Research Article

Spectral Clustering with Local Projection Distance Measurement

Chen Diao,¹ Ai-Hua Zhang,¹ and Bin Wang²

¹College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

²School of Electrical Engineering, Northwest University for Nationalities, Lanzhou 730030, China

Correspondence should be addressed to Ai-Hua Zhang; lutzhangah@163.com

Received 18 November 2014; Revised 15 March 2015; Accepted 27 March 2015

Academic Editor: Xin-She Yang

Copyright © 2015 Chen Diao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Constructing a rational affinity matrix is crucial for spectral clustering. In this paper, a novel spectral clustering via local projection distance measure (LPDM) is proposed. In this method, the Local-Projection-Neighborhood (LPN) is defined, which is a region between a pair of data, and other data in the LPN are projected onto the straight line among the data pairs. Utilizing the Euclidean distance between projective points, the local spatial structure of data can be well detected to measure the similarity of objects. Then the affinity matrix can be obtained by using a new similarity measurement, which can squeeze or widen the projective distance with the different spatial structure of data. Experimental results show that the LPDM algorithm can obtain desirable results with high performance on synthetic datasets, real-world datasets, and images.

1. Introduction

As an unsupervised classification technique, clustering has been successfully applied to exploratory data analysis, such as image segmentation [1–3], data mining [4, 5], signal analysis [6], gene expression analysis [7], sport activities analysis in sport domain [8], and other subjects [9–11]. During the last decade, a number of clustering algorithms were adequately developed. Nonetheless, many of these algorithms are not effective for data classification when applied on nonconvex data space. Compared with the classical clustering, spectral clustering (SC) [12] has been successfully used to identify irregularly shaped datasets and is supported by linear algebra theory [13].

SC can be regarded as a type of partition problem for undirected graph [14]. In partition problem, a set of data points are represented by a similarity graph. Each data point can be considered as a vertex, and the similarities between pairs of data points can be considered as the edge weights in a graph. Through a partition of the graph, the data points can be clustered into different subgraphs such that the edges between different subgraphs have relatively low weights, in comparison to those within the same subgraph. It is well known that inchoate graph partition methods, such as min-cut [15], tend to generate unbalanced solutions and they are extraordinarily sensitive to noise [16]. In order to overcome

these drawbacks, many spectral clustering algorithms were proposed, such as normalized cut [17], ratio cut [18], min-max cut [19], and Ng-Jordan-Weiss (NJW) [20], which employed diverse criteria for spectral clustering to optimize the quality of graph partition. In these methods, as we know, Gaussian kernel function is chosen as similarity function, but it cannot offer any local spatial structure of the dataset.

To further improve the performance of spectral clustering algorithms, Chen and Feng [16] presented a semisupervised SC based on Near Strangers or Distant Relatives model, which is a generalization of SC algorithm. In [21], Li and Guo proposed a novel affinity matrix generation approach, which can adaptively adjust the similarity measure of data points, based on the spatial structure of dataset. To obtain insensitive scaling parameter, Zelnik-Manor and Perona [22] developed a self-tuning SC algorithm to estimate the scaling parameter. Via M-estimation statistics, Chang and Yeung [23] proposed a robust path-based SC algorithm. In [24], the local density can be employed to adjust the scaling parameter. Nevertheless, the method needs to set the parameter empirically [13]. To solve the problem, Yang et al. [25] proposed a density sensitive function, which can either elongate the similarity measure or shorten it in different density regions.

From a graph-cut perspective, based on minimizing the sum of edge weights between data points, SC can partition an undirected weighted graph into disjoint components. The

information of the adjacency relations between data points is contained within the affinity matrix A . Most of existing methods exploited the local structure of dataset to construct a rational affinity matrix, which is one of the key issues of SC and greatly affects partition results.

As we know, data points with high similarity should have uniform density and consistent spatial characteristic [26]. Therefore, the key to estimate whether a pair of data points belong to a specific cluster is how to use the data information between them. The local projection distance measure (LPDM) presented in the current paper can reflect the local spatial structure of dataset in more depth, by which the goal of rational partition for synthetic datasets, images, and most of real-world datasets may be well achieved.

The main contributions of the current paper are three-fold. (1) The concept of Local-Projection-Neighborhood is introduced, which is a spatial area among data points and is an important source to obtain the local spatial structure of datasets. (2) A measure for local projection distance is presented, which facilitates embodying an accurate local structure of dataset. (3) A novel similarity measure is defined that can adaptively adjust the measure of similarity based on the local spatial structure of datasets and is insensitive to the parameter on UCI datasets.

The outline of the rest of this paper is as follows. In Section 2, spectral clustering algorithm is briefly discussed as preliminary. Section 3 introduces the LPDM algorithm. The performance of the presented approach is evaluated in Section 4. Section 5 is the conclusion.

2. Spectral Clustering

SC algorithms can be regarded as solving graph-cut problems which are extensively applied to exploratory data analysis. In this section, as a preliminary, we will briefly review spectral clustering, which is closely related to LPDM algorithm.

In this paper, SC can be considered as an undirected weighted graph-cut problem. For a dataset $X = \{x_1, x_2, \dots, x_n\}$, the weights of the graph can be constructed by the adjacency matrix $A \in R^{n \times n}$. Specifically, the element A_{ij} of the adjacency matrix is formulated as

$$A_{ij} = \begin{cases} \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma^2}\right) & i \neq j \\ 0 & i = j, \end{cases} \quad (1)$$

where σ is the scaling parameter determining the neighborhood and $d(x_i, x_j)$ is the distance of points x_i and x_j . If element $A_{ij} = 0$, this means that there is no link between them. The diagonal degree matrix D is constructed as $D_{ii} = \sum_{j=1}^n A_{ij}$. SC uses the similarity information to group data points into predefined K clusters.

The steps of NJW approach are listed as follows.

- (1) Calculate the similarity A_{ij} of data points by (1) to construct the affinity matrix $A \in R^{n \times n}$ and degree matrix D .
- (2) Compute the normalized affinity matrix $L = D^{-1/2} A D^{-1/2}$.

- (3) The first K largest eigenvalues and the corresponding eigenvectors $v^1, v^2, \dots, v^K$ are calculated to construct the matrix $V = [v^1, v^2, \dots, v^K] \in R^{n \times K}$ with v^i the column vector.
- (4) Normalize each row of the matrix V (i.e., $Y_{ij} = V_{ij}/(\sum_j V_{ij}^2)^{1/2}$) to construct the matrix Y .
- (5) Group data points by k -means method in a new space, which is spanned by the rows of the matrix Y .

Remark 1. Via spectral clustering, data points are mapped into a convex dataset in another space, and then k -means clustering can be used to group the image data points in the new space. In SC methods, NJW is widely applied to data analysis; thus, it is adopted in this paper.

The affinity matrix of classical spectral clustering is usually constructed by the Gaussian kernel function, but it could not well represent the space structure of datasets and lead to irrational results. Aiming at the problem, the LPDM algorithm is devised in the next section.

3. Affinity Matrix Construction for Spectral Clustering through Local Projection Distance Measure

As the most important part of the paper, some general problems on three similarity measurement algorithms are briefly discussed in the first part of this section. In the second part, to overcome these problems, a novel LPDM algorithm will be introduced.

3.1. Similarity Function Analysis. As we know, the Gaussian kernel function is employed in most of the existing SC methods. In most cases, since the scaling parameter σ is fixed and has to be set manually, Gaussian kernel function cannot objectively reflect the local spatial structure of datasets and reasonably figure out the similarity between data points, especially as the similarity function is applied to complex datasets. Figure 1 illustrates the high impact on the clustering by σ . It is evident that the results of NJW algorithm are to be affected greatly by the scaling parameter in Gaussian kernel function.

Unlike setting a fixed scaling parameter, the parameter σ_i in self-tuning spectral clustering (SC-ST) [22] can be calculated based on the neighborhood of point x_i as

$$\sigma_i = d(x_i, x_k), \quad (2)$$

where x_k is the k th neighbor of point x_i . Unfortunately, the affinity matrix in SC-ST is still constructed by Gaussian kernel function which is less valid in many cases [24]. In Figure 2, the method cannot yield better clustering result, failing to classify the Three-Spiral-Arms datasets.

Zhang et al. [24] addressed the Common-Near-Neighbor CNN(a, b) for spectral clustering (SC-DA) and defined a novel similarity function $A_{ij}(a, b)$ as

$$A_{ij}(a, b) = e^{-d(a, b)^2 / (2\sigma^2(\text{CNN}(a, b) + 1))}, \quad (3)$$

FIGURE 1: Clustering results (using NJW algorithm) with different values of σ on Three-Circle dataset. (a) $\sigma = 0.1$, (b) $\sigma = 0.2$, (c) $\sigma = 0.4$, and (d) $\sigma = 0.6$.

where $CNN(a, b)$ represents the local density of the overlapped area in data space and the region can be determined by the data points a and b with radius ϵ . The result of SC-DA algorithm on the Three-Spiral-Arms dataset is shown in Figure 3. It is evident that the algorithm produces the correct clustering result.

Utilizing CNN, the parameter σ can be adjusted adaptively. However, the approach requires setting the parameter σ manually for the correct clustering [13]. Figure 4 shows that SC-DA fails to classify the dataset and the unrevealed structures in this dataset cannot be discovered. It reveals that, in some cases, the similarity of data points cannot be properly reflected by Euclidean distance.

It is generally considered that if data points fall into the same cluster, the distribution of points should have

similar patterns and concordant density. Nevertheless, in some cases, CNN could not correctly estimate the local density of complex datasets. Let us survey the synthetic dataset. In Figure 5, it is easy to find that the data points a , b , c , and d belong to the same cluster.

For three pairs of data points, the parameters of the CNN, the Euclidean distance (L_{ED}), the similarity (S_{GA}) by (1), and the novel similarity (S_{DA}) by (3) are calculated, respectively, and the results are summarized in Table 1.

From Table 1, we can find that the L_{ED} between data point b and others are approximately the same. That is to say, those similarities between each of the three pairs of points are similar and they can be reflected by the parameter of S_{GA} . The parameter of CNN reflects the local density among data points according to SC-DA and it can be used to estimate

FIGURE 2: Three-Spiral-Arms dataset example. (a) Original dataset; (b) clustering result of SC-ST algorithm.

FIGURE 3: Results of SC-DA algorithm on Three-Spiral-Arms dataset.

TABLE 1: Attributes for four data points; here σ is set to 0.1.

Data points	CNN	L_{ED}	S_{GA}	S_{DA}
a, b	143	0.5617	1.4247×10^{-7}	0.8963
b, c	2	0.5754	0.6465×10^{-7}	0.0040
b, d	3	0.5572	1.8121×10^{-7}	0.0206

the similarity of point pairs. However, as can be seen from Table 1, the CNN and S_{DA} of the pair a and b are much larger than others' and this implicates that the two points probably do not belong to the same cluster. Apparently these four data points locate in the same cluster. CNN can adaptively adjust

the scaling parameter σ in Gaussian kernel function and reduce the impact of the fixed σ to some extent. Nonetheless, CNN merely reflects the local density of the geometric center between two data points. Therefore, the local structure of dataset cannot be fully described by CNN.

Combining the analysis of the three SC methods in this subsection, it is clear that, in some sense, the similarity of the correlative points could not be rightly reflected, based on Gaussian kernel function. How to obtain local spatial structure of datasets and construct an appropriate affinity matrix will be addressed in next subsection.

3.2. Local Projection Distance Measure. The motivation of LPDM algorithm originated from the idea that, in order to

FIGURE 5: Shared neighbors on the synthetic dataset.

construct an appropriate affinity matrix, we should know spatial structure about the neighborhoods of the correlative points as much as possible. Therefore, in this subsection, we define the Local-Projection-Neighborhood (LPN) and propose a novel density sensitive similarity measure.

With a dataset $X = \{x_1, x_2, \dots, x_n\}$ given in R^d , the LPN(x_i, x_j) of the pair x_i and x_j is the overlapped region with specified Euclidean radius $d(x_i, x_j)$, around the center points x_{cp} . The center points of the region can be calculated as

$$\begin{aligned} \|x_{cp} - x_i\| &= d(x_i, x_j), \\ \|x_{cp} - x_j\| &= d(x_i, x_j), \end{aligned} \quad (4)$$

where x_{cp} is a center point of the sphere region and $d(x_i, x_j)$ is the radius. Here, the three points x_{cp} , x_i , and x_j form an equilateral triangle with the side lengths $d(x_i, x_j)$. Therefore, the center point x_{cp} can be obtained by solving (4).

The idea of LPDM algorithm is to discover the unrevealed configuration patterns of local dataset, with the spatial structure of the data points x_l in LPN. Thus, how to obtain the points in LPN is a problem. Because the data points in LPN are located in the overlapped area between two sphere regions, they can be obtained as

$$\begin{aligned} (x_l - x_{cp1})^2 &\leq d^2(x_i - x_j), \\ (x_l - x_{cp2})^2 &\leq d^2(x_i - x_j). \end{aligned} \quad (5)$$

Now consider the data points which are dispersedly located in LPN. How to achieve the similarity measurements from the spatial structure of these points is a key for LPDM algorithm. In this study, the point x in LPN is projected onto the straight line connecting the points x_i and x_j , where the

projective point is denoted by $\tilde{x}$. The k th coordinate of point $\tilde{x}$ can be calculated as

$$\tilde{x}^k = -\frac{\sum_{l=1}^d (x_i^l - x^l)(x_j^l - x_i^l)}{\sum_{l=1}^d (x_i^l - x_j^l)^2} (x_j^k - x_i^k) + x_i^k, \quad (6)$$

where x^k is the k th coordinate of point x .

Evidently, the straight line connecting the points x_i and x_j can be divided into several line segments by these projective points. As we know, data points being close in space and with uniform density might possibly tend to belong to the same cluster. The Euclidean distance between projective points in LPN represents the local similarity among the data points. If data points are located in the same cluster, the consistent projection distances exist in LPN. The structure information of local dataset can be reflected by the quantity and the length of the line segment.

Remark 2. We know that this method stresses the local spatial structure and can avoid seeking the shortest path that connects any two nodes in an undirected graph.

In this subsection, the local structure of datasets being reflected by projection distances in LPN may seem obvious. How to obtain a meaningful similarity measure among data points by applying projective distance is crucially important in LPDM algorithm. Here, a novel similarity measure is addressed, which is motivated by the discussions in [25].

The new adjustable projection distance is defined as

$$L(x_a, x_b) = 10^{\rho d(x_a, x_b)}, \quad (7)$$

where $d(x_a, x_b)$ is the Euclidean distance between the points x_a and x_b and ρ is the flexing factor. The Euclidean distance $d(x_a, x_b)$ can be enlarged or shortened by the nonlinear function (7).

As we know, the similarity of the pair of points can be reflected by the distance between the two points. Nevertheless, the pair of points with a longer distance might still belong to the same cluster with a large number of points uniformly distributed between them. Therefore, the length of the line segment connecting projective image points in LPN can be adjusted by the nonlinear function (7). According to the spatial structure of the local dataset, a new measure of distance of the pair of points can be obtained through a summation of the values of the length of these line segments. The new distance between two points x_i and x_j can be obtained as

$$PD(x_i, x_j) = \sum_{l=0}^m L(x_l, x_{l+1}), \quad (8)$$

where m is the number of projective points in LPN. Notice that the points x_0 ($l = 0$) and x_{m+1} ($l = m$) are x_i and x_j , respectively.

The similarity of pair points is inversely proportional to their distance. Therefore, the similarity could be computed as

$$S(x_i, x_j) = \frac{1}{PD(x_i, x_j)}. \quad (9)$$

FIGURE 6: Local projection distance measure.

The novel similarity metric can highlight the diversity of the local structure of datasets and avoid seeking the shortest path in graph.

Here, the following example will be exhibited to illustrate the processes of LPDM algorithm. Consider four data points $x_a(0, 0)$, $x_b(3, 4)$, $x_c(0, 3)$, and $x_d(1, 2)$, shown in Figure 6.

The similarity of the pair of points x_a and x_b will be calculated by the following steps.

Step 1. Compute the center x_{cp} of points x_a and x_b with specified Euclidean radius $d(x_i, x_j)$ by (4):

$$\begin{aligned} \sqrt{(x_{cp}^1)^2 + (x_{cp}^2)^2} &= 5 \\ \sqrt{(3 - x_{cp}^1)^2 + (4 - x_{cp}^2)^2} &= 5. \end{aligned} \quad (10)$$

The results are $x_{cp1} = (3/2 + 2\sqrt{3}, 2 - 3\sqrt{3}/2)$ and $x_{cp2} = (3/2 - 2\sqrt{3}, 2 + 3\sqrt{3}/2)$.

Step 2. Find the data points in LPN by (5). According to the distances

$$\begin{aligned} \|x_c - x_{cp1}\| &= 2.5321, \\ \|x_d - x_{cp1}\| &= 3.9416, \\ \|x_c - x_{cp2}\| &= 6.1309, \\ \|x_d - x_{cp2}\| &= 4.7396 \end{aligned} \quad (11)$$

we can find that only the point x_d belongs to LPN.

Step 3. Compute the projective point of the point x_d by (6)

$$\begin{aligned} \tilde{x}_d^1 &= -\frac{(x_a^1 - x_d^1)(x_b^1 - x_a^1) + (x_a^2 - x_d^2)(x_b^2 - x_a^2)}{(x_b^1 - x_a^1)^2 + (x_b^2 - x_a^2)^2} (x_b^1 - x_a^1) \\ &\quad + x_a^1 = 1.320, \\ \tilde{x}_d^2 &= -\frac{(x_a^1 - x_d^1)(x_b^1 - x_a^1) + (x_a^2 - x_d^2)(x_b^2 - x_a^2)}{(x_b^1 - x_a^1)^2 + (x_b^2 - x_a^2)^2} (x_b^2 - x_a^2) \\ &\quad + x_a^2 = 1.760. \end{aligned} \quad (12)$$

Step 4. Calculate the distance between points x_a and x_b by (7) and (8):

$$PD(x_a, x_b) = L(x_a, \tilde{x}_d) + L(x_b, \tilde{x}_d) = 789.4467, \quad (13)$$

where ρ is set to 1.

Step 5. The similarity of the pair of points x_a and x_b can be estimated by (9):

$$S(x_a, x_b) = \frac{1}{PD(x_a, x_b)} = 0.0013. \quad (14)$$

It is worth mentioning that the k -nearest neighbor is adopted to construct the affinity matrix in LPDM. Employing the structural information about the neighborhoods of the correlative points and novel density sensitive similarity measure, the LPDM algorithm achieves high accuracies of spectral clustering. The clustering results on the synthetic datasets achieved via LPDM algorithm are shown in Figure 7. The algorithm can obtain desired cluster results for this dataset.

In order to reduce the computational complexity, k -nearest neighbor is used to construct the affinity matrix [27]. According to the neighbor propagation principle [21], it is unnecessary to obtain all affinity relationships among data points because neighbor propagation could fully describe the structure of the dataset.

In this subsection, the LPDM algorithm is presented. Contrary to the classical method of similarity measure based on Gaussian kernel function, the similarity among data points achieved via LPDM algorithm can be constructed directly and the value of the measure reflects more the local spatial structure of datasets. The performance of this algorithm will be further illustrated in Section 4.

4. Experimental Results and Analysis

In the section, a number of experiments are conducted to evaluate the performance of the LPDM algorithm and the sensitivity of the parameters of LPDM algorithm will be further analyzed. The experimental results distinctly manifest the advantage of LPDM algorithm. In order to illustrate the procedure of the experiment, our experiments are conducted in the following subsection. Firstly, four SC algorithms are

FIGURE 7: Clustering results of LPDM.

applied to several synthetic datasets and real-world datasets. The clustering accuracies of different algorithms can be examined with two small-size datasets. Then, LPDM algorithm is executed on larger datasets to evaluate the performance of our method. All experiments are implemented in Matlab 7.12 environment on a PC with Intel CPU 1.6 GHz and 4 GB memory.

In our clustering experiments, clustering accuracy (Acc) [28] and Rand Index (RI) [29] are used to assess the

performance of LPDM algorithm. The Acc is defined as

$$\text{Acc} = \frac{\max_{\text{map}} \sum_{i=1}^n \delta(y_i, \text{map}(c_i))}{n}, \quad (15)$$

where y_i and c_i are the true clustering results and experimental result of original data, respectively. $\delta(a, b)$ is the quantity of data points that constitute both the true clustering result a and the practical cluster b . $\text{map}(\cdot)$ is a function which can map all cluster labels to the corresponding labels.

It is a known fact that there exist $n(n-1)/2$ potential pairwise decisions to estimate whether each pair of data points belong to the same cluster, where n is the size of dataset. RI is used to evaluate clustering accuracy and its value is proportional to the clustering performance, which is defined as

$$RI = \frac{CD}{TD}, \quad (16)$$

where CD denotes the quantity of correct decisions and TD denotes the quantity of total decisions.

4.1. Parameter Selection. For NJW, SC-DA, SC-ST, and LPDM algorithm, the parameters of SC algorithm need to be set for the above experiments. In order to obtain a reasonable scale parameter, Iris dataset from UCI datasets is used to evaluate the quality of the scale parameter. In NJW and SC-DA algorithm, the scale parameter σ needs to be set. In Figures 8 and 9, we can find that NJW and SC-DA gain better performances, when scale parameters σ are set to 0.2 and 0.01, respectively. k is set to 7 in all our experiments in accordance with SC-ST in [22]. In LPDM, the flexing factor ρ is set to 12 and 7% of the size of dataset is adopted as the parameter k of neighborhood size when LPDM is implemented on each dataset.

4.2. Synthetic Data Experiments. In this subsection, four synthetic datasets of arbitrary shape and various densities are used to test the accuracy of the four SC algorithms. Experimental results are presented in Figure 10.

The first row of synthetic data is the Two-Moon dataset in Figure 10. It is evident that the data points of each moon should belong to the same cluster. Since the dataset includes nonconvex separate clusters and the two “moons” are very close, the classification of the dataset is a difficult task for SC algorithms. Figure 10 shows that SC-ST cannot rationally classify the Two-Moon dataset whereas the other SC approaches correctly identify the genuine clusters. In Figure 10, the second row of toy data includes diverse density of data and it is a challenging clustering problem. According to the results, we can find that SC and SC-DA cannot classify them effectively, implicating that both SC and SC-DA are less suitable for the multiscale clustering problem. For the remaining two synthetic datasets, four SC algorithms can obtain the expected clusters. In conclusion, the rational classifications can be obtained for all these synthetic datasets by applying LPDM. Thus, the algorithm can well handle different clustering problems.

4.3. Real Datasets Experiments. As we know, the UCI [30] datasets and the MNIST handwritten digits database [31] have been widely used for testing SC algorithms in the clustering problem.

In this subsection, both datasets are used in our experiments to evaluate the performance of proposed approach. In the UCI databases, we perform experiment on five datasets including Wilt, Wine Quality, Ionosphere, Zoo, and Abalone. The dimension of the data is the number of attributes varying

FIGURE 8: Experiment about parameter σ of NJW.

FIGURE 9: Experiment about parameter σ of SC-DA.

from 6 to 34. Table 2 describes the characteristics of these datasets. Unlike the toy data, the dimension of MNIST database is much higher. Each image of the handwritten digit has been normalized and centered to 28×28 gray-level image. In this experiment, four subsets $\{6, 9\}$, $\{1, 6\}$, $\{1, 2, 3\}$, and $\{0, 1, 3, 4\}$ are selected to test the LPDM and 200 examples for each digit are randomly chosen from the MNIST training dataset. The basic characteristics of these datasets are summarized in Table 3.

For the UCI datasets, the clustering results are summarized in Figure 11. From Figure 11, we can find that LPDM outperforms others in accordance with Acc and RI. Taking the Wine Quality dataset as the example, one will see that

FIGURE 10: (a) Clustering results of SC algorithm; (b) clustering results of SC-ST algorithm; (c) clustering results of SC-DA algorithm; (d) clustering results of LPDM algorithm.

TABLE 2: Attributes of UCI datasets.

Dataset	Number of instances	Number of attributes	Number of clusters
Wilt	4339	6	2
Wine Quality	4898	12	7
Ionosphere	351	34	2
Zoo	101	18	7
Abalone	4177	8	29

TABLE 3: Attributes of MNIST handwritten digits datasets.

Dataset	Number of instances	Number of attributes	Number of clusters
{6, 9}	400	$28 \times 28 = 784$	2
{1, 6}	400	$28 \times 28 = 784$	2
{1, 2, 3}	600	$28 \times 28 = 784$	3
{0, 1, 3, 4}	800	$28 \times 28 = 784$	4

the method can obtain the accuracy of 0.8000 in accordance with Acc, and the others are 0.3833, 0.5000, and 0.3500, respectively. For the Abalone dataset, the clustering accuracy

of LPDM is seemingly lower than other datasets, but the performance of LPDM is still superior to other methods.

The experimental results on MNIST datasets are summarized in Figure 12. As can be seen in the figure, the accuracy of LPDM is higher than SC, SC-ST, and SC-DA. For the subset {0, 1, 3, 4}, the accuracy of the four methods is 0.630, 0.6825, 0.5425, and 0.7125 by Acc, respectively. For the four subsets, despite the similarity of accuracies between SC-ST and LPDM, the accuracy of SC-ST is a little lower than the LPDM. It proves that a more reasonable affinity matrix can be constructed by LPDM.

4.4. Image Segmentation Experiments. Image segmentation is one of the applications of SC. The SC algorithm can be easily evaluated by the results of image segmentation and we can learn whether the results “look good,” whether an algorithm works only on small-size datasets, and so on. Here, LPDM algorithm is applied to image segmentation and its ability can be intuitively evaluated by observation. In Figure 13, two original images (a) and (d) with the size of 960×1280 and 64×48 in jpg format are used in this experiment, which are chosen from [20]. To reduce the cost of computation and memory space, we resize the image (a) to the size of 96×128 and the sizes of the two images (a) and (d) are

FIGURE 11: Comparison of four SC algorithms. (a) Experimental results on UCI by Acc; (b) experimental results on UCI by RI.

FIGURE 12: Comparison of four SC algorithms. (a) Experimental results on MNIST by Acc; (b) experimental results on MNIST by RI.

12288 pixels and 3072 pixels, respectively. As we know, it is difficult for SC to segment the salient objects from the complicated background, especially for images with large number of pixels. In contrast, as can be seen from Figure 13, the child and the fire hydrant are partitioned successfully from the backgrounds of images (a) and (d).

4.5. Parameter Sensitiveness. In the last part of the experiments, the parameter sensitiveness of the LPDM approach is studied and the stability of the algorithm depends on its two parameters: the flexing factor ρ and the neighborhood size k . In this algorithm, two parameters ρ and k are required to be adjusted for clustering. The setting of parameters ρ and

FIGURE 13: Testing images and segmentation results. Subplots (a) and (d) are original images. Subplots (b), (e), (d), and (f) show the components of the partition of images (a) and (d), the desired and the background objects, respectively.

k is the crucial problem of LPDM. Here, Wilt, Wine Quality, Ionosphere, Zoo, and Abalone datasets are applied to evaluate the sensitiveness of two parameters.

For the parameter ρ , the algorithm is evaluated with $k = 0.07$. The parameter interval of ρ is $[10, 210]$. Figures 14(a) and 14(b) show the Acc rate and RI rate of LPDM on the five datasets. We can see that changes in the different intervals of the parameter ρ have less impact on the Acc rate and RI rate, respectively. Apparently, the algorithm could work well under ρ in the interval $[10, 30]$. Figures 14(c) and 14(d) show that LPDM is insensitive to k , except the Wine Quality dataset. Hence, it is necessary to adopt the value of k in the interval $[0.05, 0.13]$. Experimental results show that, in most cases, LPDM is insensitive to the parameters ρ and k in the different parameter intervals recommended in this subsection.

5. Conclusion

A local projection distance measurement for spectral clustering is proposed in this paper, which utilizes projective

data points in LPN to detect the local spatial structure of the distribution of datasets. Employing a novel density sensitive similarity measure, local spatial structural information of datasets can be exploited and converted into the similarity measure of a pair of data points. Meanwhile, k -nearest neighbor sparse strategy is adopted to reduce both the computational difficulties and memory assumptions. The numerical results presented show that the local projection distance measure approach is able to correctly cluster many synthetic datasets, UCI datasets, MNIST handwritten digits database, and images and is less sensitive to parameters than other classical SC approaches.

There are still many problems awaiting us to offer solutions. For instance, how to automatically and effectively set several specific parameters in our algorithm is to be dealt with as the future work.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

FIGURE 14: Experiment about parameter sensitivity. (a) Clustering accuracy curves with Acc under different ρ values at $k = 0.07$; (b) clustering accuracy curves with RI under different ρ values at $k = 0.07$; (c) clustering accuracy curves with Acc under different k values at $\rho = 12$; (d) clustering accuracy curves with RI under different k values at $\rho = 12$.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant 81360229), the Research Fund for the Doctoral Program of Higher Education (Grant 20116201110002), the Open Project Program of the National Laboratory of Pattern Recognition (Grant 201407347), and

the Natural Science Foundation of Gansu Province (Grants 1308RJZA225 and 145RJ2A065).

References

- [1] A. Rajendran and R. Dhanasekaran, "Enhanced possibilistic fuzzy C-means algorithm for normal and pathological brain

- tissue segmentation on magnetic resonance brain image,” *Arabian Journal for Science and Engineering*, vol. 38, no. 9, pp. 2375–2388, 2013.
- [2] S. Ghaffarian and S. Ghaffarian, “Automatic histogram-based fuzzy C-means clustering for remote sensing imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 97, pp. 46–57, 2014.
 - [3] S. Zeng, R. Huang, Z. Kang, and N. Sang, “Image segmentation using spectral clustering of Gaussian mixture models,” *Neurocomputing*, vol. 144, pp. 346–356, 2014.
 - [4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–53, 1996.
 - [5] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.
 - [6] F. Mohamad, T. S. Cevat, A. Mehdi, and P. Farzad, “Fracture characteristics of AISI D2 tool steel at different tempering temperatures using acoustic emission and fuzzy C-means clustering,” *Arabian Journal for Science and Engineering*, vol. 38, no. 8, pp. 2205–2217, 2013.
 - [7] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: a survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.
 - [8] I. Fister Jr., I. Fister, D. Fister, and S. Fong, “Data mining in sporting activities created by sports trackers,” in *Proceedings of the International Symposium on Computational and Business Intelligence (ISCBI '13)*, pp. 88–91, August 2013.
 - [9] N. Cai, J.-W. Cao, H.-Y. Ma, and C.-X. Wang, “Swarm stability analysis of nonlinear dynamical multi-agent systems via relative Lyapunov function,” *Arabian Journal for Science and Engineering*, vol. 39, no. 3, pp. 2427–2434, 2014.
 - [10] N. Cai, J. Cao, M. Liu, and H. Ma, “On controllability problems of high-order dynamical multi-agent systems,” *Arabian Journal for Science and Engineering*, vol. 39, no. 5, pp. 4261–4267, 2014.
 - [11] N. Cai, J. Cao, and M. J. Khan, “A controllability synthesis problem for dynamic multi-agent systems with linear high-order protocol,” *International Journal of Control, Automation and Systems*, vol. 12, no. 6, pp. 1366–1371, 2014.
 - [12] F. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, RI, USA, 1997.
 - [13] K. Taşdemir, “Vector quantization based approximate spectral clustering of large datasets,” *Pattern Recognition*, vol. 45, no. 8, pp. 3034–3044, 2012.
 - [14] B. Mohar, “The Laplacian spectrum of graphs,” in *Graph Theory, Combinatorics, and Applications*, Y. Alavi, G. Chartrand, O. Ollermann, and A. Schwenk, Eds., vol. 2, pp. 871–898, Wiley, New York, NY, USA, 1991.
 - [15] E. L. Johnson, A. Mehrotra, and G. L. Nemhauser, “Min-cut clustering,” *Mathematical Programming*, vol. 62, no. 1–3, pp. 133–151, 1993.
 - [16] W. Chen and G. Feng, “Spectral clustering: a semi-supervised approach,” *Neurocomputing*, vol. 77, no. 1, pp. 229–242, 2012.
 - [17] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
 - [18] L. Hagen and A. B. Kahng, “New spectral methods for ratio cut partitioning and clustering,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 11, no. 9, pp. 1074–1085, 1992.
 - [19] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, “A min-max cut algorithm for graph partitioning and data clustering,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM '01)*, pp. 107–114, San Jose, Calif, USA, 2001.
 - [20] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: analysis and an algorithm,” in *Advances in Neural Information Processing Systems 14*, pp. 849–856, 2001.
 - [21] X. Y. Li and L. J. Guo, “Constructing affinity matrix in spectral clustering based on neighbor propagation,” *Neurocomputing*, vol. 97, pp. 125–130, 2012.
 - [22] L. Zelnik-Manor and P. Perona, “Self-tuning spectral clustering,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 17, pp. 1601–1608, 2004.
 - [23] H. Chang and D.-Y. Yeung, “Robust path-based spectral clustering,” *Pattern Recognition*, vol. 41, no. 1, pp. 191–203, 2008.
 - [24] X. Zhang, J. Li, and H. Yu, “Local density adaptive similarity measurement for spectral clustering,” *Pattern Recognition Letters*, vol. 32, no. 2, pp. 352–358, 2011.
 - [25] P. Yang, Q. Zhu, and B. Huang, “Spectral clustering with density sensitive similarity function,” *Knowledge-Based Systems*, vol. 24, no. 5, pp. 621–628, 2011.
 - [26] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, “Learning with local and global consistency,” in *Proceedings of the Neural Information Processing Systems (NIPS '04)*, vol. 16, pp. 321–328, 2004.
 - [27] F. Zhao, H. Liu, and L. Jiao, “Spectral clustering with fuzzy similarity measure,” *Digital Signal Processing*, vol. 21, no. 6, pp. 701–709, 2011.
 - [28] F. Zhao, L. Jiao, H. Liu, X. Gao, and M. Gong, “Spectral clustering with eigenvector selection based on entropy ranking,” *Neurocomputing*, vol. 73, no. 10–12, pp. 1704–1717, 2010.
 - [29] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
 - [30] C. C. Blake and C. J. Merz, “UCI repository of machine learning databases,” <http://www.ics.uci.edu/mllearn/MLRepository.html>.
 - [31] Y. LeCun and C. Cortes, “The MNIST database of handwritten digits,” 2009, <http://yann.lecun.com/exdb/mnist/>.