International Journal of Machine Learning and Cybernetics (2021) 12:541–553
https://doi.org/10.1007/s13042-020-01187-3

ORIGINAL ARTICLE

M-pSC: a manifold *p*-spectral clustering algorithm

Ling Ding[1] · Shifei Ding[1,2] · Yanru Wang[1] · Lijuan Wang[1] · Hongjie Jia[1,3]

Received: 30 December 2019 / Accepted: 5 August 2020 / Published online: 14 August 2020

Abstract

Since *p*-spectral clustering has good performance in many practical problems, it has attracted great attention. The Cheeger cut criterion is used in *p*-spectral clustering to do graph partition. However, due to the improper affinity measure and outliers, the original *p*-spectral clustering algorithm is not effective in dealing with manifold data. To solve this problem, we propose a manifold *p*-spectral clustering (M-pSC) using path-based affinity measure. First, we design a path-based affinity function to describe the complex structures of manifold data. This affinity function obeys the clustering assumption that the data pairs within the manifold structure share high affinities, and the data pairs between different manifold structures share low affinities. This will help us construct a good affinity matrix, which carry more category information of the points. Then we propose a M-pSC algorithm using the path-based affinity function. In the Cheeger cut criterion, the *p*-Laplacian matrix are constructed based on the manifold affinity function, and the final clustering results are obtained by using the eigenvectors of graph *p*-Laplacian. At last, the proposed algorithm is tested on several public data sets and the experiments show that our algorithm is adaptive to different manifold data. Compared with other popular clustering algorithms, our algorithm has good clustering quality and robustness.

Keywords *p*-Laplacian matrix · Clustering · Manifold distance · Affinity measure

1 Introduction

In the field of data mining, clustering is one of the most widely used methods of exploratory data analysis. In many scientific field, when dealing with the collected data, people try to identify groups with similar behavior in the data. Clustering analysis can discover the distribution structure among data objects.[1]. Based on a similarity measure, the cluster analysis method groups data by maximizing the similarity of data objects located in the same cluster and minimizing the similarity of data objects between different clusters. The final data clusters are structured such that data objects located in the same cluster have relatively high affinities, while objects between different clusters have low affinities [2]. Cluster analysis methods are widely used in many fields, including pattern recognition, market research, image segmentation, etc. It can be used as an independent analysis tool, and can also be used as a preprocessing in other data analysis methods.

Traditional clustering analysis algorithms, such as *k*-means like clustering algorithms and hierarchical clustering algorithms, are all built on the convex spherical sample space [3]. These algorithms are suitable for clustering in convex sample space. But for non-convex data space, these algorithms are easy to fall into local optimum and cannot correctly identify clusters. In order to cluster in the sample space of arbitrary shape, a new clustering algorithm—spectral clustering algorithm has begun to attract the attention of scholars. The theoretical basis of spectral clustering is spectral theory, which transforms data clustering into graph segmentation, especially for non-convex data sets [4–6]. It converts the clustering problem into the optimal partitioning problem of an undirected weighted graph [7]. Spectral

✉ Shifei Ding
dingsf@cumt.edu.cn

✉ Hongjie Jia
jiahj@ujs.edu.cn

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2] Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou 221116, China

[3] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

clustering regards the data samples as the vertices on the graph, and uses the similarities between the samples to represent the weight of the edge of the graph. Since the optimal solution of the solution graph partition is a NP hard problem, the spectral clustering method relaxes the problem to a continuous space to solve it. According to the spectral graph theory, this solution is essentially solving the spectral decomposition problem of the Laplacian matrix of the graph [8]. Therefore, the spectral clustering can be considered as a the continuous relaxation solution for the graph partition problem, which is an approximate clustering method.

Spectral clustering is a series of clustering algorithms based on different graph cut criteria. One of the graph cut criteria is Cheeger cut, which tends to segment the balanced subgraphs [9]. As mentioned above, the optimal partition of Cheeger cut is also a NP hard problem. We can use eigenvectors of *p*-Laplacian matrix to solve Cheeger cut problem and obtain the approximate optimal graph partition. This method is called *p*-spectral clustering [10]. The *p*-Laplacian matrix used in *p*-spectral clustering is a generalized version of normalized graph Laplacian matrix.

Manifold is a generalization of Euclidean space in the case of large-scale analysis. Euclidean space is its special case. The manifold has a homeomorphic mapping locally with the Euclidean space. Specifically, a one-dimensional manifold is a curve, and a two-dimensional manifold is a curved surface. High-dimensional manifolds are analogs of curves and surfaces in high-dimensional situations. The manifold can usually be considered a hyperplane in a small scale. A large number of high-dimensional data have lower-dimensional manifold embedding, and scholars have studied the essential features of these massive data to reveal the inherent laws of things. Gaussian kernel is based on Euclidean distance, but Euclidean distance is not a proper distance measure for manifold data [11, 12]. Although *p*-spectral clustering has a good theoretical basis, it doesn't perform well on manifold data sets [10]. Because *p*-spectral clustering needs to calculate the affinities between data points and Gauss kernel function is a commonly used similarity measure. Gauss kernel function is based on Euclidean distance, which is only suit for the data points in Euclidean space. When dealing with the data set with manifold structures, Gauss kernel function is not a good affinity measure. Real world data sets often has complex structures. The affinities simply calculated by Euclidean distance may lose important information about the data. As *p*-spectral clustering is based on the affinity matrix, improper affinity measure may reduce the clustering quality. The development of the modern society has put forward higher request on data analysis technology. Traditional *p*-spectral clustering algorithms are unable to meet these requirements and the Euclidean distance may cause bad clustering results. Therefore, designing and developing effective affinity measurement to describe the complex data has become a hotspot in *p*-spectral clustering research.

Suppose a good clustering should reflect the local and global structure of clustering. Under this assumption, we design a path-based affinity measure for manifold data. In affinity measurement, we design a density-adjustable length and use the shortest path to describe the affinities. In this way, the internal affinities of the manifold are high, while the external affinities of different manifold are low. Then we apply the manifold affinity measure to *p*-spectral clustering to improve its clustering quality, and present a M-pSC algorithm using path-based affinity measure (M-pSC). M-pSC algorithm can correctly describe the complex relationship of manifold data, and use graph *p*-Laplacian solve Cheeger cut problem. The effectiveness of the proposed algorithm is verified through a lot of experiments.

The paper structures are: Sect. 2 introduces the related works of *p*-spectral clustering and affinity measure; Sect. 3 presents a manifold *p*-spectral clustering (M-pSC) method using path-based affinity measure; Sect. 4 gives the theoretical analysis of the proposed M-pSC algorithm; Sect. 5 utilizes benchmark data sets to test the performance of the proposed algorithm; the last section is conclusion.

2 Related work

The graph partitioning criteria of *p*-spectral clustering is Cheeger cut, Cheeger cut considers the connections within the cluster and the connections outside clusters at the same time, so it can produce balanced clusters. Because of the good theoretical support, there are a lot of research works about *p*-spectral clustering. We summarize these related works from three aspects: eigen-decomposition, affinity measure, manifold learning, as shown in Table 1.

p-Spectral clustering needs to calculate the eigenvectors of *p*-Laplacian matrix, but this process is very time-consuming. In order to reduce the complexity of eigen-decomposition, Frederix and Van Barel [13] use the approximate Laplacian matrix and sparse method in spectral clustering. Binkiewicz et al. [14] study the potential clusters in a graph with the covariates of vertex and propose a node-contextualized randomized block model in spectral clustering. Ariascastro et al. [15] construct a weighted graph using the local principal components analysis to solve the intersection problems of spectral clustering. Law et al. [16] develop a new deep metric learning model for spectral clustering, in which the clustering complexity is reduced by linear gradient method.

The affinity measure plays an important role for *p*-spectral clustering. Nie et al. [17] measure the affinity relationship by searching the appropriate neighbors for every point

Table 1 Related works about *p*-spectral clustering

Research theme	Algorithm	Author
Eigen-decomposition	Sparse spectral clustering algorithm based on the incomplete Cholesky decomposition	Frederix and Van Barel [13]
	Covariate-assisted spectral clustering algorithm	Binkiewicz et al. [14]
	Spectral clustering algorithm based on local PCA	Ariascastro et al. [15]
	Deep spectral clustering algorithm	Law et al. [16]
Affinity measure	Projected clustering algorithm with adaptive neighbors	Nie et al. [17]
	Approximate spectral clustering algorithm with utilized similarity information	Tasdemir et al. [18]
	Fuzzy similarity measure based spectral clustering algorithm	Goyal et al. [19]
Manifold learning	Spectral clustering algorithm for multiple manifolds	Wang et al. [20]
	Adaptive kernel spectral clustering algorithm	Langone et al. [21]
	Constrained spectral clustering algorithm via the randomized projected power method	Zhi et al. [22]

based on the local distance of data. Tasdemir et al. [18] propose a hybrid geodesic affinity measure, in which the geodesic relationships is formed by data topology, distance and density information at the same time. To deal with noise-contaminated images, Goyal et al. [19] add local spatial information to the fuzzy similarity measure to remove noise in the image and improve the quality of image segmentation.

In order to improve the performance of *p*-spectral clustering on manifold data, Wang et al. [20] design a multi-view clustering model, which utilize multiple geometric information to describe the manifold structure of data. Langone et al. [21] improve the kernel spectral clustering algorithm to detect outliers in manifold data and apply the proposed algorithm in health monitoring applications. Zhi et al. [22] use pairwise constraints and balanced constraints to describe the relations of manifold data, and then use random projection method to optimize the objective function of constrained clustering to obtain clustering results.

Although many research results of *p*-spectral clustering have been achieved, it still has many problems worth studying. In traditional *p*-spectral clustering, the affinity matrix is usually formed by radial basis function, which cannot describe complex data structures very well, especially the manifold data structure. To enhance the clustering quality of *p*-spectral clustering on manifold data, we design a path-based affinity function to measure the affinity relationship between data objects. The proposed affinity measure will assign high affinities for adjacent data objects and the objects with the same manifold structures. Then the affinity function is combined with *p*-spectral clustering to process the data set with complex structures. This method is insensitive to parameters and can recognize non-linear distribution of data points or overlapping areas. Experimental results demonstrate that our algorithm has good performance on synthetic datasets and benchmark datasets.

3 Proposed approach

3.1 *p*-Spectral clustering

The theoretical basis of spectral clustering is algebraic graph theory. In spectral clustering, we generally use the graph $G=(V, E)$ to describe the data set X. The data points in X are viewed as the vertices of graph G. These vertices form set V and the edges connecting any two vertices in V form set E. In graph G, each edge has a weight that represents the affinity between the two vertices of the edge. Suppose V_1 and V_2 are two subsets of V, and $V_1 \cup V_2 = V$. The graph cut of set V is given by Eq. (1):

$$\mathrm{cut}(V_1, V_2) = \sum_{i \in V_1, j \in V_2} w_{ij} \tag{1}$$

where w_{ij} is the affinity between vertex i and vertex j.

In *p*-spectral clustering, w_{ij} is usually calculated by radial basis function:

$$w_{ij} = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{2}$$

where $\left\|x_i - x_j\right\|$ is the Euclidean distance between point x_i and point x_j, σ is the scale parameter, controlling the changes of distances.

A graph partition can be obtained by minimizing Eq. (1). However, it may separates a single point from the others in some cases and result in bad graph partitions. Since clusters should be reasonably large groups of points, one way to avoid this problem is to explicitly request that the sets V_1, …, V_k are “reasonably large”. Cheeger cut criterion is designed to encode this. Cheeger cut is denoted as Ccut [23]:

$$\mathrm{Ccut}(V_1, V_2) = \frac{\mathrm{cut}(V_1, V_2)}{\min\{|V_1|, |V_2|\}} \tag{3}$$

where $|V_1|$ is the number of data points in set V_1, $\min\{|V_1|, |V_2|\}$ is used to measure the size of subset V_1 and V_2 to get more balanced clusters.

Minimizing Eq. (3) will lead to the optimal graph partition that the similarities within a cluster are as large as possible, while the similarities between clusters are as small as possible. But calculating the optimal Cheeger cut is an NP-hard problem because of the balancing conditions. Wagner and Wagner [24] point out that “the more balanced the partition, the harder the problem". Relaxation is a way to solve the Cheeger cut problem. We will see that relaxing Ccut leads to p-spectral clustering,

According to the Rayleigh quotient principle, the relaxed solution of Cheeger cut is related to the eigenvectors of p-Laplacian matrix. Liu et al. [25] define the inner product form of graph p-Laplacian Δ_p as follows:

$$\langle f, \Delta_p f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^p \tag{4}$$

where $p \in (1,2]$, f is the eigenvector of p-Laplacian matrix.

The Rayleigh quotient shows that we can obtain the eigenvalues of p-Laplacian Δ_p as local minima of the functional F_p:

$$F_p(f) = \frac{\langle f, \Delta_p f \rangle}{\|f\|_p^p} \tag{5}$$

where $\|f\|_p^p = \sum_{i=1}^{n} |f_i|^p$.

Theorem 1 in Sect. 4.2 gives the mathematical connection between $F_p(f)$ and Cheeger cut objective function, which indicate that Cheeger cut problem can be converted to the eigenvalue problem of function $F_p(f)$. Therefore, we can get a relaxed approximate solution of $\mathrm{Ccut}(V_1, V_2)$ by minimizing $F_p(f)$, and the best partition is obtained when $F_p(f)$ reaches the minimum value:

$$\lambda_p = \arg\min_{p \to 1} F_p(f) \tag{6}$$

where λ_p is the eigenvalue of graph p-Laplacian corresponding to vector f.

p-Spectral clustering provides an effective way to solve Cheeger cut problem in polynomial time using p-Laplacian operator. Specifically, if we set an appropriate threshold, the second eigenvector $v_p^{(2)}$ of p-Laplacian matrix will lead to a good bipartition of the graph. The optimal threshold is determined by minimizing the corresponding Cheeger cut. For each partition V_1 and V_2, the threshold t should satisfy

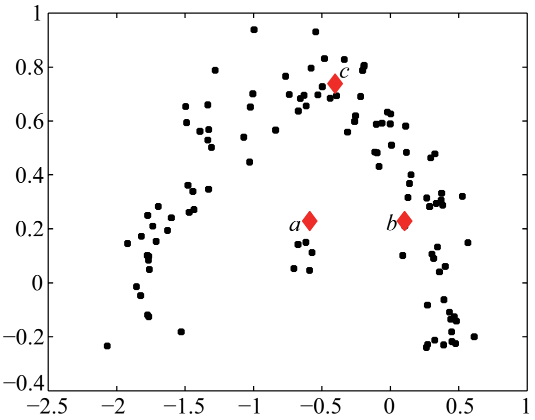


Fig. 1 Euclidean distance for manifold data

$$t = \arg\min_{V_1=\{i\in V | v_p^{(2)}(i)>t\}} \mathrm{Ccut}(V_1, V_2) \tag{7}$$

3.2 Path-based affinity measure

The standard p-spectral clustering algorithm uses Radial basis function to measure the affinity between data points. Gaussian kernel is based on Euclidean distance which describes the absolute distance of each point in space, and it is connected with the position coordinates of data points. However, in practice, the distribution of data usually has an unpredictable complex structure, and the global consistency of the data set cannot be reflected by the Euclidean distance affinity [26], that is, the data points on the same manifold should have high affinity.

Figure 1 is an example to illustrate the shortcomings of Euclidean distance. It can be seen from Fig. 1 that point b and point c are on the same manifold, point a and point b are on different manifolds. We hope that the affinity between point b and point c is greater than the affinity between point a and point b, so that it is possible to group b and c into the same cluster. However, the Euclidean distance between point a and point b is significantly smaller than the Euclidean distance between point b and point c. If we use Euclidean distance to measure the similarities between data points, point b is more similar to point a instead of point c, which means that point a and point b has higher probability to be grouped together than that of point b and point c. In other words, Euclidean distance is insufficient to represent the global consistency of the data shown in the figure. Therefore, if we simply use Euclidean distance to measure affinity, it can cause bad performance for complex clustering problems in the real world.

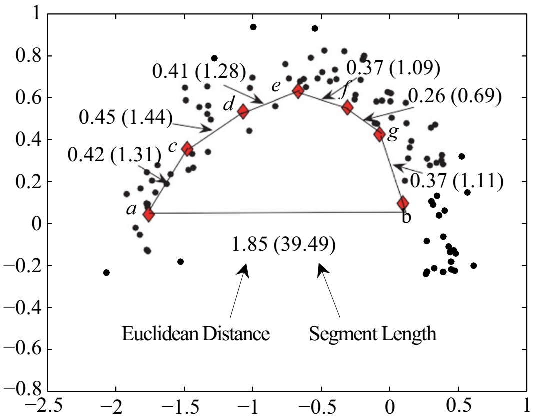


Fig. 2 Euclidean distance and the segment length on manifold

The global clustering assumption requires high similarities of data pairs in the same manifold structure, and low similarities of data pairs in different manifold structures [27, 28]. In order to keep the global consistency of the clustering, this paper presents a new affinity measurement—the path-based affinity function to evaluate the relationship of data points on manifold data sets. Its concrete form is described as follows.

Definition 1 The length of line segment on manifold:

$$L(x_i, x_j) = e^{\rho d(x_i, x_j)} - 1 \tag{8}$$

where $d(x_{i,}\, x_j) = \|x_i - x_j\|$ is the Euclidean distance between the data points x_i and x_j; ρ is called the scaling factor. We can enlarge or shorten the length of line segment between two points by adjusting the density factor ρ, and enhance the adaptability of the algorithm to different data sets. Figure 2 below shows the advantages of such definition.

In Fig. 2, the value outside the parentheses is the Euclidean distance between the two points, and the value in parentheses is its corresponding segment length on manifold (when the density factor $\rho = 2$). From Definition 1, it is known that the length of line segment on manifold is exponentially related to the Euclidean distance, and it has the effect of enlarging Euclidean distance. Observing the distance between the seven points a, b, c, d, e, f, g, when the Euclidean distance between two points is small, the corresponding segment length is also small; if two points have large Euclidean distance, their segment length will be significantly larger. This indicates that the defined segment length is sensitive to large distance changes. Although the Euclidean distance $d(a, c) + d(c, d) + d(d, e) + d(e, f) + d(f, g) + d(g, b) > d(a, b)$, the length of the line segment on manifold $L(a, c) + L(c, d) + L(d, e) + L(e, f) + L(f, g) + L(g, b) < L(a, b)$. Therefore, if two points are on the same manifold, assuming that there is a path inside the manifold to connect the two points, we can use the length of the path as the manifold high-density region. According to the length of line segment on manifold, a new distance measure—path-based distance is further defined.

Definition 2 Path-based distance: Given an undirected weighted graph $G = (V, E)$, let $path = \{v_1, v_2, \ldots, v_m\} \in V_{path}$ denote the path between vertex v_1 and v_m, where m is the number of vertices contained in $path$, the edge $(v_k, v_{k+1}) \in E$, $1 \le k < m$. Let $\Omega_{i,j}$ represent the set of all paths connecting the point pair $\{x_i, x_j\}$ $(1 \le i, j < n)$, then the path-based distance between x_i and x_j is

$$D(x_i, x_j) = \frac{1}{\rho^2} \ln\left(1 + d_{sp}(x_i, x_j)\right)^2 \tag{9}$$

where $d_{sp}(x_i, x_j) = \min\limits_{path \subset \Omega_{i,j}} \sum_{k=1}^{m-1} L(v_k, v_{k+1})$ is the distance of the shortest path between nodes x_i and x_j on graph G; $L(v_k, v_{k+1})$ is the length of line segment of two adjacent points on the shortest path from x_i to x_j on graph G. It is easy to see that the path-based distance satisfies the four constraints of distance measure:

1 Reflexivity: $D(x_i, x_j) = 0$, if and only if $x_i = x_j$.
2 Symmetry: $D(x_i, x_j) = D(x_j, x_i)$.
3 Non-negative: $D(x_i, x_j) \ge 0$.
4 Triangular inequality: $D(x_i, x_j) \le D(x_i, x_k) + D(x_k, x_j)$.

Definition 3 According to the above path-based distance measure, the manifold affinity of data points x_i and x_j is defined as

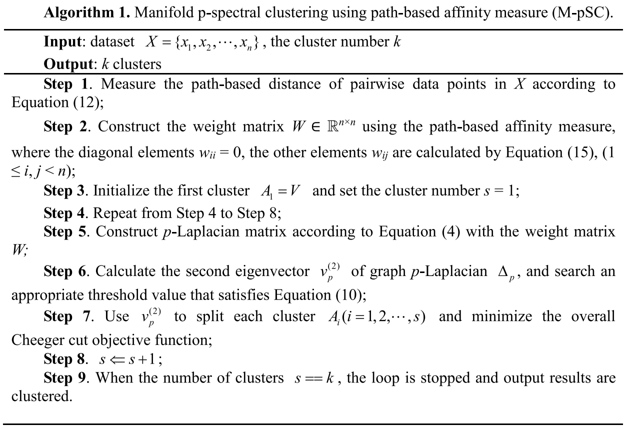
$$w_{ij} = \exp\left(-\frac{D(x_i, x_j)}{2\sigma_i \sigma_j}\right) \tag{10}$$

where the scale parameter $\sigma_i = d(x_i, x_{il}) = \|x_i - x_{il}\|$, x_{il} is the lth neighbor node of x_i. Equation (10) uses the adaptive adjustment method [29] to determine the manifold distance kernel parameter σ. σ_i adaptively changes with the neighborhood distribution to ensure that the affinity within the same cluster increases and the affinity between different clusters decreases. It is not difficult to see that this distance kernel also satisfies the positive definite condition.

3.3 The proposed manifold *p*-SC algorithm

Calculating the similarities between data points and constructing the weight matrix W is a key step in p-spectral clustering. Traditional p-spectral clustering algorithms measures the similarities of data based on the Euclidean distance, but Euclidean distance cannot describe the manifold structure of data sets [30]. Therefore, the clustering performance of p-spectral clustering algorithm are not good on manifold data sets. In addition, real data sets usually contain noise and irrelevant features that interfere with clustering processes and affect the accuracy of clustering results. To solve the above problems, we design a path-based affinity measure for the data sets with manifold structures in Sect. 3.2. This affinity measurement can well describe the spatial distribution of complex data, increasing the similarities between data points on the same manifold, and reducing the similarities of data pairs located on different manifolds.

Then we use the manifold affinity measure to improve p-spectral clustering, and use path-based affinity measurement (M-pSC) to propose M-pSC. The main idea of M-pSC algorithm is: first, compute the path-based distance between data points; measure the manifold similarities of data points using the path-based distance to construct the weight matrix; then calculate the p-Laplacian matrix based on the weight matrix; finally, divide the graph into multiple sub-graphs with the eigenvectors of p-Laplacian matrix. When the Cheeger cut criterion is minimized. we can get high quality clustering results. The detailed steps of the M-pSC algorithm are given in Algorithm 1.



Algorithm 1. Manifold p-spectral clustering using path-based affinity measure (M-pSC).

```
Input: dataset $X = \{x_1, x_2, \cdots, x_n\}$, the cluster number $k$
Output: $k$ clusters
Step 1. Measure the path-based distance of pairwise data points in $X$ according to Equation (12);
Step 2. Construct the weight matrix $W \in \mathbb{R}^{n \times n}$ using the path-based affinity measure, where the diagonal elements $w_{ii} = 0$, the other elements $w_{ij}$ are calculated by Equation (15), ($1 \le i, j < n$);
Step 3. Initialize the first cluster $A_1 = V$ and set the cluster number $s = 1$;
Step 4. Repeat from Step 4 to Step 8;
Step 5. Construct $p$-Laplacian matrix according to Equation (4) with the weight matrix $W$;
Step 6. Calculate the second eigenvector $v_p^{(2)}$ of graph $p$-Laplacian $\Delta_p$, and search an appropriate threshold value that satisfies Equation (10);
Step 7. Use $v_p^{(2)}$ to split each cluster $A_i (i = 1, 2, \cdots, s)$ and minimize the overall Cheeger cut objective function;
Step 8. $s \Leftarrow s + 1$;
Step 9. When the number of clusters $s == k$, the loop is stopped and output results are clustered.
```

4 Theoretical analysis

4.1 Feasibility of the path-based affinity measure

The distance measure given by Definition 2 uses the shortest path on the manifold, which can well reflect the inner manifold structure of the dataset. On the same manifold, we can use many short edges to connect two data points, while two data points from different manifolds need to be connected by a long edge passing through the low density region. Since the manifold distance of the short-edge combinations is small and the manifold distance of the single long edge is usually large, it is possible to reduce the distance of data points on the same manifold and enlarge the space between data points on different manifolds. As can be seen from the above definition, this distance measure is able to describe the local density characteristics of the data.

Consider the two extreme cases of Eq. (9):

1 When $\rho \to 0$, using the equivalent infinitesimal theorem, we get the limit:

$$D(x_i, x_j) = \min_{path \subset \Omega_{i,j}} \sum_{k=1}^{m-1} d(v_k, v_{k+1}) \tag{11}$$

which is the shortest path between nodes based on Euclidean distance. According to the triangular inequality of distance measure, the minimum path is the Euclidean distance between two points, namely $\left\| x_i - x_j \right\|$. Therefore the derived affinity measure cannot reflect the global consistency of the manifold structure. For the sparse connection matrix, this distance measure takes full consideration of the path length, so it can prevent the influence of the noise data on the boundary. But because of the limitations of Euclidean distance, the result distance is not suitable for describing the affinity of manifold data.

2 When $\rho \to \infty$, with the Lobida law, we get the situation:

$$D(x_i, x_j) = \min_{path \in \Omega_{i,j}} \max_{k<m} d(v_k, v_{k+1}) \tag{12}$$

$D(x_i, x_j)$ is called the connection distance, which is the minimum value among the maximum lengths between two adjacent points on all paths. The kernel matrix calculated by Eq. (12) is the connection kernel. It should be noted that this distance measure only considers the maximum distance between two neighbors in the connection path. It takes full account of the influence of local density to ensure the global consistency, but not considers the length of the path. So the affinity may be disturbed by the manifold boundary points.

Therefore, the scaling factor ρ is very important for the distance measure of Definition 2. By controlling the scaling factor ρ, we may consider the manifold structure of the data, and prevent the influence of boundary noise at the same time, thus ensure the global consistency of the manifold distance. Figure 3 shows the Euclidean distance and the path-based distance of ab and bc when $\rho = 2$. The value outside brackets (d_1) is Euclidean distance, and the value inside brackets (d_2) is the path-based distance.

We assume that the shortest path from point b to point c only passes point d. According to the calculation equation of manifold distance in Definition 2, the manifold distance of bc is the sum of two adjacent line segments bd and dc on manifold in the shortest path of bc. It can be seen from

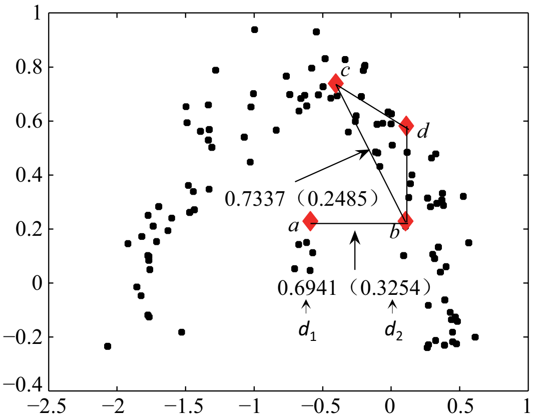


Fig. 3 Euclidean distance and path-based distance

Fig. 3 that the Euclidean distance of *ab* is less than that of *bc* ($0.6941 < 0.7337$). But in the case where the scaling factor is set to 2, the manifold distance of *ab* is larger than that of *bc* ($0.3254 > 0.2485$). This is because point *a* and point *b* are on different manifolds, while point *b* and point *c* are on the same manifold. The manifold distance defined in this paper can increase the distance between two points on different manifolds and reduce the distance between two points on the same manifold. The above results are based on the assumption that there is only one intermediate node *d* in the shortest path from *b* to *c*. From Fig. 4, it is not difficult to find that the shortest path of *bc* may have many nodes. In this case, using the manifold distance method defined in this paper, the distance of *bc* will be far less than the distance of *ab*, and thus increasing the affinity between point *b* and point *c*, reducing the affinity between point *a* and point *b*, and ensuring the global consistency of manifold.

In addition, by adjusting the proportional factor ρ, the manifold distance between two points on different manifolds can be extended. When ρ is set to ∞, it can be seen from Eq. (12) that the distance between two points is the maximum Euclidean distance between any two nodes in the shortest path. Figure 3 presents an intuitive example that the distance of *bc* in high-density region will be far less than the distance of *ab* connected through the low density area. Therefore, the proposed path-based affinity measure can shorten the distance between two points on the manifold, which fully reflects the internal manifold structure of the data set. It effectively solves the problem that the Euclidean distance as the affinity measure cannot embody the global consistency of data.

4.2 Feasibility of the graph *p*-Laplacian

The proposed M-pSC algorithm minimizes Cheeger cut by graph *p*-Laplacian. Theorem 1 analyzes the relationship between Cheeger cut and graph *p*-Laplacian.

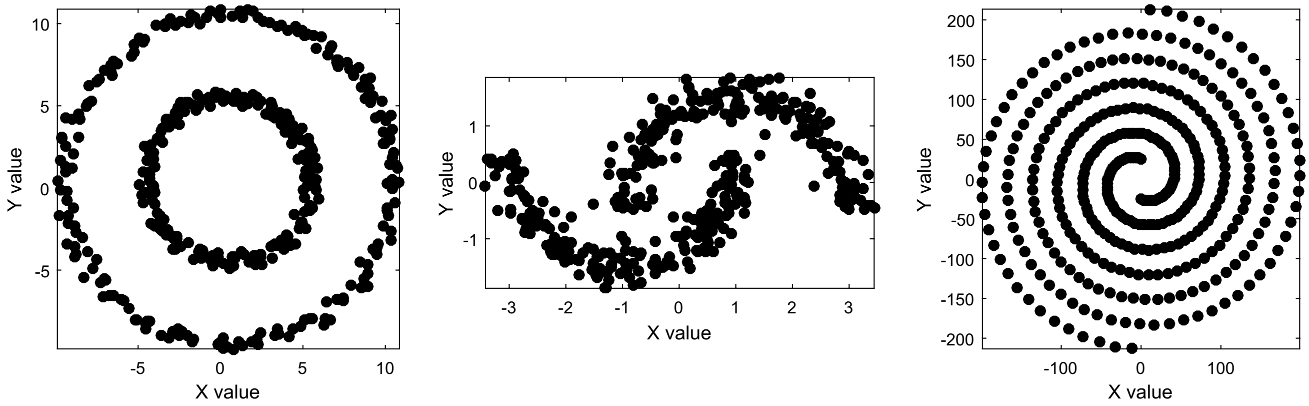
Theorem 1 *Assume the bipartition of V are V_1 and V_2. For $p > 1$, there exists a indicate vector f(p) such that the functional $F_p(f)$ associated to the p-Laplacian satisfies*:

$$F_p(f) = \mathrm{cut}(V_1, V_2)\left|\frac{1}{|V_1|^{\frac{1}{p-1}}} + \frac{1}{|V_2|^{\frac{1}{p-1}}}\right|^{p-1} \tag{13}$$

Equation (13) has a special case that can be viewed as a balanced graph partition criterion:

$$\lim_{p\to 1} F_p(f) = \mathrm{Ccut}(V_1, V_2) \tag{14}$$

Proof First we define a function *f*(*p*) to describe the partition (V_1, V_2) of *V*:



(a) two circles **(b)** two moons **(c)** two spirals

Fig. 4 Original synthetic datasets

$$f(p)_i = \begin{cases} 1 \Big/ \left|V_1\right|^{\frac{1}{p-1}} & i \in V_1 \\ -1 \Big/ \left|V_2\right|^{\frac{1}{p-1}} & i \in V_2 \end{cases}, \quad (p > 1) \tag{15}$$

Bring Eq. (15) into $\langle f, \Delta_p f \rangle$ and $\|f\|_p^p$, we have

$$\langle f, \Delta_p f \rangle = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left|f_i - f_j\right|^p = \sum_{i \in V_1, j \in V_2} w_{ij}\left|\frac{1}{\left|V_1\right|^{\frac{1}{p-1}}} + \frac{1}{\left|V_2\right|^{\frac{1}{p-1}}}\right|^p$$

$$\|f\|_p^p = \sum_{i=1}^{n}\left|f_i\right|^p = \sum_{i \in A}\left|1 \Big/ \left|V_1\right|^{\frac{1}{p-1}}\right|^p + \sum_{i \in \bar{A}}\left|-1 \Big/ \left|V_2\right|^{\frac{1}{p-1}}\right|^p$$
$$= \frac{1}{\left|V_1\right|^{\frac{1}{p-1}}} + \frac{1}{\left|V_2\right|^{\frac{1}{p-1}}}$$

Then, replace the numerator and denominator of Eq. (5) using the above two equations respectively:

$$F_p(f) = \frac{\langle f, \Delta_p f \rangle}{\|f\|_p^p} = \frac{\sum_{i \in V_1, j \in V_2} w_{ij}\left|\frac{1}{\left|V_1\right|^{\frac{1}{p-1}}} + \frac{1}{\left|V_2\right|^{\frac{1}{p-1}}}\right|^p}{\frac{1}{\left|V_1\right|^{\frac{1}{p-1}}} + \frac{1}{\left|V_2\right|^{\frac{1}{p-1}}}}$$
$$= \sum_{i \in V_1, j \in V_2} w_{ij}\left|\frac{1}{\left|V_1\right|^{\frac{1}{p-1}}} + \frac{1}{\left|V_2\right|^{\frac{1}{p-1}}}\right|^{p-1}$$
$$\leq \sum_{i \in V_1, j \in V_2} w_{ij}\left|\frac{2}{\min\{\left|V_1\right|, \left|V_2\right|\}^{\frac{1}{p-1}}}\right|^{p-1} = 2^{p-1}\frac{\mathrm{cut}(V_1, V_2)}{\min\{\left|V_1\right|, \left|V_2\right|\}}$$

Comparing the above inequality and the objective function of Cheeger cut, we can see that if p approaches 1, the special case of $F_p(f)$ is $\mathrm{Ccut}(V_1, V_2)$:

$$\lim_{p \to 1} F_p(f) = \mathrm{Ccut}(V_1, V_2) \qquad \square$$

M-pSC algorithm computes the global minimizer of the second eigenvector $v_p^{(2)}$ of p-Laplacian matrix. Since the functional $F_p(f)$ is continuous in p, if two values p_1 and p_2 are close, the global minimizer of $v_{p_1}^{(2)}$ and $v_{p_2}^{(2)}$ are also close. M-pSC algorithm uses a mixture of gradient and Newton steps to minimize the functional $F_p(f)$, because Newton-like methods have super-linear convergence close to the global optima.

5 Experimental analysis

5.1 Clustering on synthetic datasets

In the experiments, the clustering performances of spectral clustering algorithm (SC) [31], p-spectral clustering algorithm (p-SC) [23] and M-pSC algorithm are compared on three challenging synthetic datasets: "two circles", "two moons" and "two spirals". These datasets are illustrated in Fig. 4. The clusters in these datasets are distributed on manifold data structures and they are not easy to be separated by linear partition method.

The clustering results of SC algorithm, p-SC algorithm and M-pSC algorithm on these three synthetic data sets are presented in Fig. 5. The scale parameter σ of SC algorithm is 0.5; the maximum iteration of p-SC algorithm is maxit = 100; the density factor of M-pSC algorithm is $\rho = 2$ and the number of nearest neighbors is $l = 5$, which is used to calculate the scale factor σ.

Form Fig. 5, we can see that SC algorithm can recognize 'two circles' data set, but it doesn't perform well on 'two moons' and 'two spirals' data sets. p-SC algorithm can generate balanced clusters on 'two moons' data set. But similar to SC algorithm, p-SC algorithm measures the affinity between points based on Euclidean distance and it cannot recognize complex manifold structure of the data set, such as 'two spirals'. In contrast, the performance of the proposed M-pSC algorithm is much better. M-pSC algorithm inherits the advantage of p-SC algorithm that using Cheeger cut criterion to find clusters. With the help of manifold affinity measurement, M-pSC algorithm is applicable to the clustering problem of various datasets. For M-pSC algorithm, the data points on the same manifold have high affinity and the data points on different manifolds are dissimilar with each other. Therefore M-pSC algorithm can find the appropriate exemplars for each data point and assign the corresponding data points to the right clusters.

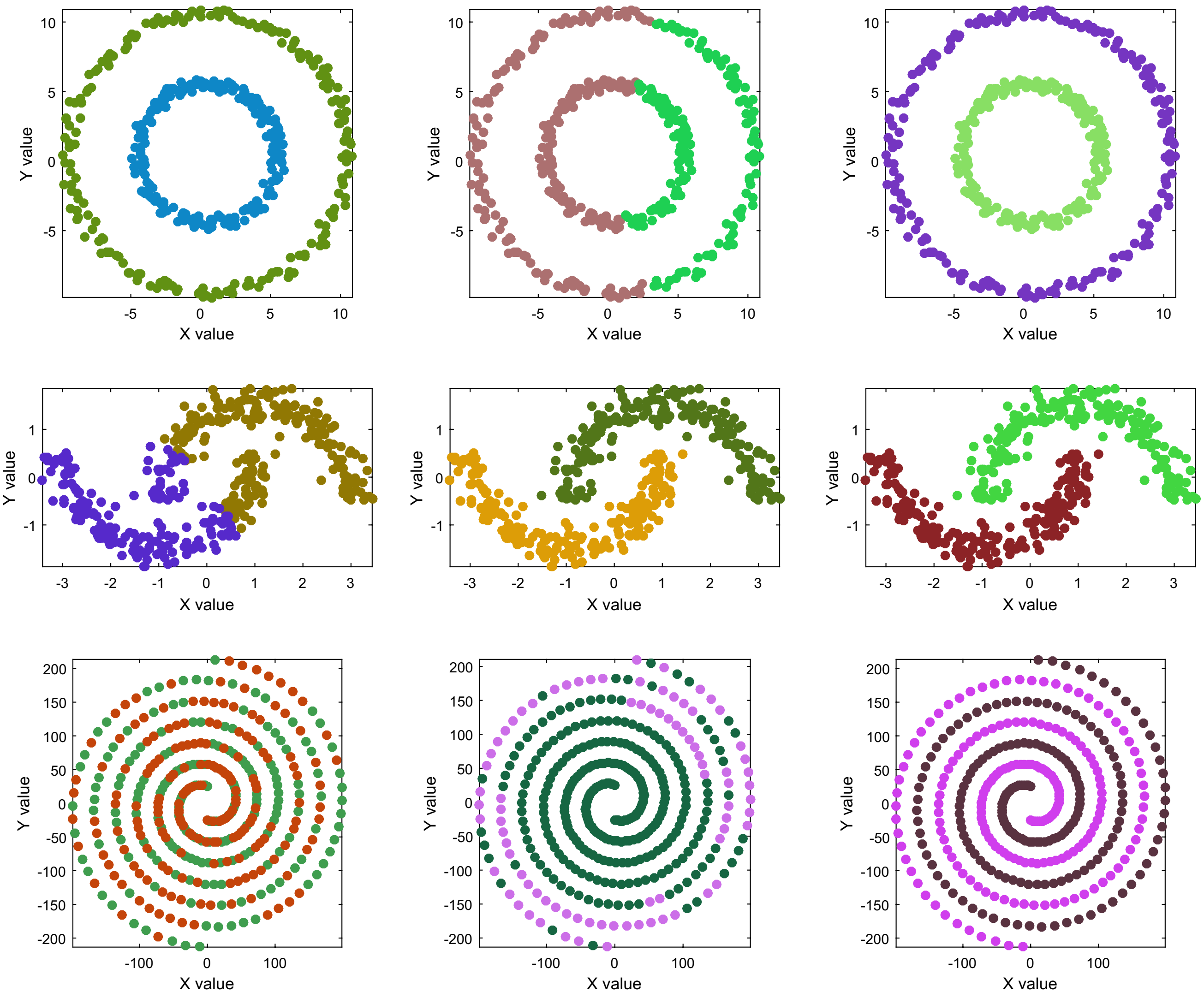
5.2 Clustering on real world datasets

1 Data sets

In order to test the effectiveness of the proposed M-pSC algorithm, six benchmark datasets are used for experiments.The information of these data sets are shown in Table 2. Ionosphere, WDBC, Madelon, Gisette are come from UCI machine learning repository.[1] Colon cancer, leukemia are all cancer data sets, which are available on the LIBSVM data page.[2]

[1] https://archive.ics.uci.edu/ml/.

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.



(a) SC algorithm **(b)** *p*-SC algorithm **(c)** M-pSC algorithm

Fig. 5 Clustering results of different algorithms on synthetic datasets

2 Evaluation metric

There are a lot of methods to measure the merits of clustering results [32]. The F measure and normalized mutual information (NMI) are commonly used evaluation metrics in clustering analysis.

F measure [33] F measure comes from the field of information retrieval. The F value contains the accuracy and recall rate. These two indicators describe the difference between cluster results and actual classes from different perspectives. Assuming there are k classes in the data set, class i is associated with cluster $i*$ in the cluster results. We may compute the F-score of class i using the following three equations:

$$F(i) = \frac{2 \times P(i) \times R(i)}{P(i) + R(i)} \tag{16}$$

where $P(i) = N_{ii*}/N_{i*}$ is the accuracy rate and $R(i) = N_{ii*}/N_i$ is the recovery rate; N_{ii*} is the size of the intersection of class i and cluster $i*$; N_i is the size of class i; N_{i*} is the size of cluster $i*$.

The F-index of the clustering result is the weighted average of the F values for each class:

$$F = \frac{1}{n} \sum_{i=1}^{k} [N_i \times F(i)] \tag{17}$$

where n is the number of sample points; k is the class number of data set; N_i is the size of class i. $F \in [0, 1]$, the larger

Table 2 Data sets used in the experiments

Data set	Instance number	Condition attribute number	Decision attribute number	Class number
Ionosphere	351	34	1	2
WDBC	569	30	1	2
Colon cancer	62	2000	1	2
Leukemia	72	7129	1	2
Madelon	2600	500	1	2
Gisette	7000	5000	1	2

the F index, means that the clustering results of the algorithm closer to the real data category.

NMI [34] NMI is a normalization of the mutual information (MI). Let U_c be the membership matrix of clustering results and U_t be the membership matrix of true data labels. The NMI of U_c and U_t is:

$$\mathrm{NMI}(U_c, U_t) = \frac{\mathrm{I}(U_c, U_t)}{\sqrt{\mathrm{H}(U_c) \cdot \mathrm{H}(U_t)}} \tag{18}$$

where $\mathrm{I}(U_c, U_t)$ is the mutual information, $\mathrm{H}(U_c)$ and $\mathrm{H}(U_t)$ are information entropies used to normalize the mutual information. Usually, NMI is estimated by Eq. (19):

$$\mathrm{NMI}(U_c, U_t) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j}^{c,t} \log\left(\frac{n \cdot n_{i,j}^{c,t}}{n_i^c \cdot n_j^t}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i^c \log \frac{n_i^c}{n}\right)\left(\sum_{j=1}^{k} n_j^t \log \frac{n_j^t}{n}\right)}} \tag{19}$$

where n_i^c is the size of cluster i, n_j^t is the size of class j, and $n_{i,j}^{c,t}$ is the size of the intersection between class j and cluster i. If the clustering results and the true data labels are the same, their NMI is 1; if the data points are grouped randomly, their NMI tends to be 0. A higher NMI corresponds to better clustering results.

3. Clustering results

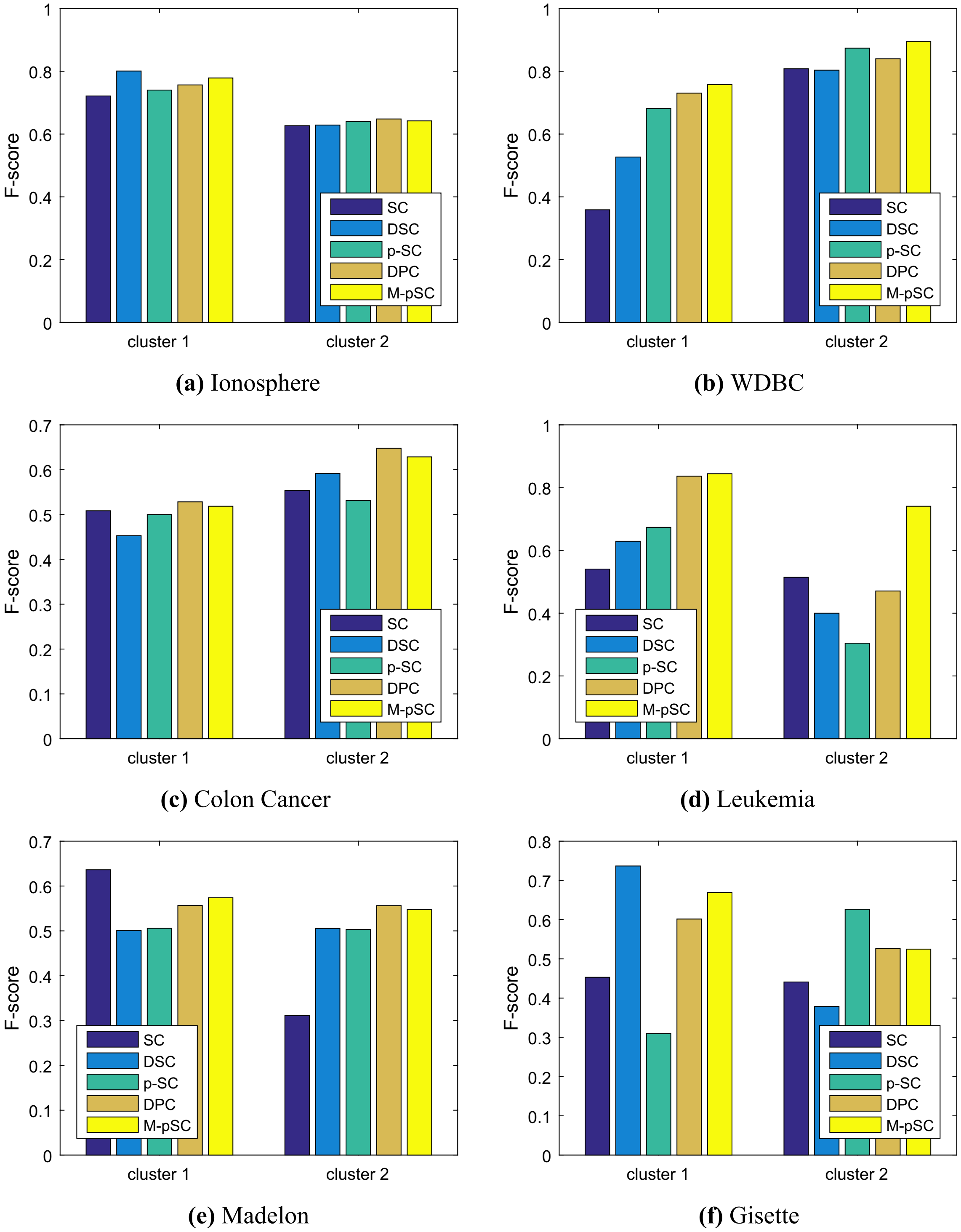
   In the experiments, M-pSC algorithm is compared with the spectral clustering algorithm (SC) [31], density adaptive spectral clustering algorithm (DSC) [35], p-spectral clustering algorithm (p-SC) [23] and the density peaks clustering algorithm (DPC) [36]. All algorithms are implemented by MATLAB, running on a high-performance workstation with 3.20 GHz CPU. The clustering F-score of these five algorithms on each data set are shown in Fig. 6. The horizontal axis of the graph is the cluster label, and the vertical axis is the F-score for each cluster.

From Fig. 6 we can see that the performance of SC algorithm is close to that of p-SC algorithm. This is mainly because that their affinity matrix is based on Euclidean distance. DSC algorithm uses local density adaptive affinity measure to calculate the similarities between data points, so its clustering results are better than SC algorithm on most data sets. p-SC algorithm turns the clustering problem into a graph partitioning problem with the balanced Cheeger cut criterion and it works well on Ionosphere and WDBC data sets. However, for multi-cluster problems, their F-values are lower than the DPC algorithm and the proposed M-pSC algorithm. Because the information in each attribute of the instance is different, and they also have different contributions to the cluster. Inappropriate affinity metrics can have a negative impact on cluster results. Traditional p-spectral clustering algorithms are susceptible to interference from noise and extraneous properties, so it is not suitable to cluster data sets with complex structures. DPC algorithm groups data points according to the local density of data, but the global consistency of data is ignored. In contrast, the proposed M-pSC algorithm is able to handle manifold clustering problems and recognize more complex data structures with the help of path-based affinity measure.

For further comparison, Table 3 lists the F index, NMI index and clustering time of each algorithm. Table 3 shows that M-pSC algorithm can handle the data sets with different structures well. It can generate more accurate clustering results compared with the p-SC and other clustering algorithms. In M-pSC algorithm, the affinities between data points are measured by the path-based affinity function, which can well describe the manifold structure of data. Therefore, the data points within the same cluster are more compact, while the data points between different clusters are more separate. Therefore, in most cases, M-pSC algorithm has higher clustering accuracy. The M-pSC algorithm is applicable to data sets distributed on manifolds. It has good robustness and strong generalization ability. However, the path-based affinity measure also increases the clustering time of M-pSC. How to improve the efficiency of M-pSC algorithm needs further study.

6 Conclusions

When dealing with manifold data, traditional p-spectral clustering does not perform well. In this paper, we design a path-based affinity measure to evaluate the relationship between data points with manifold structures. In this method, we first define the segment length of two points on manifold. Then we use the shortest path between two points to represent their distance. The path-based distance is used to calculate the manifold similarities of pairwise points. Then we use the path-based affinity measure to improve the performance



(a) Ionosphere

(b) WDBC

(c) Colon Cancer

(d) Leukemia

(e) Madelon

(f) Gisette

Fig. 6 Clustering F-score on different datasets

of *p*-spectral clustering on manifold data and propose the M-pSC algorithm.

The advantages of M-pSC are: (1) the affinity matrix constructed by path-based affinity measure in M-pSC can better describe the distribution of manifold data; (2) M-pSC will enlarge the similarities within the manifold and reduce the similarities between manifolds. The drawbacks of M-pSC are: (1) the computation of the shortest path between two nodes will increase the time cost of affinity measure; (2) the

Table 3 Clustering results of different algorithms

Data set	Evaluation index	Algorithm				
		SC	DSC	*p*-SC	DPC	M-pSC
Ionosphere	*F* index	0.6874	**0.7490**	0.7040	0.7177	0.7235
	NMI index	0.1170	**0.1743**	0.1305	0.1312	0.1333
	Time (s)	0.2030	0.2369	2.2776	0.1603	3.2815
WDBC	*F* index	0.6408	0.7260	0.8019	0.8139	**0.8527**
	NMI index	0.3114	0.3569	0.4223	0.4651	**0.5524**
	Time (s)	0.5332	1.4189	8.8077	0.4224	10.0767
Colon cancer	*F* index	0.5377	0.5423	0.5202	**0.6055**	0.5911
	NMI index	0.0089	0.0175	0.0033	**0.0881**	0.0232
	Time (s)	0.0307	0.1738	0.2401	0.0429	0.6455
Leukemia	*F* index	0.5314	0.5496	0.5453	0.7094	**0.8142**
	NMI index	0.0028	0.0055	0.0039	0.0084	**0.0161**
	Time (s)	0.1254	0.8025	0.3406	0.1681	1.3562
Madelon	*F* index	0.4738	0.5030	0.5045	0.5565	**0.5606**
	NMI index	0.0002	0.0003	0.0005	0.0018	**0.0032**
	Time (s)	17.1829	21.8166	15.4910	10.5509	31.1646
Gisette	*F* index	0.4456	0.5074	0.4594	0.5238	**0.5435**
	NMI index	0.3686	0.4266	0.4083	0.4427	**0.4676**
	Time (s)	25.2036	29.6294	23.5818	6.0951	40.1548

The bold value means the best result on each dataset

$n \times n$ affinity matrix will occupy a lot of memory space when the number of data points n is very large.

Experiments on benchmark data sets show that the clustering quality of M-pSC algorithm is superior to that of the original p-spectral clustering algorithm. In the future, we will study how to apply M-pSC algorithm to network data mining, information retrieval, social network analysis, and other scenarios.

Acknowledgements This work is supported by the National Natural Science Foundation of China (no. 61672522, no. 61976216, and no. 61906077), the Natural Science Foundation of Jiangsu Province (no. BK20190838), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (no. 18KJB520009).

References

1. Liu R, Wang H, Yu X (2018) Shared-nearest-neighbor-based clustering by fast search and find of density peaks. Inf Sci 450:200–226
2. Zhang H, Lu J (2010) SCTWC: an online semi-supervised clustering approach to topical web crawlers. Appl Soft Comput 10(2):490–495
3. Du T, Qu S, Liu F et al (2015) An energy efficiency semi-static routing algorithm for WSNs based on HAC clustering method. Inf Fusion 21:18–29
4. Jia H, Ding S, Xu X et al (2014) The latest research progress on spectral clustering. Neural Comput Appl 24(7–8):1477–1486
5. Zhang H, Cao L (2014) A spectral clustering based ensemble pruning approach. Neurocomputing 139:289–297
6. Du M, Ding S, Xu X et al (2018) Density peaks clustering using geodesic distances. Int J Mach Learn Cybern 9(8):1335–1349
7. Cheng D, Nie F, Sun J et al (2017) A weight-adaptive Laplacian embedding for graph-based clustering. Neural Comput 29(7):1902–1918
8. Li Z, Nie F, Chang X et al (2018) Rank-constrained spectral clustering with flexible embedding. IEEE Trans Neural Netw Learn Syst 29(12):6073–6082
9. Bresson X, Szlam AD (2010) Total variation and cheeger cuts. In: Proceedings of the 27th international conference on machine learning, pp 1039–1046
10. Jia H, Ding S, Du M (2015) Self-tuning p-spectral clustering based on shared nearest neighbors. Cogn Comput 7(5):622–632
11. Zhang L, Wei W, Bai C et al (2018) Exploiting clustering manifold structure for hyperspectral imagery super-resolution. IEEE Trans Image Process 27(12):5969–5982
12. Jia H, Wang L, Song H et al (2018) A K-AP clustering algorithm based on manifold similarity measure. In: IIP2018, IFIP AICT, vol 538, pp 20–29
13. Frederix K, Van Barel M (2013) Sparse spectral clustering method based on the incomplete Cholesky decomposition. J Comput Appl Math 237(1):145–161
14. Binkiewicz N, Vogelstein JT, Rohe K (2017) Covariate-assisted spectral clustering. Biometrika 104(2):361–377
15. Ariascastro E, Lerman G, Zhang T (2017) Spectral clustering based on local PCA. J Mach Learn Res 18(9):253–309
16. Law MT, Urtasun R, Zemel RS (2017) Deep spectral clustering learning. In: International conference on machine learning, pp 1985–1994
17. Nie F, Wang X, Huang H (2014) Clustering and projected clustering with adaptive neighbors. In: ACM SIGKDD international conference on knowledge discovery & data mining. ACM, 2014, pp 977–986
18. Tasdemir K, Yalcin B, Yildirim I (2015) Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. Pattern Recognit 48(4):1465–1477

19. Goyal S, Kumar S, Zaveri MA et al (2017) Fuzzy similarity measure based spectral clustering framework for noisy image segmentation. Int J Uncertain Fuzziness Knowl Based Syst 25(04):649–673
20. Wang Y, Jiang Y, Wu Y et al (2011) Spectral clustering on multiple manifolds. IEEE Trans Neural Netw 22(7):1149–1161
21. Langone R, Reynders E, Mehrkanoon S et al (2017) Automated structural health monitoring based on adaptive kernel spectral clustering. Mech Syst Signal Process 90:64–78
22. Zhi W, Qian B, Davidson I (2017) Scalable constrained spectral clustering via the randomized projected power method. IEEE Int Conf Data Min 2017:1201–1206
23. Trillos NG, Slepčev D, Von Brecht J et al (2016) Consistency of cheeger and ratio graph cuts. J Mach Learn Res 17(1):6268–6313
24. Wagner D, Wagner F (1993) Between min cut and graph bisection. In: Proceedings of the 18th international symposium on mathematical foundations of computer science (MFCS), pp 744–750
25. Liu W, Ma X, Zhou Y et al (2019) p-Laplacian regularization for scene recognition. IEEE Trans Cybern 49(8):2927–2940
26. Wang B, Zhang J, Liu Y et al (2017) Density peaks clustering based integrate framework for multi-document summarization. CAAI Trans Intell Technol 2(1):26–30
27. Wang D, Wei Q, Bai X et al (2020) Fractal characteristics of fracture structure and fractal seepage model of coal. J China Univ Min Technol 49(232):103–109 **(+122)**
28. Pu Y, Apel DB, Liu V et al (2019) Machine learning methods for rockburst prediction-state-of-the-art review. Int J Min Sci Technol 29(4):565–570
29. Kang R, Zhang T, Tang H et al (2016) Adaptive region boosting method with biased entropy for path planning in changing environment. CAAI Trans Intell Technol 1(2):179–188
30. Dyke MV, Klemetti T, Wickline J (2020) Geologic data collection and assessment techniques in coal mining for ground control. Int J Min Sci Technol 30(1):131–139
31. Hadjighasem A, Karrasch D, Teramoto H et al (2016) Spectral-clustering approach to Lagrangian vortex detection. Phys Rev E 93(6):063107
32. Shi X, Li Y, Zhao Q (2019) Remote sensing image segmentation combining hierarchical Gaussian mixture model with M-H algorithm. J China Univ Min Technol 48(228):668–675
33. Jeong J, Kim H, Kim S (2018) Reconsideration of F1 score as a performance measure in mass spectrometry-based metabolomics. J Chosun Nat Sci 11(3):161–164
34. Jia H, Ding S, Du M et al (2016) Approximate normalized cuts without Eigen-decomposition. Inf Sci 374:135–150
35. Zhang X, Li J, Yu H (2011) Local density adaptive similarity measurement for spectral clustering. Pattern Recognit Lett 32(2):352–358
36. Du M, Ding S, Jia H (2016) Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowl Based Syst 99:135–145

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.