Spectral Clustering on Manifolds with Statistical and Geometrical Similarity

Yong Cheng[1] and Qiang Tong[2]

[1] Department of Computer Sciences,
Beijing University of Chemical Technology
chengyong@ieee.org
[2] School of Information Technology,
University of International Business and Economics
tongqiang@uibe.edu.cn

Abstract. The problem of clustering data has been driven by a demand from various disciplines engaged in exploratory data analysis, such as medicine taxonomy, customer relationship management and so on. However, Most of the algorithms designed to handle data in the form of point clouds fail to cluster data that expose a manifold structure. The high dimensional data sets often exhibit geometrical structures which are often important in clustering data on manifold. Motivated by the fact, we believe that a good similarity measure on a manifold should reflect not only the statistical properties but also the geometrical properties of given data. We model the similarity between data points in statistical and geometrical perspectives, then a modified version of spectral algorithm on manifold is proposed to reveal the structure. The encouraging results on several artificial and real-world data set are obtained which validate our proposed clustering algorithm.

Keywords: Clustering; Manifold Learning; Spectral Clustering.

1 Introduction

Clustering data with similar features into clusters has been studied in a wide range of literatures and many effective algorithms have been proposed[1]. Despite the success they fail to cluster data that expose a manifold structure, like speech, motion and images, that generally exist in the form of paths in a high-dimensional space. Manifold learning is a learning scheme that characterizes a possibly non-linear manifold on which the data would lie. Popular manifold learning techniques include Locally Linear Embedding (LLE)[2], Hessian LLE[3], ISOMap[4], Laplacian Eigenmaps[5] and so on.

Differing from the manifold learning that primarily discovers a manifold embedding of input data, manifold clustering attempts to partition a set of data into several different clusters each of which contains data points originating from a separate low-dimensional manifold. There exist several works that cluster the data on manifolds. In [6], R. Souvenir et al. presented an approach to factor low-rank manifold of data that originate from multiple, intersecting low-dimensional

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part I, LNCS 6063, pp. 422–429, 2010.

manifolds. In their method, two novel technical contributions are highlighted: node-weighted multidimensional scaling and a fast algorithm for weighted low-rank approximation for rank-one weight matrices. Q. Guo et al. proposed a manifold clustering approach with energy minimization strategy [7]. In which an energy function has been defined by weighted components on Euclidean distance between two consecutive and discrete curvature of manifolds. The function then be minimized by tabu search to find the locally optimal sequence of the data. Finally, clusters are generated by breaking the sequence and merging some isolated points. It is worthy of noting that the approach only works on 2-D and 3-D data space. In [8], R. Haralick et al. described a new cluster model LMCLUS which is based on the concept of linear manifolds. In order to detect clusters embedded in lower dimensional linear manifolds, LMCLUS uses the strategy of hierarchical-divisive procedure and random projection via sampling and histogram thresholding to construct trial linear manifolds of various dimensions.

In this paper, we are motivated by the fact that high dimensional data sets often exhibit geometrical structures which are important to be considered in clustering data on manifold. We believe that a good similarity measure on a manifold should reflect not only the statistical properties but also the geometrical properties of data sets. We then propose a manifold clustering algorithm SCM which uses spectral analysis to drive the clusters and evaluate the proposed algorithm on several synthetic and real-world data set, the results obtained indicate the improvement in the model quality and give additional insights into the data.

The remainder of this article is organized as follows. The problem of manifold clustering is formulated in Section 2. In Section 3, we first proposed a variant of spectral clustering in which the graph is constructed with the statistical and geometrical similarity, then, the algorithm SCM is described in details. The experimental results on synthetic and real-world are reported in Section 4. Finally, Section 5 concludes the paper.

2 Problem Formulation

Generally speaking, the goal of manifold clustering is to find clusters with an intrinsic dimensionality that is much smaller than the dimensionality of the data set. The problem of clustering data on multiple complex manifold structure can be described as follows: Suppose a set of points $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ that derived from S intersecting manifolds, where $\mathbf{x}_i \in \mathbb{R}^d$, n is the size of X, d is the dimension and S is the possible number of manifold. The goal of clustering is to partition the data into S clusters each of which corresponds to a manifold. The output of clustering generally requires a set of labels $C = \{c_1, c_2, \cdots, c_n\}$, where $c_i \in \{1, 2, \cdots, S\}, 1 \leq i \leq n$.

3 SCM: Spectral Clustering on Manifold

3.1 Motivation

The basic goal in clustering analysis is to group data objects with similar features together. In manifold clustering, most existing algorithms utilize certain traditional similarity measure that emphasizes the statistical properties in given data set. However, we believe that, in additional to statistical properties, high dimensional data sets often exhibit geometrical structures which are often important to be considered in clustering data on manifold: similar data samples should have similar geometrical properties. Furthermore, a good similarity measure on a manifold should reflect not only the statistical properties but also the geometrical properties of given data set, which is the direct motivation of proposed approach in this paper. In our approach, the statistical property of similarity is often implemented with nearest neighborhood technique and the geometrical property of given data sample is measured with data objects whose distances are smaller than a threshold value ε.

3.2 Modeling the Similarity

As introduced in the above, we consider the statistical properties and geometrical properties to model the similarity between data objects. Like most of the traditional clustering methods, the k nearest neighborhood techniques is chosen to calculate the similarity in the perspective of statistical properties. In addition, the similarity in geometrical property of given data object is modeled with ε neighborhood in our consideration. The k nearest neighborhood and ε neighborhood are defined as follows:

Definition 1 (k-neighborhood). *$\mathcal{N}_i(k)$ is a set of data points in dataset X that contains k nearest points of $\mathbf{x}_i$.*

Definition 2 (ε-neighborhood). *$\mathcal{N}_i(\varepsilon)$ is a set of data points in dataset X that contains points satisfying $\|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon$, i.e.*

$$\mathcal{N}_i(\varepsilon) = \{\mathbf{x}_j | \|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon, \mathbf{x}_j \in X\} \tag{1}$$

k-neighborhood and ε-neighborhood describe the local relationship of a data point $\mathbf{x}_i$ in statistical and geometrical aspects respectively. Thus, we can define the k-neighborhood and ε-neighborhood graph as follows:

Definition 3 (k-neighbor Graph). *Given a set of points $X = \{\mathbf{x}_i, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, and the k-neighborhood of each point $\mathbf{x}_i, 1 \le i \le n$. A weighted undirected graph $G_k = \langle V_k, E_k \rangle$ is constructed from the given data set X where,*

- *$V_k = X$, i.e. that each vertex in the G_k corresponds a data object in X;*
- *$E_k = \{\mathbf{x}_i\mathbf{x}_j | \mathbf{x}_i \in \mathcal{N}_j(k)\}$, that is two data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected if and only if one is in the k-neighborhood of the other;*
- *The weight w_{ij}^k of edge connecting $\mathbf{x}_i$ and $\mathbf{x}_j$ is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$;*

Definition 4 (ε-neighbor Graph). *Given a set of points $X = \{\mathbf{x}_i, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, and the ε-neighborhood of each point $\mathbf{x}_i, 1 \le i \le n$. A weighted undirected graph $G_\varepsilon = \langle V_\varepsilon, E_\varepsilon \rangle$ is constructed from the given data set X where,*

- *$V_\varepsilon = X$, i.e. that each vertex in the G_ε corresponds a data object in X;*
- *$E_\varepsilon = \{\mathbf{x}_i\mathbf{x}_j | \mathbf{x}_i \in \mathcal{N}_j(\varepsilon)\}$, that is two data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected if and only if one is in the ε-neighborhood of the other;*
- *The weight w^{ε}_{ij} of edge connecting $\mathbf{x}_i$ and $\mathbf{x}_j$ is the weighted similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$;*

In general, there exists three possible choices to weight the edge in graph G. The simplest method for building the weighted graph is the *binary* weighting approach, the second approach is *Gaussian Kernel(GK)* which is often applied in many application, the final method to estimate the edge is given by the locally linear embedding technique and the weight w_{ij} can be calculated by solving an optimization problem [2].

Definition 5 (Similarity Graph). *Given a set of points $X = \{\mathbf{x}_i, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, and the k-neighborhood and ε-neighborhood of each point $\mathbf{x}_i, 1 \le i \le n$. The k-neighbor graph $G_k = \langle V_k, E_k \rangle$ and ε-neighbor graph $G_\varepsilon = \langle V_\varepsilon, E_\varepsilon \rangle$ can be constructed according to definition (3) and definition (4). Thus, a weighted graph that combines the statistical and geometrical similarity will be created as $G = \langle V, E \rangle$, where*

- *$V = X$, i.e. that each vertex in the G corresponds a data object in X;*
- *$E = \{E_k \cup E_\varepsilon\}$, that is two data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected if one is in the k-neighborhood or ε-neighborhood of the other;*
- *The weight w_{ij} of edge connecting $\mathbf{x}_i$ and $\mathbf{x}_j$ is the linearly combined weighted similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, and*

$$w_{ij} = \alpha w^k_{ij} + (1-\alpha) w^{\varepsilon}_{ij} \quad \alpha (0 \le \alpha \le 1) \textit{ is the coefficient} \tag{2}$$

As mentioned above, the graph $G = \langle V, E \rangle$ considers the two local relationship of a data point $\mathbf{x}_i$, i.e. the k-neighborhood and ε-neighborhood which are incorporated into a unified graph representation via linear combination. It is worthy of noting that the graph is not symmetrical, like most approaches in the literature, here, we also use the symmetrical graph by $W' = \frac{1}{2}(W + W^\top)$. Without loss the generality, W denotes the symmetrical similarity in the remainder of the paper. After the data points and local relationship representation in graph is obtained, we can derive the clusters using the spectral analysis.

3.3 SCM Algorithm

Spectral clustering works by detecting the clusters in data set via analyzing the eigenvectors of graph Laplacian. In other words, the multiplicity k of the eigenvalue 0 of unnormalized graph Laplacian L or normalized graph Laplacian L_{norm} equals the number of connected components in the graph. The graph Laplacian L and L_{norm} are computed as follows,

$$L = D - W \tag{3}$$

$$L_{norm} = D^{-1/2} L D^{-1/2} \tag{4}$$

where D is the degree matrix of W, $D_{ii} = \sum_j w_{ij}, 1 \le i \le n$.

Following the general framework of spectral clustering, we can derive our spectral clustering variant that clusters data on a manifold considering the statistical and geometrical similarity. The proposed algorithm is described in Figure (1).

Algorithm 1. SCM: Spectral Clustering Algorithm on Manifold

```
Input: Data set $X$, statistical and geometrical neighborhood $k, \varepsilon$, the
       bandwidth of kernel $\sigma$, the number of clusters $K$;
Output: Clusters $C_1, C_2, \cdots, C_K$;
Construct the $k$-neighborhood graph with statistical similarity, Let $W_k$ be its weighted matrix;
Construct the $\varepsilon$-neighborhood graph with geometrical similarity, Let $W_\varepsilon$ be its weighted matrix;
Construct a similarity graph by considering the statistical and geometrical similarity with Equation (2). Let $W = \alpha W_k + (1 - \alpha) W_\varepsilon$ be its weighted matrix;
Compute the unnormalized Laplacian $L$ or the normalized Laplacian $L_{norm}$;
Compute the first $K$ generalized eigenvectors $u_1, u_2, \cdots, u_K$ of the generalized eigen problem $Lu = \lambda Du$;
Let $U \in \mathbb{R}^{n \times K}$ be the matrix containing the vectors $u_1, u_2, \cdots, u_K$ as columns;
For $i = 1, \cdots, n$, let $y_i \in \mathbb{R}^K$ be the vector corresponding to the $i$-th row of $U$;
Cluster the points $(y_i)_{i=1,\cdots,n}$, in $\mathbb{R}^K$ with the $k$-means algorithm into clusters $C_1, \cdots, C_K$;
```

In the first step, the complexity for compute statistical k-neighborhood for all points is $O(n^3 \cdot \log(n))$, while the complexity is $O(n^3)$ for the geometrical ε-neighborhood, where n is the number of data points. The later three steps are implemented in a spectral clustering algorithm, the time complexity of the implemented algorithm depends on the complexity of the eigenvalue decomposition algorithm is about $O(n^3)$, where n is the number of rows/columns. In most cases, it is possible to reduce the time complexity, since the algorithm needs certain eigenvectors only (which are corresponding to smallest or largest eigenvalues in magnitude). Thus, the proposed algorithm has about the same time complexity with the classic spectral clustering.

4 Experimental Results

We experiment the proposed algorithm on several synthetic and real-world data from UCI machine-learning repository [9]. Our spectral is in fact a general variant of classic spectral clustering. There have several parameters need to be set. In all experiments, we select the Gaussian Kernel to measure similarity between two data points and the bandwidth of kernel σ is set to 2. In addition, we treat the statistical similarity and geometrical similarity equally, so the coefficient α is set to 0.5. For the purpose of comparing performance of clustering, RandIndex is selected to evaluate the performance of our algorithm. In each experiment, we run the algorithm on each data set for 10 times, then the mean RandIndex value is calculated.

4.1 Synthetic Data

We select two synthetic data sets, namely “twomoons” and “spiral” data for evaluating our proposed algorithm. The “twomoons” and and “spiral” data set have become the standard benchmarks in numerous other manifold related experiments. Note that these synthetic data cannot be clustered in a meaningful way by certain methods that assume the data form a compact shape. The original data set of “twomoons” and “spiral” are illustrated in Figure (1) and (5), respectively.

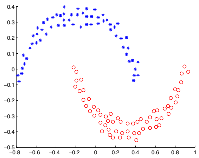

Fig. 1. Two moons

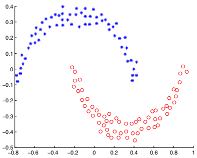

Fig. 2. $k = 4, \varepsilon = 0.2$

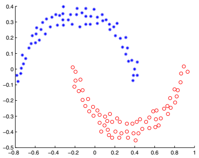

Fig. 3. $k = 1, \varepsilon = 0.2$

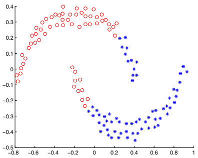

Fig. 4. $k$means

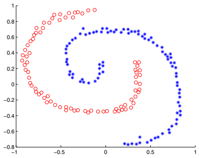

Fig. 5. Spiral

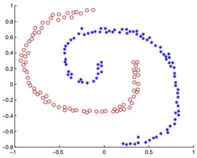

Fig. 6. $k = 4, \varepsilon = 0.2$

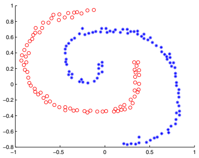

Fig. 7. $k = 1, \varepsilon = 0.2$

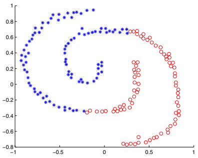

Fig. 8. $k$means

The clustering results of our proposed algorithm are shown in figure (2) to (3) for twomoons data and figure (6) to (7) for spiral data. We can see our algorithm work very well with different configuration of parameters. Figure (4) and (8) are the results that the k-means algorithm returns. It is obvious that our algorithm can find much more meaningful clusters than the classic k-means for the data on the manifold.

4.2 Real-World Dataset

We also test our algorithms on several real-world data sets from the UCI machine-learning repository [9]. Four data, namely “Soybean”, “Vowel”, “Iris” and “Zoo” are selected to measure the performance with RandIndex. All data are multi-attirubte and multi-class that are considered to situate on the manifolds. The basic information of these data sets is described in the Table (1).

In our experiment, different configuration of parameters are tested to compare the performance improved. Four k-neighborhood, i.e. $k = 0, 10, 20, 30$ nearest neighbors are constructed as the statistical similarity, and four ε-neighbors, i.e.

Table 1. Data set description from UCI

Name	Num of Insts	Num of Attrs	Num of Disc.	Num of Cont.	Num of Class
Soybean	683	36	36	0	19
Vowel	990	14	4	10	11
Iris	150	5	4	1	3
Zoo	101	18	17	1	7

$\varepsilon = 0, 0.5, 2, 5$ geometrical neighborhoods are constructed as the geometrical similarity. Thus, our proposed algorithm is tested on the possible statistical and geometrical similarity of 15 combination to measure the clustering performance, the coefficient is set to 0.5 in all experiments (except for $k = 0$ and $\varepsilon = 0$). The experiment results are shown in the Table. (2) to (5).

Table 2. Soybean data

k	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 2$	$\varepsilon = 5$
$k = 0$	N/A	.8307	N/A	.7552
$k = 10$	.8334	.8338	.8494	.8099
$k = 20$	.7760	.7761	.7779	.7775
$k = 30$	.7994	.8030	.6860	.8503

Table 3. Vowel data

k	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 2$	$\varepsilon = 5$
$k = 0$	N/A	N/A	N/A	.8286
$k = 10$	.8011	.8160	.8141	.8277
$k = 20$	.8090	.7951	.8157	.8312
$k = 30$	.8164	.8342	.8359	.8280

Table 4. Iris data

k	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 2$	$\varepsilon = 5$
$k = 0$	N/A	.7766	.8797	.8797
$k = 10$	.8759	.8859	.8797	.8797
$k = 20$	.9017	.9055	.8797	.8797
$k = 30$	.9124	.9124	.8787	.8737

Table 5. Zoo data

k	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 2$	$\varepsilon = 5$
$k = 0$	N/A	.6945	.6669	.9002
$k = 10$	.8996	.8976	.8996	.8996
$k = 20$	.7826	.7869	.8994	.8994
$k = 30$	.8924	.8994	.8994	.8994

From these figures, we can see the proposed algorithm can certainly improve the performance of clustering on manifolds. For the “soybean” data, the best clustering results is obtained when $k = 30, \varepsilon = 5$. Our algorithm uses the statistical and geometrical similarity to describe the local neighborhood relationship and outperform classic spectral clustering (using k-neighborhood or ε-neighborhood) and k-means algorithm whose randIndex is .5427.

5 Conclusions and Future Work

In this paper, we proposed a variant of spectral clustering that perform well on several data sets. Our work is motivated by the fact that similar data points in high dimensional data sets often exhibit similar geometrical structures. Differing from most existing algorithms that emphasize the statistical properties in given data set, we consider a good similarity measure on a manifold should reflects not only the statistical properties but also the geometrical properties of given data set. We use the k-neighborhood and ε-neighborhood to model the statistical and geometrical aspects respectively and derived the new variant of spectral clustering. An extensive experiments have been conducted to test the quality of the clusters produced by proposed algorithm on a varied collection of artificial and real-world data set. The results indicate promising performance that validate our proposed algorithm.

Acknowledgments

This work is supported by the young teacher project in BUCT under contract (No. QN0824). Qiang Tong thanks the 211 project in UIBE under contract (No. 32046).

References

1. Han, J., Kamber, M.: Data Mining - Concepts and Techniques, 2nd edn. China Machine Press, Beijing (2007)
2. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
3. Donoho, D.L., Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. In: Falkow, S. (ed.) Proceedings of the National Academy of Sciences of the United States of America, May 2003, vol. 100, pp. 5591–5596 (2003)
4. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
5. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
6. Souvenir, R., Pless, R.: Manifold clustering. In: Proceedings of the 10th International Conference on Computer Vision (ICCV 2005), October 17-21, vol. 1, pp. 648–653 (2005)
7. Guo, Q., Li, H., Chen, W., Shen, I.F., Parkkinen, J.: Manifold clustering via energy minimization. In: Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA 2007), December 13-15, pp. 375–380. IEEE Computer Society, Los Alamitos (2007)
8. Haralick, R., Harpaz, R.: Linear manifold clustering. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), vol. 3587, pp. 132–141. Springer, Heidelberg (2005)
9. Asuncion, A., Newman, D.: UCI machine learning repository (2009)