minimized and the distances from the data points outside these patches are maximized simultaneously. The proposed LTPI method can effectively discover the intrinsic local topologies embedded in original high-dimensional space. The comparison experiments for document clustering demonstrate its effectiveness.

Keywords: *spectral clustering, graph embedding, dimensionality reduction*

1. Introduction

Spectral clustering refers to a class of techniques which rely on the eigen-structure of a similarity matrix to partition points into disjoint clusters. It has many applications in machine learning, data analysis, computer vision and speech processing.

The researchers have developed various spectral clustering algorithms. Particularly, Latent Semantic Indexing (LSI) is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original feature space by minimizing the global reconstruction error, rather than the local error [1]. However, in recent years, some studies suggest that the discriminative information is often embedded in the local topology of data space [2].

Furthermore, a certain representation of data usually resides on a nonlinear low-dimensional manifold embedded in original high-dimensional space. So an effective clustering method must be able to find a low-dimensional representation of the sample data that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method applies a weighted function to each pairwise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the sample data [3]. However, the selection of the weighted functions is often a difficult task.

Recently, graph embedding has become a topic of significant interest for dimensionality reduction [4, 5]. It usually constructs a graph to encode the geometrical information in the data. For some applications like Web search, the graph can be pre-defined by using hyperlinks. Using the notion of graph Laplacian, one can find a lower-dimensional representation which respects the graph structure [6]. Many state-of-the-art dimensionality reduction algorithms

such as Isomap [7], Laplacian Eigenmap [8], Locally Linear Embedding [9], Neighborhood Preserving Embedding [10] and Locality Preserving Projections [2], as well as canonical algorithms like Principle Component Analysis and Linear Discriminant Analysis, can be interpreted in a general graph embedding framework with different choices of the graph structure.

This way, the data points belonging to the same class are merged together in the embedding space in which better classification or clustering performance can be obtained. A key problem in graph embedding is the out-of-sample extension. We further propose that an explicit mapping function can be learned which is defined everywhere.

In this paper, we propose a new spectral clustering method based on local topology preserving indexing (LTPI), which explicitly considers the manifold structure embedded in the similarities between the samples. It aims to find an optimal low-dimensional subspace by simultaneously maximizing the distances between the data points in the local patches and minimizing the distances between the data points outside these patches. This is different from LSI and LPI, which are focused on detecting the intrinsic structure between widely separated data points rather than on detecting the intrinsic structure between nearby data points. Since the intrinsic structure is embedded in the low-dimensional manifold, the LTPI method focuses on detecting the intrinsic structure between nearby data points rather than on detecting the intrinsic structure between widely separated points.

The rest of the paper is organized as follows. The proposed spectral clustering based on LTPI is presented in Section 2. In Section 3, experimental results are provided to illustrate the performance of the LTPI method. Finally, conclusions are given in Section 4.

2. Spectral Clustering based on LTPI

2.1 Constructing Graphs and Weights

Denote the sample set as $X = [x_1, x_2, \ldots, x_N]$, $x_i \in R^m$. Here, we define two types of graph: intrinsic graph G and penalty graph G^p. Let $G = \{X, W\}$ be an undirected weighted graph with vertex set X and the intrinsic matrix $W \in R^{N \times N}$. $G^p = \{X, W^p\}$ is the other undirected weighted graph with vertex set X and the penalty matrix $W^p \in R^{N \times N}$. W and W^p are both symmetric $N \times N$ matrices with each element having the weight of the edge joining vertices x_i and x_j.

For G, the vertex x_i and the vertex x_j are connected if x_j is among the k nearest neighbors of x_i. In contrast, for G^p, the vertex x_i and the vertex x_j are connected if x_j is not among the k nearest neighbors of x_i The k nearest neighbors are searched for with Euclidean distance. Next, weights are assigned to the connected edges. The same scheme is used for the two type of weights. That is the weight is $e^{-\frac{d^2(x_i,x_j)}{2}}$, where $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j, if x_i and x_j are connected.

The G and W characterize certain statistical or geometrical properties of the data set. The purpose of LTPI is to represent each vertex of the graph as a low dimensional vector that preserves similarities between the vertex pair, where similarity is measured by the edge weight.

2.2 Graph Preserving Criterion

In this work, we assume that the graph G characterize the similarity relationship between the data pairs. The LTPI aims to preserve local intrinsic topology of the original data in the objective low dimensional embedding space. It is respected that y_i and y_j are close in the embedding space if vertex x_i and x_j are close in the original space, and meanwhile, y_i and y_j are apart if vertex x_i and x_j are apart. Then it leads to the local topology preserving criterion as follows

$$y^* = \arg \min_{y^T By=c} \sum_{i \neq j} \|y_i - y_j\|^2 W_{ij} = \arg \min_{y^T By=c} y^T Ly \tag{1}$$

where c is a constant, B is the Laplacian matrix of a penalty graph G^p. That is $B = L^p = D^p - W^p$, where $L^p = D^p - W^p$, $D^p_{ii} = \sum_{j \neq i} W^p_{ij}, \forall i$. Accordingly, $L = D - W$, $D_{ii} = \sum_{j \neq i} W_{ij}, \forall i$.

Furthermore, we assume that the low dimensional vector representation can be obtained from linear projections as $y = A^T x$, where A is a transform matrix $A = (a_1, a_2, \ldots, a_r)$ and a_i's are column vectors. Then, the objective function (1) is changed to

$$a^* = \arg \min_{a^T XBX^T a=c} a^T XL^I X^T a \tag{2}$$

2.3 Algorithm Derivation

With Lagrange Multiplier Method, the solutions of Eq. (2) can be obtained by solving the generalized eigenvalue decomposition problem as

$$XLX^T a = XBX^T a \tag{3}$$

It is easy to show that the matrices XLX^T and XBX^T are symmetric and positive semi-definite. The vectors a_i ($i = 1, 2, \ldots, r$) that minimize the objective function are given by the minimum eigenvalue solutions to the generalized eigenvalue problem.

2.4 Clustering Algorithm with LTPI

Based on local topology preserving indexing, the proposed spectral clustering can be summarized as follows:

1) Construct the local neighbor patch, and construct the intrinsic graph G and the penalty graph G^p with their respective weight W and W^p.
2) Perform Principle Component Analysis (PCA) on original data X to avoid the well-known singular problem. Supposing the PCA transform matrix is U, we can obtain the projected data by using $\tilde{X} = U^T X$.
3) Compute LTPI projection on $\tilde{X}$. Let $A_{LTPI} = (a_1, a_2, \ldots, a_r)$, and $a_1, a_2, \ldots, a_r$ are the eigenvectors associated with the r smallest eigenvalues to the generalized eigenvalue problem $XLX^T a = XBX^T a$. Then the low-dimensional representation of the document can be computed by

$$Y = A^T_{LTPI} \tilde{X} \tag{4}$$

4) Cluster in the LTPI embedding space. We seek a partitioning $\{\pi_j\}_{j=1}^k$ of the samples using the maximization of the following objective function:

$$Q(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^{k} \sum_{x \in \pi_j} x^T c_j \tag{5}$$

with $c_j = \frac{m_j}{\|m_j\|}$, where m_j is the mean of the sample vectors constrained in the cluster π_j.

3. Experiments on Document Clustering

In this section, we apply the proposed LTPI method to document clustering. The performance of the LTPI method is evaluated and compared with other competing clustering methods. The accuracy and the normalized mutual information are used to measure the clustering performance. The two measures are defined as the reference [11].

The 20 newsgroups corpus consists of roughly 20000 documents that come from 20 specific Usenet newsgroups. We repeated the experiments in Zha et al. and Cheng, *et al.*, [12] to illustrate the performance of the proposed CPI algorithm and other competing algorithms. Experiments of c-way clustering with $c = 5$ and $c = 10$ are performed. In each experiment, we randomly chose 50 or 100 documents from the c selected newsgroups and 100 runs were conducted for each algorithm to obtain statistically reliable clustering result. The means and standard deviations of the test results were recorded in Table 1. The number in the parenthesis (50 or 100) indicates the number of random documents chosen from the newsgroups sets.

In all experiments, each document is represented as a term frequency vector. Let $\Pi = \{f_1, f_2, \ldots, f_m\}$ be the complete vocabulary set of the document corpus after the stopwords removal and words stemming operations, The term frequency vector X_i of document d_i is defined as

$$X_i = [x_{1i}, x_{2i}, \ldots, x_{mi}]^T \tag{6}$$

$$x_{ji} = t_{ji} \cdot \log(\frac{n}{idf_j}) \tag{7}$$

where t_{ji}, idf_j, n denote the term frequency of word $f_j \in \Pi$ in document d_i, the number of documents containing word f_j, and the total number of documents in the corpus, respectively. Using n documents from the corpus, we construct an $m \times n$ term-document matrix X. This process can be completed by using the text to matrix generator (TMG) [13] code.

The comparable methods were also implemented under the same experimental setting, including Kmeans, p-Kmeans [14], p-QR [14], Spectral [12], and LPI [3]. It is noted that LTPI achieves the best accuracy and normalized mutual information in all six random datasets. Under accuracy measure, p-Kmeans and p-QR perform better than Kmeans. Kmeans outperforms the p-Kmeans and p-QR methods in two datasets under normalized mutual information measure. The results of statistical significance test show that LTPI is significantly more accurate than the other methods for most of the datasets.

Table 1. Performance Comparison of Different Clustering Methods under the Accuracy Measure

Data sets	Accuracy (%)					
	Kmeans	p-Kmeans	p-QR	Spectral	LPI	LTPI
NG2/NG3/NG4/NG5/NG6 (50)	42.32±6.91	45.76±6.34	46.10±5.83	49.04±6.95	54.83±8.03	56.34±5.37
NG2/NG3/NG4/NG5/NG6 (100)	40.51±7.86	43.42±7.07	42.93±5.02	46.82±4.91	56.29±6.23	61.53±5.29
NG2/NG9/NG10/NG15/NG18 (50)	56.37±8.52	57.59±9.38	58.71±9.14	66.57±8.79	80.55±11.72	82.31±9.98
NG2/NG9/NG10/NG15/NG18 (100)	54.21±10.35	55.83±9.18	54.97±10.61	71.53±9.63	83.26±11.43	85.79±10.35
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (50)	45.82±7.39	47.91±6.75	48.79±6.76	53.75±6.09	61.42±7.86	62.59±6.58
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (100)	42.87±6.63	43.04±7.26	45.81±5.42	55.73±4.97	61.58±7.21	64.71±6.42
Statistical significance test based on accuracy values (p-value = 0.05)						
NG2/NG3/NG4/NG5/NG6 (50)	Kmeans< p-QR< p-Kmeans Spectral LPI LTPI					
NG2/NG3/NG4/NG5/NG6 (100)	Kmeans< p-QR< p-Kmeans (Kmeans) Spectral LPI LTPI					
NG2/NG9/NG10/NG15/NG18 (50)	Kmeans< p-Kmeans< p-QR (Kmeans) Spectral LPI LTPI					
NG2/NG9/NG10/NG15/NG18 (100)	Kmeans< p-Kmeans< p-QR (Kmeans) Spectral LPI LTPI					
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (50)	Kmeans< p-Kmeans< p-QR (Kmeans) Spectral LPI< LTPI					
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (100)	Kmeans< p-Kmeans< p-QR Spectral LPI LTPI					

Note that " " (" ") indicates that schems on the right are significantly better (worse) than the schemes on the left, and "<" (">") indicates that the relationship is not significant. In all results of statistical significance tests, the expression A < B < C means the relationship between A and C is A < C.

Table 2. Performance Comparison of Different Clustering Methods under the Normalized Mutual Information Measure

Data sets	Normalized mutual information (%)					
	Kmeans	p-Kmeans	p-QR	Spectral	LPI	LTPI
NG2/NG3/NG4/NG5/NG6 (50)	28.64±9.38	28.97±8.71	29.02±8.56	32.41±7.18	37.89±7.35	42.32±9.53
NG2/NG3/NG4/NG5/NG6 (100)	21.87±7.36	21.09±9.42	21.36±9.61	29.39±5.82	43.75±7.94	44.94±8.36
NG2/NG9/NG10/NG15/NG18 (50)	40.37±9.53	39.87±10.84	40.28±9.73	52.49±8.31	73.61±9.58	75.01±9.46
NG2/NG9/NG10/NG15/NG18 (100)	32.65±11.21	36.81±8.95	37.69±8.70	61.52±10.27	78.84±11.83	83.26±10.72
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (50)	50.74±6.29	51.03±6.04	52.51±5.68	55.62±6.22	60.76±6.81	62.89±7.35
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (100)	39.79±7.17	42.86±7.53	43.18±6.25	57.34±5.73	63.12±6.06	65.41±6.80
Statistical significance test based on accuracy values (p-value = 0.05)						
NG2/NG3/NG4/NG5/NG6 (50)	Kmeans< p-Kmeans < p-QR Spectral LPI LTPI					
NG2/NG3/NG4/NG5/NG6 (100)	p-Kmeans < p-QR< Kmeans Spectral LPI LTPI					
NG2/NG9/NG10/NG15/NG18 (50)	p-Kmeans < p-QR< Kmeans Spectral LPI LTPI					
NG2/NG9/NG10/NG15/NG18 (100)	Kmeans p-Kmeans< p-QR Spectral LPI LTPI					
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (50)	Kmeans< p-Kmeans< p-QR (Kmeans) Spectral LPI< LTPI					
NG1/NG5/NG7/NG8/NG11/ NG12/NG13/NG14/NG15/NG17 (100)	Kmeans< p-Kmeans< p-QR Spectral LPI LTPI					

4. Conclusion

In this paper, we present a new spectral clustering based on local topology preserving indexing. It performs clustering in the embedding space, which is obtained by the proposed LTPI. The LTPI captures the local topology information embedded in the low-dimensional manifold. It simultaneously minimizes the distances between the data points in the same local patches and maximizes the distances between the distances outside these patches. The experiments for document clustering on NG20 data set show that the proposed LTPI method outperforms other classical clustering methods. The construction of graphs and weights in graph embedding is still an open problem, which would be a possible future work.

Acknowledgements

This work was supported by the project of the University Research Program in Education Deptartment of Hainan Province (HJKJ 2012-59) and the Research Program of Qiongtai Teachers College (QTKY 201020).

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, vol. 41, **(1990)**, pp. 391-407.
[2] X. H. P. Niyogi, "Locality preserving projections", Proceedings of Conference on Advances in Neural Information Processing Systems, **(2003)**, pp. 585-591; Vancouver, Canada.
[3] D. Cai, X. F. He and J. W. Han, "Document clustering using locality preserving indexing", IEEE Transactions on Knowledge and Data Engineering, vol. 17, **(2005)**, pp. 1624-1637.
[4] G. Lu and J. Zou, "Feature Extraction Using a Complete Kernel Extension of Supervised Graph Embedding", Neural Processing Letters, vol. 35, **(2012)**, pp. 159-175.
[5] H. Wang, "Structured sparse linear graph embedding", Neural Networks, vol. 27, **(2012)**, pp. 38-44.
[6] C. Wang, C. Chen, L. J. Zhang, J. J. Bu and W. Chen, "Constrained laplacian eigenmap for dimensionality reduction", Neurocomputing, vol. 73, **(2010)**, pp. 951-958.
[7] J. B. Tenenbaum, V. de Silva and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction", Science, vol. 290, **(2000)**, pp. 2319-2323.
[8] M. B. P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering", Proceedings of Conference on Advances in Neural Information Processing Systems, **(2001)**, pp. 585–591, MIT Press, Cambridge.
[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, **(2000)**, pp. 2323-2326.
[10] D. Cai, X. He, S. Yan and H. Zhang, "Neighborhood preserving embedding", Proceedings of International Conference on Computer Vision (ICCV'05), **(2005)**, Beijing, China.
[11] C. Park, B. C. Kang, Z. W. Sur and M. G. Cho, "Document clustering of medline abstracts based on non-negative matrix factorization using local confidence assessment", Biochip Journal, vol. 4, **(2010)**, pp. 336-349.
[12] G. Wang, D. Cheng, R. Kannan and S. Vempala, "A divide-and-merge methodology for clustering", ACM Transactions on Database System, vol. 31, **(2006)**, pp. 1499-1525.
[13] E. G. D. Zeimpekis, "Design of a matlab toolbox for term-document matrix generation", Proceedings of Workshop on Clustering High Dimensional Data and its Applications at the 5th SIAM International Conference on Data Mining (SDM05), **(2005)**, pp. 38-48; Newport Beah, Canada.
[14] X. He, H. Zha, C. Ding, H. Simon and M. Gu, "Spectral relaxation for k-means clustering" Neural Information Processing Systems, **(2001)** December, pp. 1057-1064; Vancouver, Canada.

Author

Jieqing Xing

He received the M.S. degree in Computer Techniques and Application from Chongqing University, China in 2005. He is currently the Associate Professor of the Dept. of Information Technology at Qiongtai Teachers College in China. His research interests include membrane computing and spectral clustering.