A new clustering algorithm based on the connected region generation

Liuwei Feng[1,2], Dongxia Chang[1,2] and Yao Zhao[1,2]
[1] Institute of Information Science, Beijing jiaotong University
Beijing 100044 - China
[2] School of Computer and Information Technology, Beijing Jiaotong University
Beijing 100044 - China
[e-mail: dxchang@bjtu.edu.cn]
*Corresponding author: Dongxia Chang

Received July 6, 2017; revised November 7, 2017; revised December 4, 2017; accepted January 24, 2018; published June 30, 2018

Abstract

In this paper, a new clustering algorithm based on the connected region generation (CRG-clustering) is proposed. It is an effective and robust approach to clustering on the basis of the connectivity of the points and their neighbors. In the new algorithm, a connected region generating (CRG) algorithm is developed to obtain the connected regions and an isolated point set. Each connected region corresponds to a homogeneous cluster and this ensures the separability of an arbitrary data set theoretically. Then, a region expansion strategy and a consensus criterion are used to deal with the points in the isolated point set. Experimental results on the synthetic datasets and the real world datasets show that the proposed algorithm has high performance and is insensitive to noise.

Keywords: connected region, nearest neighbors, importance index, seed point, region expansion

This work was supported in part by National Natural Science Foundation of China (No. 61532005, 61572068).

http://doi.org/10.3837/tiis.2018.06.010 **ISSN : 1976-7277**

1. Introduction

Clustering, defined as an unsupervised learning task, aims at decomposing data points into different groups such that the data points within the same group are similar to each other, while the data points in different groups are dissimilar [1]. Clustering is a fundamental tool in unsupervised learning and has been widely applied in various scientific and engineering disciplines [2-5], such as pattern recognition, artificial intelligence, web mining, image segmentation and biology [6].

Many clustering algorithms have been proposed over the years. Generally, these algorithms include partitioning methods [7], hierarchical method [8], density-based methods [9], spectral clustering method [10] , affinity propagation method [11] and other clustering algorithm based on the k nearest neighbors [12]. The partitioning method aims at classifying the data points into a number of clusters that are usually optimal in terms of some predefined criterion functions. Among the partitioning mtheod, the K-means algorithm [7] is the most representative and widely usded algorithm. Although the K-means algorithm is fast and simple, it still has some defects. One of the defects is it can only find spherical clusters and performs worse for the non-spherical data. In order to improve the performance of the K-mean algorithm, many improved algorithms been proposed [13-14]. For the hierarchical method [8,15,16], it mainly includes the agglomerative and divisive methods. The Single-Link [8] is the most representative algorithm. Whether the hierarchical algorithm is agglomerative or divisive method, a cut-off condition needs be provided artificially.

Generally, a good clustering algorithm should be able to discover clusters with different shapes and different distributions. In fact, the density-based and the spectral algorithms are easily to detect an arbitrary shape. Among the density-based methods, DBSCAN [9] is one of the more widely used algorithm which can discover clusters of arbitrary distribution automatically. However, it is difficult to distinguish the small clusters close to a large clusters. In order to improve the performance of DBSCAN, several algorithms have been proposed [17-19]. In fact, these algorithms still get bad results when there are no obvious boundaries between two classes in the dataset. These algorihtms will recognize the two classes without obvious boundaries as one class. For the spectral clustering algorithm, several spectral clustering algorithms [10,20-28] which can deal with both spherical and non-spherical data have been proposed. Ng-Jordan-Weiss (NJW) algorithm [20] is a popular spectral clustering algorithm. NJW algorithm uses the k largest eigenvectors of the affinity matrix to project the original dataset into a new space and then cluster the mapped data by using the K-means clustering algorithm. In fact, the affinity matrix is crucial to the performance of the algorithm. It usually obtains by using the Gaussian kernel function and there is a parameter σ need to be specified manually. Therefore, some algorithms have been proposed to build a better affinity matrix. Zenlnik-Manor et al proposed a self-tuning spectral clustering [21] which computes the affinity between pair of points by using the local scale. The local scaling leads to a better clustering especially when the data includes multiple scales. In order to improve the robust of the algorithm, a

robust path-based spectral clustering algorithm [22] is proposed by defining a robust path-based similarity measure. Zhang et al proposed a spectral clustering with local density adaptive measure [23] which is called common-near-neighbor (CNN). The new similarity defined in the CNN is adaptive assign the scale factor by using the local density. Although the spectral algorithm and its improved algorithm can deal with arbitrary shape data, there are some detects in these algorithms. One of the detects is that these algorithms are not robust by using the K-means algorithm to cluster the mapped data in the last step. Moreover, these algorithms are sensitive to the noise.The affinity propagation (AP) proposed by Frey [11] views each data point as a node in a network. In the network, a real-valued message is defined and recursively transmited along edges of the network until a good set of exemplars and corresponding clusters emerges. The AP algorihtm is good at dealing with the lage scale data, but it will fail to clustering some complex data sets, such as the inseparable data set. Therefore, some improved algorithms [29-34] are proposed. In [32], a semi-supervised AP algorihtm was proposed. This algorithm guides the affinity propagation process by using the prior information. Cyril proposed the hierarchical affinity propagation (HAP) which combining BP neural network with AP, but the HAP algorihtm leads to the decline of algorithm precision[33].

Recently, the clustering approach developed by Rodriguez and Laio [35], termed Rodriguez-Laio clustering (RLClu) in this paper. The RLClu algorithm is based on the idea that the cluster centers usually have higher density and relatively large distance to other cluster centers. However, there are three parameters should be provided by the user and the RLClu algorithm is sensitive to these parameters. In order to reduce the number of the setting parameters, Wang and Song design an improved algorithm called STClu which can detect the clustering centers automatically via statistical testing [36]. STClu algorithm defines a new local density by using the k nearest neighbors which makes the STClu algorithm be more robust to the preassigned parameter k. And STClu algorithm also adds a new definition $\hat{\gamma}$ which is the product of the local density and the minimum density-based distance of every point, and it is used to choose the cluster center. Then, STClu algorithm chooses the final cluster centers via an outward statistical testing method. Although the STClu algorithm only need one preassigned parameter and is less than the RLClu algorithm, the STClu algorithm only can deal with the data sets with specific distributions. The performance of the STClu algorithm will get bad if the data set is not belong to the specific distributuions.

In order to improve the performances of the existing algorithms, a new clustering algorithm based on the connected region generation (CRG-clustering) is proposed. Within the CRG-clustering algorithm, a connected region generating (CRG) algorithm is developed according to the connectivity between the points with their neighbors. The CRG algorithm can be used to obtain the K connected regions and an isolated point set. Then, the region expansion strategy is used to cluster the data points belonging to the isolated point set. Moreover, a consensus criterion is used to deal with the isolated points.

The rest of this paper is organized as follow. Section 2 provides some definitions necessary for our approach and the connected region generating algorithm. Section 3 describes details

of our CRG-clustering algorithm. The performance of proposed algorithm is vealuated and compared with related work in Section 4. Finally, the paper is concluded in Section5.

2. Related work

In this section, we first briefly review the k-NN Based Clustering Algorihtm (KNBC) which is proposed based on the assumption that if $\boldsymbol{x}_i$ is one of the nearest neighbors of $\boldsymbol{x}_j$, then $\boldsymbol{x}_j$ is one of the nearest neighbors of $\boldsymbol{x}_i$ [12]. Let $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ be a finite subset of a N-dimensional vector space, $S_1,S_2,\cdots,S_m$ be the clustering result. Define $N_i=\{x\,|\,x \text{ is one of the nearest neighbors of } x_i\}$, S_j is the j-th cluster subsets, $N_{S_m}=\{x\,|\,x \text{ is one of the nearest neighbors of the points in } S_m\}$. The KNBC algorithm is presented in the following:

(1) Set $m=1$.

(2) Add $x_1=\{x\,|\,x \text{ is the first element in } \boldsymbol{X}\}$ into S_m, namely $S_m=\{x_1\}$.

(3) Get N_{S_m}, the neighbors subset of S_m, add all the points in N_{S_m} into S_m, and let $\boldsymbol{X}=\boldsymbol{X}-N_{S_m}$.

(4) Repeat step (3), until the number of S_m is not change.

(5) If $\boldsymbol{X}=\varnothing$ then stop,else set $m=m+1$ and goto step (2).

The KNBC algorithm is based on the nearest neighbors , so it shows very good perfomance in many area. However, the performance of the KNBC algorithm will be influenced by the initialization and termination criterion. The KNBC selects the first point in the dataset as the initial cluster and adds all the nearest neighbors of the first point. Then all nearest neighbors of these data points are founded and added. In Ref. [12], the KNBC algorihtm is used as the pretreatment in the process of the film editing. The points in the film dataset is sequential, therefore the KNBC algorihtm starts from the first element in the dataset is reasonale. But in the real word, most dataset is non-sequential, the performance of the KNBC algorihtm is effected by the seletion of the first sample. When the number of points in a cluster stops increasing, the cluster generation is terminated. This termination criterion is good at clustering the separable dataset, but it will become worse when the edge of the clusters are overlapping. When the clusters whose edge is overlapped or close to each other, the KNBC algorihtm will divide the two overlapped cluster into one cluster. Moreover, the KNBC algorihtm does not consider the circumstances that there are some noise in the dataset. If there are some noise points in the dataset, the cluster number of the KNBC algorihtm will increase and there will some clusters whose members are several noise points. In order to improve the performance of the KNBC algorihtm, a connected region generating (CRG) algorithm is proposed. In the CRG algorithm, the seed point is defined and selected as the initial point of the cluster. Moreover, a new termination criterion is proposed.

3. The connected region generating algorithm

In this section, a connected region generating (CRG) algorithm is proposed. First, some definitions used in the CRG algorithm are given. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a finite subset of a N-dimensional vector space, K be the specified number of clusters, and C_i be the i-th cluster where $i = 1, \ldots, K$.

3.1 Preliminary

Definition 1 (k-th adjacency matrix). Given a data set $\boldsymbol{X}$, let k be the number of the nearest neighbors. Then the k-th adjacency matrix can be defined as:

$$A_{kNN}(i,j) = \begin{cases} 1 & if\ \boldsymbol{x}_i \in \mathrm{N}_k(\boldsymbol{x}_j) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathrm{N}_k(\boldsymbol{x}_j)$ is the k nearest neighbors set of $\boldsymbol{x}_j$, and $i, j = 1, \ldots, n$.

Difinition 2 (Importance Index). The importance index of a data point $\boldsymbol{x}_i$ is defined as :

$$I(\boldsymbol{x}_i) = \sum_{j=1}^{n} A_{kNN}(i,j) \tag{2}$$

Definition 3 (Seed point). Point $\boldsymbol{x}_i$ is a seed point if its importance index satisfies

$$I(\boldsymbol{x}_i) = \max_{\boldsymbol{x}_j \in \boldsymbol{X}} \{I(\boldsymbol{x}_j)\} \tag{3}$$

This means that $\boldsymbol{x}_i$ is adjacent to many points. In fact, these data points look more like the centroids of the data set.

Definition 4 (Connected region) Let $\boldsymbol{x}$ be a seed point in $\boldsymbol{X}$, then a connected region generated by $\boldsymbol{x}$ is defined as:

$$R(\boldsymbol{x}) = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{m+1} \,|\, \boldsymbol{x}_0 \in \mathrm{N}_k(\boldsymbol{x}), \boldsymbol{x}_i \in \mathrm{N}_k(\boldsymbol{x}_{i-1}), i \in \{1, 2, \cdots, m\}\} \tag{4}$$

Therefore, connected region can be built between the connected points using Definition 4. All the points in the same connected region belong to the same cluster.

3.2 The connected region generating algorithm

In the following, the connected region generating (CRG) algorithm is proposed. The CRG algorithm is designed according to the connectivity between the points with their k nearest neighbors. The skeleton of the CRG algorithm is presented in **Table 1**.

Table 1. The connected region generating algorithm

Input: data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, the number of clusters K, the number of nearest neighbors k, threshold $T_{number} = n/K$

Output: K connected region and an isolated points set

Compute the adjacency matrix A_{kNN} and the importance index according to Definition 1 and Definition 2

$N_{cr} = 0$ (the number of the local connected region)

$S_{cr} = \varnothing$ (the connected region set)

$S_d = \varnothing$ (the points set waiting to generate in region generating process)

$S_U = \varnothing$ (isolated points set)

For $i = 1$ to K

Find the seed point x_s according to Definition3

$N_{cr} = N_{cr} + 1$, $S_{cr}(N_{cr}) = \{\boldsymbol{x}_s\}$, $I(x_s) = 0$

$S_d = \{x_s\}$

While $S_d \neq \varnothing$

Set x_{wait} with the first element of the S_d waiting to generate set

Set $\boldsymbol{X}_{neighbor} = \{x_{nei}^1, x_{nei}^2, \ldots x_{nei}^k\}$ to save the neighbors of the point waiting to generate

$\boldsymbol{X}_{neighbor} = \mathrm{N}_k(x_{wait})$

For $j = 1$ to k

if x_{nei}^j is not classified

Insert x_{nei}^j into the N_{cr} connected region set and the waiting to generate set, and set the impartance index of x_{nei}^j with 0

$S_{cr}(N_{cr}) = S_{cr}(N_{cr}) \bigcup \{\boldsymbol{x}_{nei}^j\}$, $S_d = S_d \bigcup \{x_{nei}^j\}$, $I(x_{nei}^j) = 0$

End

End

Delete the point having generated from the waiting to generate set

$S_d = S_d - \{x_{wait}\}$

If ($|S_{cr}(N_{cr})| > T_{number}$) or (there is no unlabeled data point)

Break

End

End

End

For $i = 1$ to n

If the $i-$th point is not belong to any connected region,

$S_U = S_U \bigcup \{\boldsymbol{x}_i\}$

End

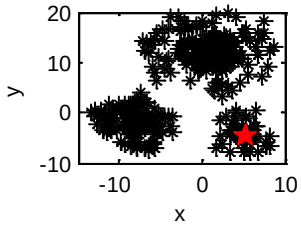
End

The connected region generating algorithm is a region generating algorithm which is based on the k nearest neighbors. This is similar to the KNBC algorithm. However, there are some difference between our CRG and the KNBC algorihtm. First, the CRG algorihtm chooses the seed point according to the importance index as the initial point in every connected region instead of choose the first point. Second, a new termination condition is added in our proposed algorithm. In order to avoid the over-growth of the connected region, the upper bound of the number in each region is set and when the numbers reach this bound the region generating process will stop. Third, the KNBC algorihtm separates all the points based on the nearest neighbors. While some isolated points can be left in our region generation process and this makes our algorithm is insensitive to the noise.

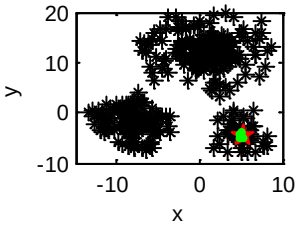
After the identification of the connected region, the data points belonging to the same region can be defined as a subset $S_{cr} \neq \varnothing$ and all the isolated points (i.e. not belonging to any connected region) form the subset S_U. Then, the data set $\boldsymbol{X}$ is partitioned into a number K of connected region sets $S_{cr}(1), S_{cr}(2), \cdots, S_{cr}(K)$, and an isolated points set S_U

$$\boldsymbol{X} = \left(\bigcup_{i \in \{1,2,\cdots,K\}} S_{cr}(i) \right) \bigcup S_U$$

Here, we use the example shown in **Fig. 1** to see the procedure of the CRG. As shown in **Fig. 1 (a)**, the point which has the highest importance index was chosen as the first seed point and marked with star. Then, the connected region will be generated according to Definition 4. Firstly, the k nearest neighbors of the first seed point are found and are added into the connected region set, as shown **Fig. 1(b)**. Then, the nearest neighbors found procedure are repeated until the terminal condition is satisfied. Therefore, the first connected region is generated and is shown in **Fig. 1(c)**. This process was repeated until all the K connected regions are generated and the results are shown in **Fig. 1(d)**. Here, the seed points are marked with stars and the different connected region are marked with different shapes. Moreover, the points marked with the black "*" mean that they are not classified into any connected region.



(a)



(b)

(c) (d)

Fig. 1. The CRG algorithm. (a) The first seed point; (b) The k nearest neighbors of the first seed point;(c) The first connected region; (d) All the K connected region.

4. The clustering algorithm based on CRG

The CRG algorithm can classify the data points according to the neighborhood relationship and it will stop when the terminal condition is satisfied. However, the CRG algorithm can't ensure to deal with all the data points. In fact, as shown in **Fig. 1 (d)**, there always left some points which not belonging to any connected region. In order to solve this problem, a region expansion strategy and a consensus criterion are proposed to deal with these unconnected points. Then, a clustering algorithm based on the CRG (CRG-clustering) is given.

4.1 Region expansion strategy

For a data set, we can get K connected region sets and an isolated point set using the CRG algorithm in the former section. In order to get a better clustering result, the points belonging to the isolated set need to be processed. Here, a region expansion strategy is proposed and the skeleton of the region expansion strategy is given in **Table 2**.

Definition 5 (Expansion radius) Expansion radius of a connected region is calculated according to the distance between the points in this connected region and their k-th nearest neighbor. For the i-th connected region, its expansion radius, $rad(i)$, is defined as:

$$rad(i) = \frac{\sum_{x_j \in S_{cr}(i)} distance(x_j, x_j^k)}{\left|S_{cr}(i)\right|} * 2 \tag{5}$$

where x_j^k is the k-th nearest neighbor of point x_j, $S_{cr}(i)$ is the i-th connected region, and $\left|S_{cr}(i)\right|$ is the number of the points in the i-th connected region.

Table 2. The region expansion strategy

Input: K connected region $S_{cr}(i)$, $i=1,2,\cdots,K$ and an isolated point set S_U
For $i=1$ to K
Calculate the expansion radius $rad(i)$ of the i-th connected region $S_{cr}(i)$ according to Definition 5.
$j=1$
While $j \le \left\|S_{cr}(N_{cr})\right\|$
Find all the points who are in the circle-domain of $x_j \in S_{cr}(i)$ with the radius $rad(i)$, if there are some points belonging to S_U, then remove these points from S_U to $S_{cr}(i)$
End
End

4.2 Consensus criterion

For a data set, if there always left some points that not belonging to any connected region after the connected region generation and the region expansion. Then a consensus criterion will be used to deal with these data points.

Definition 6 (Nearest voting score) $\boldsymbol{V}_i$ is denoted as the nearest voting score matrix of the i-th point, $\boldsymbol{V}_i$ is a K-dimensional vector and every element represent the score result of the k-th nearest neighbor, is defined as:

$$\boldsymbol{V}_i(j)=\left|\left\{x_m \mid x_m \in \mathrm{N}_k(x_i), x_m \in S_{cr}(j)\right\}\right| \tag{6}$$

where $\mathrm{N}_k(\boldsymbol{x}_i)$ is the k nearest neighbors set of $\boldsymbol{x}_i$, and $S_{cr}(j)$ is the j-th connected region, and $\left|\cdot\right|$ represents a function to get the number of the elements in the set.

Definition 7 (Victor region) $R(x_i)$ is denoted as the victor region by using the nearest voting, is defined as:

$$R\left(x_i\right)=\begin{cases}\left\{r \mid \boldsymbol{V}_i(r)=\max\limits_{j=1,2,\ldots K}(\boldsymbol{V}_i(j)), r=1,2,\ldots K\right\} & if \ \max\limits_{j=1,2,\ldots K}(\boldsymbol{V}_i(j))>0 \\ 0 & \text{otherwise}\end{cases} \tag{7}$$

In Definition7, $\max\limits_{j=1,2,\ldots K}(\boldsymbol{V}_i(j))=0$ represents that the k nearest neighbors of $\boldsymbol{x}_i$ all unvisited.

The consensus criterion is presented in the following:

(1) Find the points still in S_U.

(2) Calculate the nearest voting score of every point in S_U according to Definition 6.

(3) Divided the sample into the victor region which is got by using the Difiniton7.

(4) If its k nearest neighbors all unvisited, we need to divide the unvisited point into the region which its nearest labeled neighbor belongs to.

4.3 The clustering algorithm based on CRG

In the new clustering algorithm, the CRG algorithm is used to obtain the K connected regions, firstly. Then the region expansion strategy is used to deal with the data points which are belonging to the isolated data set. Moreover, a consensus criterion is used for the remaining data points. Finally, the clusters of the data set are acquired. The flow chart of the CRG-clustering algorithm is given in **Fig. 2**.

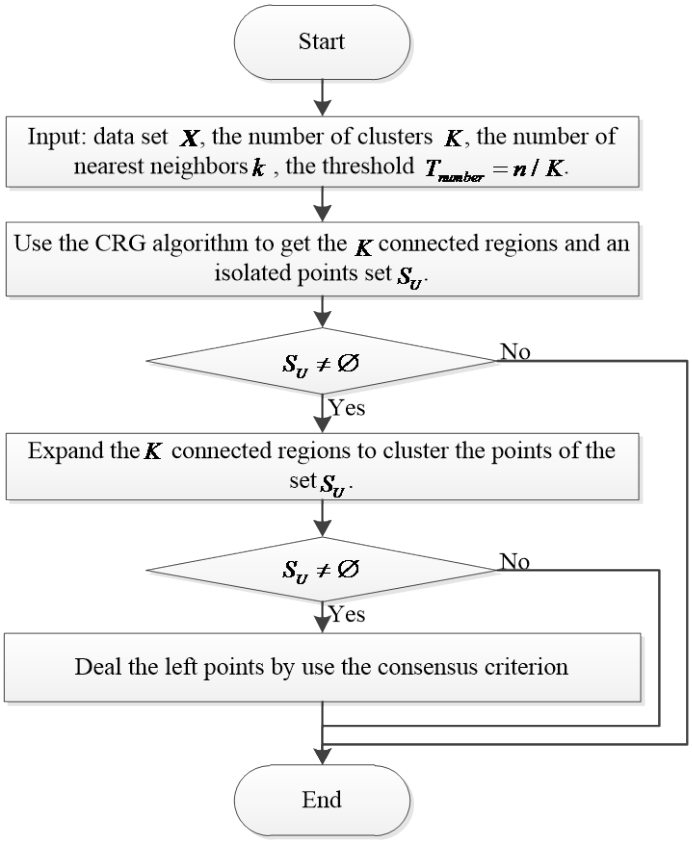


Fig. 2. The flow chart of the CRG-clustering algorithm

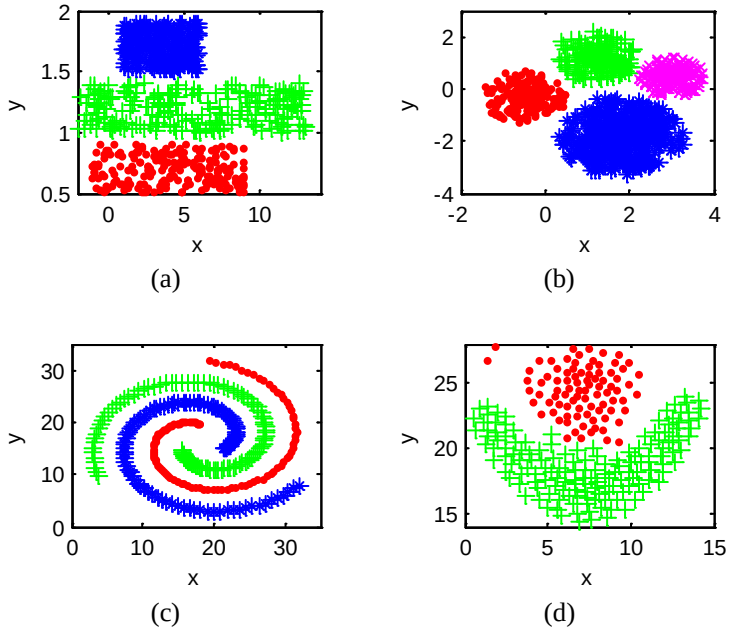
The CRG-clustering algorithm mainly includes three steps. The first step is to obtain the K connected regions by using the CRG algorithm. The CRG algorithm will find out the seed point whose important index is the highest and generate a region from the seed point according to the connectivity of the point and its k nearest neighbors. If there are points not belonging to any region left, then the second step will be used. The second step will deal with the points classify into the isolated data set in the first step, and divide the points into the corresponding connected region by using the region expansion strategy. In certain condition, there are some points left after the second part. So, the third step is used for the remaining data points to ensure all points being clustered. The consensus criterion of the k nearest neighbors will ensure every point belongs to a connected region.

5. Experiments

In order to validate the performance of the CRG-clustering clustering algorithm, a set of experiments are conducted on both artificial data sets and real-life data from the UCI Machine Learning Repository. The performances of the CRG-clustering clustering, K-means [7], DBSCAN [9], RLClu [35], NJW [20], and SC-DA [23] algorithm are compared through the experiments. The results show that the CRG-clustering clustering algorithm has high performance and flexibility.

5.1 Data sets

Several data sets, both artificial and real, have been utilized in our experiments. The eight artificial datasets are referred to as D1~D8, respectively. These artificial data sets, as shown in **Fig. 3**, have different degrees of difficulty for clustering. The first three data sets to be relatively simple: the clusters are separable. Clusters in D4, D5 and D6, have rather different shapes and a few of them are overlapped. For D7 and D8, there are some noise data points, scattered between the clusters, which have been added to increase further the difficulty of the problem to be solved. The real-life data sets considered are UCI data sets.



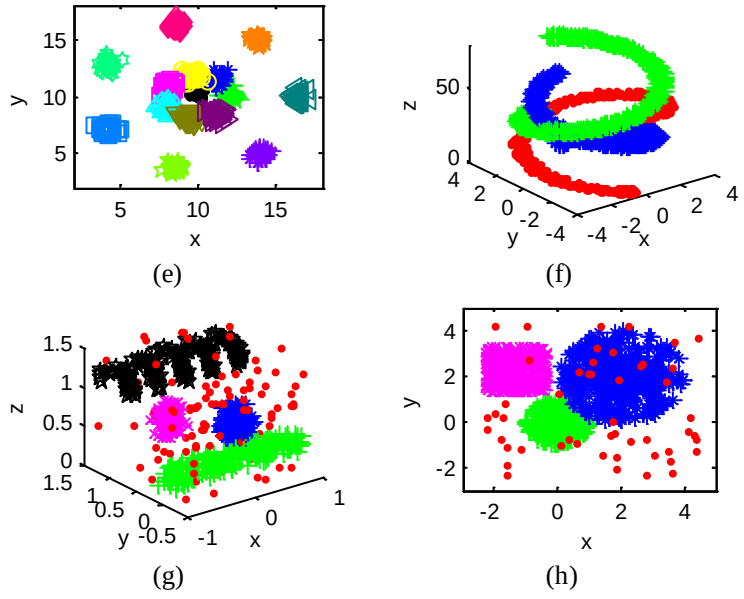


Fig. 3. The artificial data sets used in the experiments: (a)D1, (b)D2, (c)D3, (d)D4, (e)D5, (f)D6,(g)D7, and (h)D8

For convenience, we summarize the fifteen sets in **Table 3** with the characteristics of the data sets. The four columns show the number of data points N, the number of classes k, the dimension of the feature space d, and the number of points in every cluster for each data set.

Table 3. Fifteen data sets used in the experiments

Dataset	N	k	d	points per cluster
D1	600	3	2	200, 200, 200
D2	1000	4	2	200, 200, 200, 400
D3	312	3	2	101,105,106
D4	240	2	2	87,153
D5	600	15	2	Each class owe 40 objects
D6	1200	3	3	400,400,400
D7	900	4	3	200,200,200,200,100(noise)
D8	1250	3	2	400,400,400,50(noise)
Iris (IR)	150	3	4	50, 50, 50
Vehicle (VE)	846	4	18	199, 212, 217, 218
Soybean_small (SO)	47	4	35	10, 10, 10, 17
Seeds (SE)	210	3	7	70, 70, 70
Chess (CH)	2130	2	36	1102, 1028
Liverdisorder (LI)	345	2	6	145, 200
Pima-indians-diabetes(PI)	768	2	8	268, 500

5.2 Results

In the experiments, we have executed all algorithms 20 times independently with random initialization on each data set. Firstly, the clustering results of the artificial data sets obtained by K-means, DBSCAN, RLClu, NJW, SC-DA and CRG-clustering clustering are given in **Fig. 4-11**.

From the results of **Fig. 4** to **Fig. 11**, it may be realized that the CRG-clustering algorithm performs superior to K-means, DBSCAN, RLClu, NJW and SC-DA algorithms. On 6 out of 8 data sets, CRG-clustering performs better than the other algorithms and on the remaining data sets CRG-clustering exhibits a better performance. For the first three separable data sets which include D1, D2 and D3, the CRG-clustering algorithm and the DBSCAN algorithm play the best performance and can divided the three data sets into the correct categories. But the other three algorithms fail in doing so. The RLClu algorithm can divide the points of D2 and D3 into the correct categories, but plays a worst performance in D1. The K-means, NJW and SC-DA can't divide all the samples of three separable data sets into the correct categories. For D4 and D6, NJW, SC-DA and CRG-clustering succeed in providing the all the clusters, while the other three algorithms all have worst results. For D5, only RLClu and CRG-clustering get a better performance. For D7 and D8, there are very noisy backgrounds added to these data sets. For D7, the CRG-clustering algorithm and the DBSCAN algorithm both can divide the non-noisy points into the correct categories. While the K-means, RLClu, NJW and SC-DA algorithms all play a worse performance because of the influence by the noise points. For D8, only DBSCAN algorithm divides the two inseparable spherical clustering into one class due to the connectivity of the boundary samples. And other algorithms are able to find all the clusters of the data set.

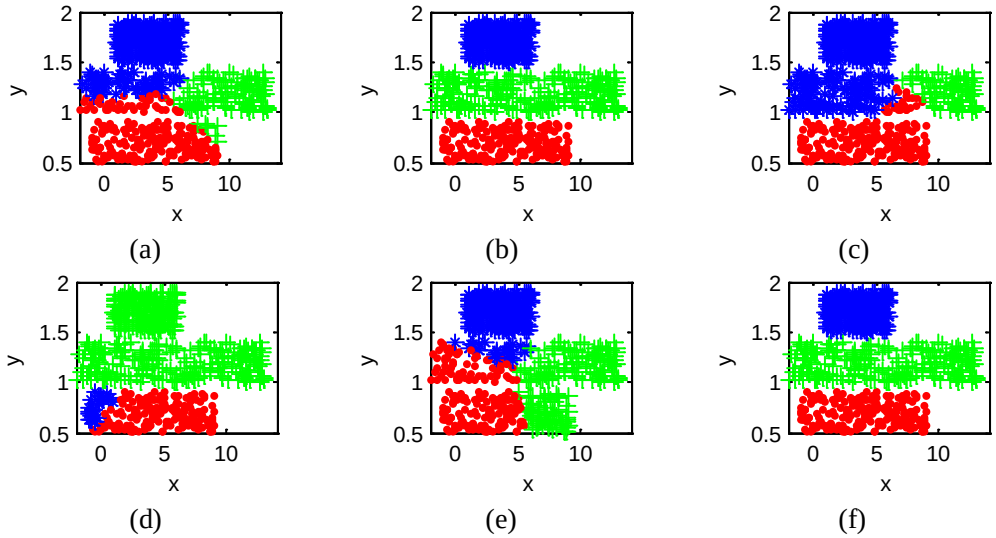


Fig. 4. The clustering results of D1 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

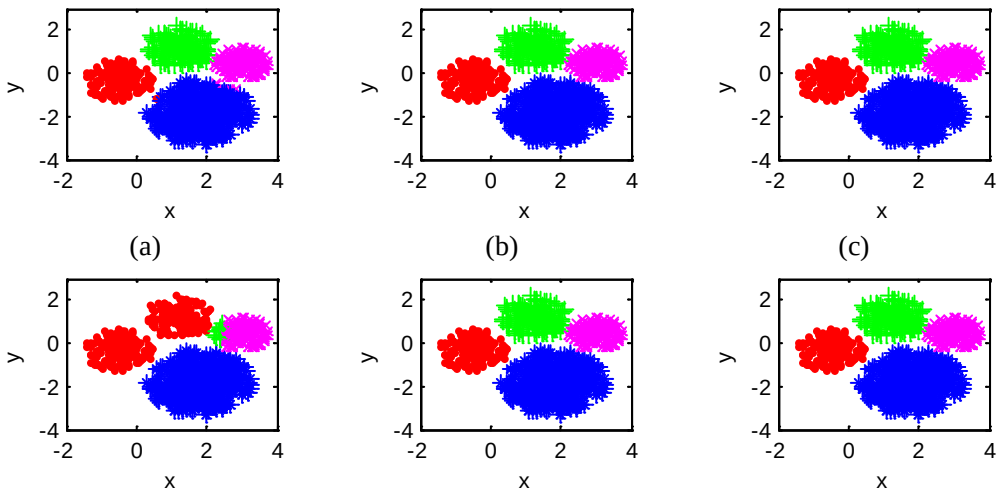


Fig. 5. The clustering results of D2 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

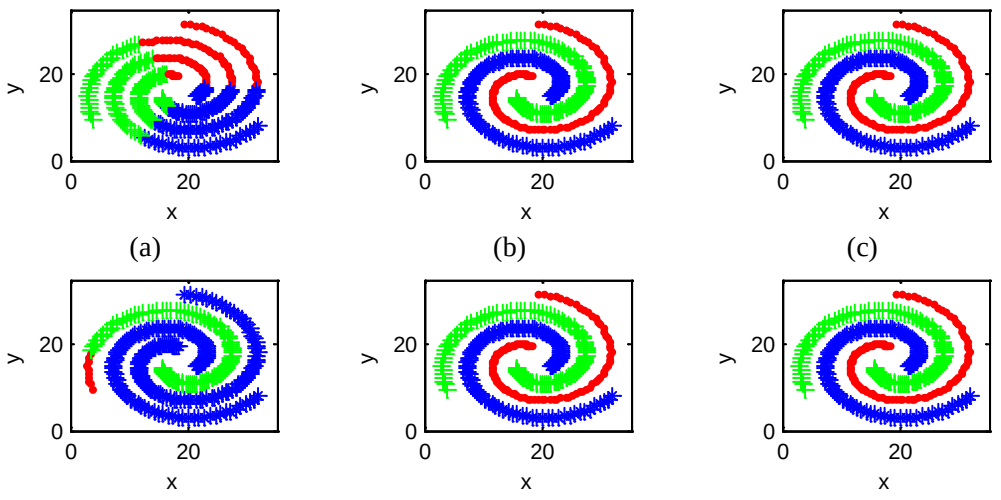


Fig. 6. The clustering results of D3 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

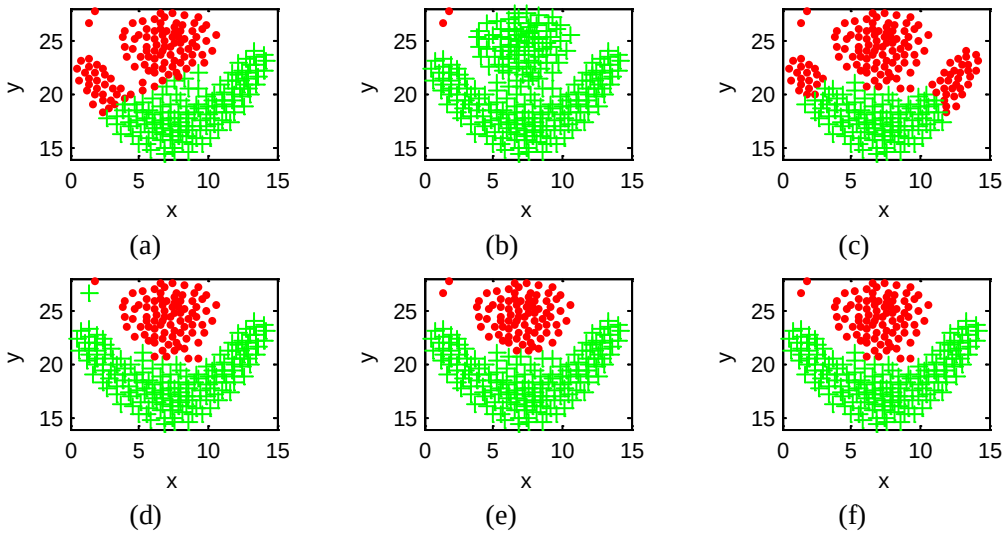


Fig. 7. The clustering results of D4 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

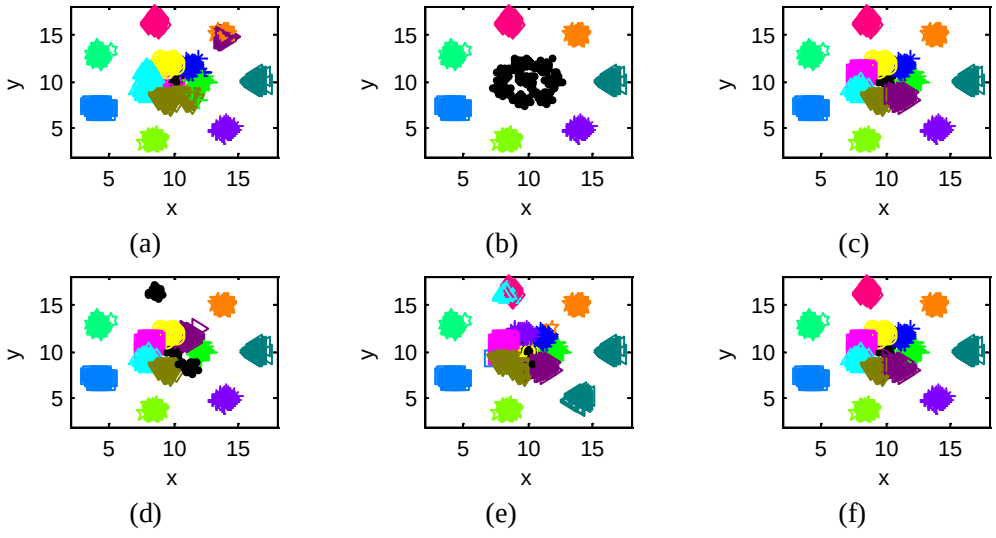


Fig. 8. The clustering results of D5 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

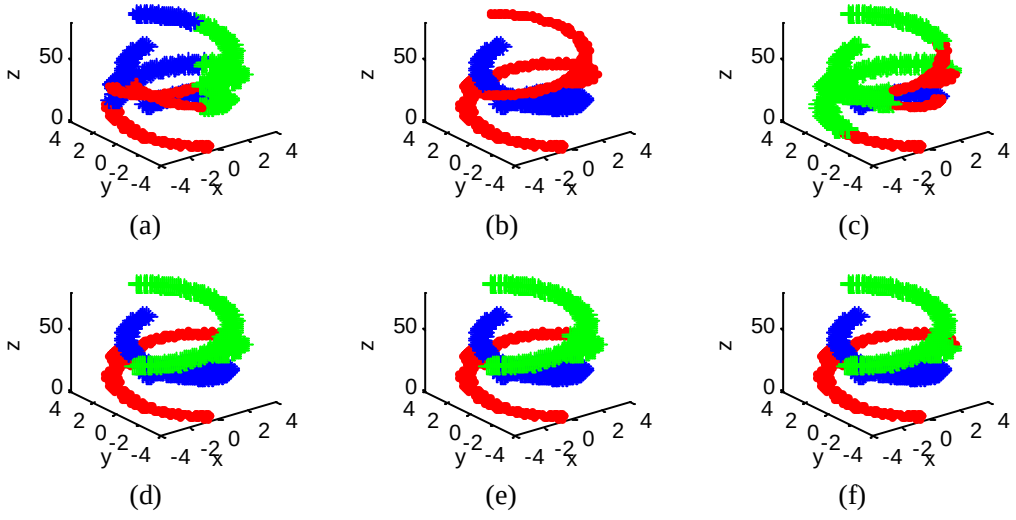


Fig. 9. The clustering results of D6 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

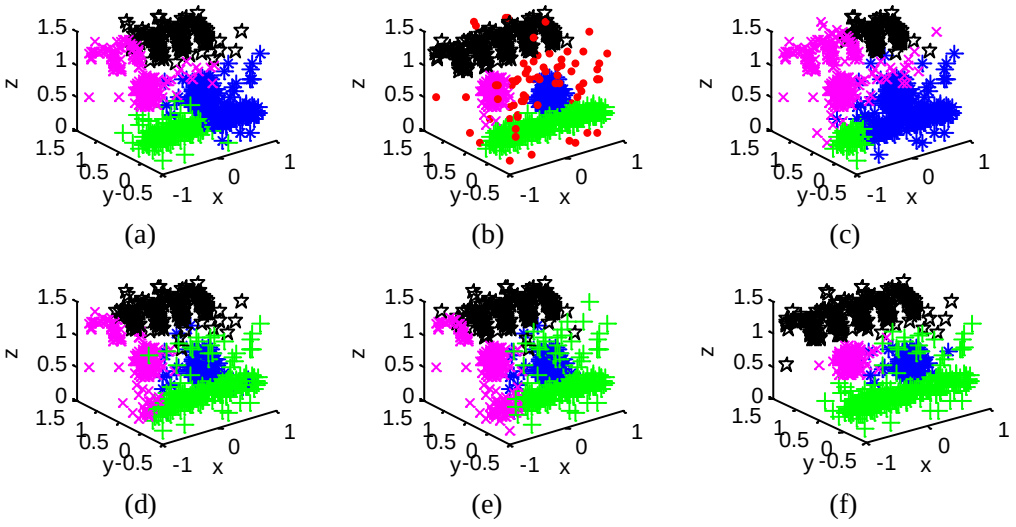


Fig. 10. The clustering results of D7 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

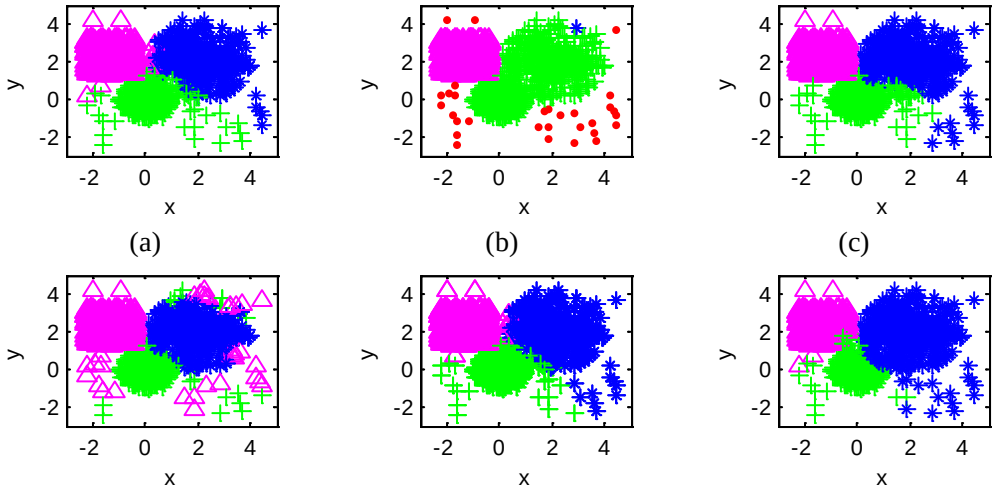


Fig. 11. The clustering results of D8 obtained by the six algorithms for (a) K-means; (b) DBSCAN; (c) RLClu; (d)NJW; (e)SC-DA; (f)CRG-clustering.

To further verifying the clustering performance of the algorithms, three statistical score functions (overall Accuracy, Rand index and Kappa index) are employed. Let n be the total number of data points, and n_{ij} denote the number of points classified into class i as produced by the algorithm which also are in cluster j in the true cluster structure. Then $n_{i\cdot} = \sum_j n_{ij}$ be the number of the points classified into cluster i in the data set under experiment, and $n_{\cdot j} = \sum_i n_{ij}$ be the number of points classified into class j in the real cluster structure.

(1) Overall Accuracy

The Overall Accuracy (OA) is the most widely used statistical score function for the validation. It denotes the percentage of correctly classified data points in the dataset. It is given by:

$$OA = \frac{\sum_i n_{ii}}{n} \tag{8}$$

The OA returns values in the interval $[0,1]$, and the optimum score is 1, with higher scores being better.

(2) Kappa Index

The Kappa Index (KI) measures the agreement between the clustering results produces by algorithm and the true cluster structure. It can be calculated as:

$$KI = \frac{n\sum_i n_{ii} - \sum_i n_{i+} n_{+i}}{n^2 - \sum_i n_{i+} n_{+i}} \tag{9}$$

(3) Rand index

The Rand Index [37] measures the agreement of the clustering result with the true cluster structure. It counts the number of pair-wise co-assignments of data items between the two partitions. Let n_s be the number of pairs of patterns that are assigned to the same cluster in both the resultant partition and the true cluster structure, and n_d be the number of pairs of patterns that are assigned to different clusters in both the resultant partition and the true cluster structure. The Rand index is defined as:

$$RI = \frac{n_s + n_d}{C_n^2} \tag{10}$$

where $C_n^2 = n(n-1)/2$. The value of Rand index will range from 0 to 1, with 0 indicating that the original classification and the clustered classification do not agree on any pair of points and 1 indicating that the original classification and the clustered classification are exactly the same.

In the following, the three statistical measures described above are used to measure the clustering performance of the algorithms. **Table 4** and **Table 5** show the mean and standard deviation of the overall accuracy (in %), the kappa index (in %), and the rand index for the artificial data sets (D1~D8) and the seven real data sets, respectively. All of which are calculated over 20 runs of the six clustering algorithms.

From **Table 4** and **Table 5**, one may observes that our approach outperforms K-means, DBSCAN, RLClu, NJW, and SC-DA in a statistically significant manner. For D1, D2, D3 which are the separable data sets with arbitrary shape, the proposed algorithm and DBSCAN can cluster the data sets accurately since the two algorithm both cluster the points according to the neighborhood relationship and never mind the limitation caused by the shape of the clustering data. For D4, D5, D6 which are inseparable data sets with arbitrary shape, the performance of K-means is poor due to it is not suitable for finding the non-spherical cluster. DBSCAN, RLClu, NJW, and SC-DA will perform worst when meet the inseparable data sets, the measure value of the clustering results of the five algorithms is lower than our algorithm. For D7 with noise, the performance of our algorithm and DBSCAN do not affected by the noise. Because our CRG-clustering picked the points with the most importance as the seed point of every regions and passed the label according to the neighborhood relationship. For D8 owning the inseparable data sets with noise, the performance of DBSCAN became worst when meets the inseparable data sets and the K-means, RLClu, NJW, and SC-DA performed worse than our CRG-clustering algorithm due to the noise data points.

Table 4. Mean and standard deviation (in parentheses) of four statistical validity measures produced by the six algorithms for the artificial data sets

Data	Measure	Mean and std. dev. of the validity measure					
		K-means	DBSCAN	RLClu	NJW	SC-DA	CRG-clustering
D1	OA	77.73 (1.53e-1)	100.00 (0)	79.00 (1.17e-29)	79.75 (3.53)	71.41 (7.24e-5)	100.00 (0)
	KA	66.60 (3.44e-1)	100.00 (0)	68.50 (5.19e-30)	64.51 (11.55)	57.11 (1.63e-4)	100.00 (0)
	RI	0.7814 (5.90e-4)	1.0000 (0)	0.7938 (1.17e-31)	0.8662 (1.54e-2)	0.7613 (6.28e-7)	1.0000 (0)
D2	OA	88.56 (2.77)	100.00 (0)	100.00 (0)	85.57 (1.39)	93.65 (1.46)	100.00 (0)
	KA	84.30 (5.22)	100.00 (0)	100.00 (0)	77.89 (3.33)	90.84 (2.97)	100.00 (0)
	RI	0.9440 (5.27e-3)	1.0000 (0)	1.0000 (0)	0.9293 (4.51e-3)	0.9702 (4.08e-3)	1.0000 (0)
D3	OA	34.38 (6.35e-4)	100.00 (0)	100.00 (0)	79.09 (3.85)	83.85 (5.12)	100.00 (0)
	KA	1.53 (1.30e-3)	100.00 (0)	100.00 (0)	61.11 (13.66)	73.20 (14.37)	100.00 (0)
	RI	0.5541 (3.06e-9)	1.0000 (0)	1.0000 (0)	0.8623 (1.64e-2)	0.9040 (1.80e-2)	1.0000 (0)
D4	OA	84.48 (7.65e-3)	64.58 (1.30e-32)	78.75 (5.19e-30)	99.58 (0)	98.33 (1.30e-30)	100.00 (0)
	KA	68.54 (3.25e-2)	2.92 (5.07e-35)	59.18 (5.19e-30)	99.10 (1.30e-30)	96.36 (1.17e-29)	100.00 (0)
	RI	0.7368 (1.49e-4)	0.5406 (1.30e-32)	0.6639 (5.19e-32)	0.9917 (5.19e-32)	0.9671 (5.19e-32)	1.0000 (0)
D5	OA	75.03 (1.43)	53.33 (0)	99.67 (5.19e-30)	68.18 (6.80e-1)	79.58 (5.21e-1)	98.50 (5.19e-32)
	KA	72.98 (1.67)	48.28 (5.19e-30)	99.64 (1.30e-30)	65.37 (7.99e-1)	77.95 (6.20e-1)	98.39 (5.19e-32)
	RI	0.9639 (3.53e-4)	0.7507 (1.30e-32)	0.9991 (1.30e-32)	0.8920 (2.00e-3)	0.9715 (8.13e-5)	0.9961 (1.17e-31)
D6	OA	38.28 (1.56e-4)	66.75 (5.19e-30)	47.83 (3.24e-29)	77.59 (1.30)	77.16 (1.04)	86.58 (0)
	KA	7.43 (3.52e-4)	50.13 (0)	21.75 (3.24e-30)	66.30 (2.99)	65.74 (2.34)	79.88 (0)
	RI	0.5578 (7.66e-9)	0.7776 (5.19e-32)	0.5478 (1.30e-32)	0.8106 (1.45e-3)	0.8062 (1.24e-3)	0.8570 (5.19e-32)
D7	OA	80.26 (9.02e-2)	100.00 (0)	66.88 (5.19e-30)	86.07 (8.91e-2)	75.66 (1.65)	100.00 (0)
	KA	73.44 (2.09e-1)	100.00 (0)	55.83 (1.30e-30)	81.43 (1.58e-1)	67.32 (3.04)	100.00 (0)
	RI	0.8431 (2.71e-4)	1.0000 (0)	0.7793 (0)	0.8793 (4.14e-4)	0.8433 (1.29e-3)	1.0000 (0)
D8	OA	96.67 (5.19e-30)	66.83 (5.19e-30)	97.42 (1.30e-30)	85.33 (2.27)	97.27 (2.63e-5)	98.00 (5.19e-30)
	KA	95.00 (1.17e-29)	50.25 (5.19e-30)	96.13 (1.30e-30)	78.00 (1.70e-3)	95.90 (5.92e-5)	97.00 (0)
	RI	0.9574 (3.24e-31)	0.7776 (0)	0.9669 (0)	0.8880 (9.35e-3)	0.9648 (4.03e-7)	0.9740 (1.30e-32)

Table 5. Mean and standard deviation (in parentheses) of four statistical validity measures produced by the six algorithms for real data sets

Data	Measure	Mean and std. dev. of the validity measure					
		K-means	DBSCAN	RLClu	NJW	SC-DA	CRG-clustering
IR	OA	80.50 (1.88)	66.00 (0)	88.67 (0)	88.57 (9.64e-1)	89.47 (7.08e-1)	92.00 (0)
	KA	70.75 (4.24)	49.00 (1.30e-30)	83.00 (1.30e-30)	81.60 (3.38)	84.20 (1.59)	88.00 (5.19e-30)
	RI	0.8323 (4.45e-3)	0.7719 (0)	0.8737 (2.08e-31)	0.8906 (2.80e-3)	0.8941 (2.45e-3)	0.9055 (5.19e-32)
VE	OA	43.29 (2.04e-2)	26.10 (1.30e-30)	45.51 (3.24e-31)	40.43 (3.17e-1)	42.30 (1.04e-2)	45.60 (3.24e-31)
	KA	24.42 (4.16e-2)	0.60 (3.17e-34)	19.44 (3.24e-31)	20.67 (8.68e-1)	19.87 (3.15e-1)	27.67 (0)
	RI	0.6512 (4.18e-4)	0.2564 (0)	0.6535 (1.30e-32)	0.5673 (1.48e-2)	0.6186 (2.47e-4)	0.6189 (5.19e-32)
SO	OA	77.23 (2.23)	82.98 (1.17e-29)	57.45 (1.30e-30)	88.72 (2.48)	82.34 (1.19)	97.87 (1.30e-30)
	KA	67.64 (4.45)	75.96 (0)	46.86 (3.24e-31)	83.12 (6.00)	74.61 (2.68)	97.11 (1.30e-30)
	RI	0.8805 (5.38e-3)	0.8594 (5.19e-32)	0.9121 (1.30e-32)	0.9340 (1.34e-2)	0.8772 (3.40e-3)	0.9760 (0)
SE	OA	88.74 (5.43e-4)	34.76 (1.30e-30)	90.00 (0)	89.52 (1.30e-30)	89.52 (1.30e-30)	91.43 (1.30e-30)
	KA	83.11 (1.22e-3)	2.14 (5.07e-33)	85.00 (1.30e-30)	84.29 (0)	84.29 (0)	87.14 (5.19e-30)
	RI	0.8660 (6.24e-6)	0.3430 (3.24e-33)	0.8817 (1.30e-32)	0.8741 (0)	0.8741 (0)	0.8944 (1.30e-32)
CH	OA	54.72 (1.74e-1)	51.60 (0)	51.88 (0)	52.13 (8.45e-5)	53.19 (1.30e-30)	59.72 (5.19e-30)
	KA	8.82 (6.04e-1)	0.07 (4.95e-36)	0.30 (0)	4.41 (3.71e-4)	3.12 (3.17e-32)	18.32 (3.24e-31)
	RI	0.5075 (8.26e-5)	0.5003 (5.19e-32)	0.5005 (1.30e-32)	0.5007 (6.31e-9)	0.5018 (0)	0.5187 (0)
LI	OA	55.12 (7.37e-3)	57.68 (5.19e-30)	56.81 (1.30e-30)	57.10 (5.19e-30)	58.10 (1.10e-3)	58.84 (1.30e-30)
	KA	-0.44 (1.43e-2)	-0.58 (0)	-2.01 (0)	-1.73 (0)	3.02 (4.78e-3)	2.60 (1.27e-33)
	RI	0.5039 (3.03e-6)	0.5104 (1.30e-32)	0.5078 (0)	0.5087 (1.30e-32)	0.5117 (1.03e-6)	0.5142 (5.19e-32)
PI	OA	66.80 (0)	65.23 (0)	65.23 (0)	64.71 (1.30e-30)	65.35 (1.61e-5)	71.22 (1.30e-30)
	KA	25.94 (2.27e-5)	0.49 (7.92e-35)	0.49 (7.92e-35)	2.72 (0)	5.19 (1.99e-4)	38.22 (1.30e-30)
	RI	0.5559 (0)	0.5458 (5.19e-32)	0.5458 (5.19e-32)	0.5427 (1.30e-32)	0.5465 (6.03e-8)	0.5896 (1.30e-32)

5.3 Effect of the number of the nearest neighbors

As we known, the CRG-clustering algorihtm need two parameters, the number of the clusters K and the number of the nearest neighbors k. Here, the number of the clusters K should be specified in advance. Therefore, we only discuss the influence of k (the number of the nearest neighbors).

In order to analyze the sensitivity of the parameter k, we need to find the range of the

parameter k. Here, the range of the number of nearest neighbors k needs to meet two constrains. The first one is that k should range from 1 to $n-1$ where n is the number of the data points. Second, the CRG-clustering algorithm need divide the seed point and its k neighbors into the same cluster, so k should be small. So we choose $[1,\sqrt{n}]$ as the range of k. In order to see the effect of the number of nearest neighbors, we conduct a series of experiments, in which we vary the value of k. The OA results with different k for the data sets are given in **Fig. 12.**

From the result of the **Fig. 12**, we can see that the performance of CRG-clustering is insensitive to the choice of k. OA of the clustering result first increase to the maxium value and holds the maxium value for some k. Thus, we set $k \in [3,9]$ in our experiments.

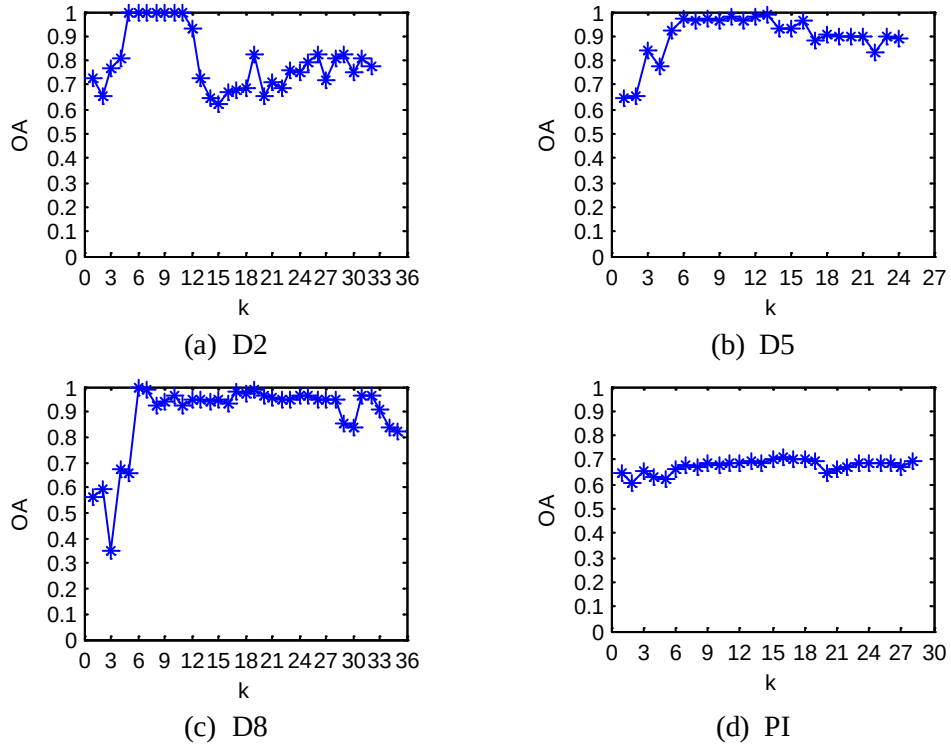


Fig. 12. The trend graph of the measure value with different nunber of k.

6. Conclusion

In this paper, a clustering algorithm based on connected region generation algorithm has been developed for unsupervised classification. The algorithm generates the K connected regions from the seed points by using a CRG algorithm. Then, the region expansion strategy and the consensus criterion are used to expand the connected regions to ensure that

all the points are included in a connected region. The CRG-clustering algorithm classifies the points based on the connectivity between point and its neighbors, therefore it can be used to discover clusters of arbitrary shape. Experimental results both on the synthetic datasets and the real world datasets show that the proposed algorithm is effective and insensitive to the noise.

Although the results presented here are extremely encouraging, there is an issue that deserves in-depth study in the future. The desired number of clusters should be specified in advance. A mechanism that can estimate the number of clusters should be investigated.

References

[1] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification, 2nd Edition," *Wiley*, New York, 2000. Article (CrossRef Link)

[2] T. Senthil and B. Kannapiran, "EETCA: Energy Efficient Trustworthy Clustering Algorithm for WSN," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 11, pp. 5437-5454, November, 2016. Article (CrossRef Link)

[3] F. Aadil, S. Khan, K. B. Bajwa, M. F. Khan and A. Ali, "Intelligent Clustering in Vehicular ad hoc Networks," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 8, pp. 3512-3528, August, 2016. Article (CrossRef Link)

[4] Y. Su, X. Zhu and W. Z Nie, "Multiple Person Tracking based on Spatial-temporal Information by Global Graph Clustering," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 6, pp. 2217-2229, June, 2015. Article (CrossRef Link)

[5] B. A. Galitsky, G. Dobrocsi, J. L. D. L. Rosa and S. O. Kuznetsov, "Using Generalization of Syntactic Parse Trees for Taxonomy Capture on the Web," in *Proc. of the 19th international conference on Conceptual structures for discovering knowledge*, pp. 104-117, July 25-29, 2011. Article (CrossRef Link)

[6] B. Everitt, S. Landau and M. Leese, "Cluster Analysis," Arnold, London, 2001. Article (CrossRef Link)

[7] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, January, 1967. Article (CrossRef Link)

[8] R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," *Computer Journal*, vol. 16, no. 1, pp. 30-34, January, 1973. Article (CrossRef Link)

[9] M. Ester, H. Kriegel, S. Jiirg and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, August 2-4, 1996. Article (CrossRef Link)

[10] 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, August, 2000. Article (CrossRef Link)

[11] B. J. Frey, D. Dueck, "Clustering by Passing Messages between Data Points," *Science*, vol. 315, no. 5814, pp. 972-976, February, 2007. Article (CrossRef Link)

[12] Y. Liu, Y. Liu, and K. C. C. Chan, "Dimensionality reduction for heterogeneous dataset in rushes editing," *Pattern Recognition*, vol. 42, no. 2, pp. 229-242, 2009. Article (CrossRef Link)

[13] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proc. of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027-1035, January 7-9, 2007. Article (CrossRef Link)
[14] P. Bradley, O. Mangasarian, and W. Street, "Clustering via Concave Minimization," *Advances in Neural Information Processing Systems*, pp. 368-374, January, 1996. Article (CrossRef Link)
[15] D. Defays, "An Efficient Algorithm for a Complete Link Method," *Computer Journal*, vol. 20, no. 4, pp. 364-366, January, 1977. Article (CrossRef Link)
[16] J. A. García , J. Fdez-Valdivia, F. J. Cortijo and R. Molina, "A dynamic approach for clustering data," *Signal Processing*, vol. 44, no. 2, pp. 181-196, June, 1995. Article (CrossRef Link)
[17] S. Feng, J. Fan, H. Tan, Y. He, H. Mao, W. Luo and D. Ma, "MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce," in *Proc. of IEEE International Conference on Parallel and Distributed Systems. IEEE Computer Society*, pp. 473-480, December 7-9, 2011. Article (CrossRef Link)
[18] Y. Kim, K. Shim, M. Kim and J. S. Lee, "DBCURE-MR: An efficient density-based clustering algorithm for large data using Map Reduce," *Information Systems*, vol. 42, pp. 15-35, June, 2014. Article (CrossRef Link)
[19] C. Cassisi, A. Ferro, R. Giugno, G. Pigola and A. Pulvirenti, "Enhancing density-based clustering: Parameter reduction and outlier detection," *Information Systems*, vol. 38, no. 3, pp. 317-330, May, 2013. Article (CrossRef Link)
[20] A. Y. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," *Proceedings of Advances in Neural Information Processing Systems*, vol. 14, pp. 849-856, April, 2002. Article (CrossRef Link)
[21] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1601-1608, January, 2004. Article (CrossRef Link)
[22] C. Hong and D. Y. Yeung, "Robust path-based spectral clustering with application to image segmentation," in *Proc. of 10th IEEE International Conference on Computer Vision, IEEE Computer Society*, pp. 278-285, October 17-20, 2005. Article (CrossRef Link)
[23] 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, January, 2011. Article (CrossRef Link)
[24] C. H. Q. Ding, X. He, H. Zha, M. Gu and H. D. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. of IEEE International Conference on Data Mining. IEEE Computer Society*, pp. 107-114, November 29-December 2, 2001. Article (CrossRef Link)
[25] I. Fischer and J. Poland, "Amplifying the block matrix structure for spectral clustering," *Idsia*, pp. 21-28, January, 2005. Article (CrossRef Link)
[26] 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, November, 2006. Article (CrossRef Link)
[27] T. Xia, J. Cao, Y. Zhang and J. Li, "On defining affinity graph for spectral clustering through ranking on manifolds," *Neurocomputing*, vol. 72, no. 13–15, pp. 3203-3211, August, 2009. Article (CrossRef Link)
[28] J. Cao, P. Chen, W. K. Ling, Z. Yang and Q. Dai, "Spectral Clustering with Sparse Graph Construction Based on Markov Random Walk," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 7, pp. 2568-2584, July, 2015. Article (CrossRef Link)

[29] I. E. Givoni, B. J. Frey, "A binary variable model for affinity propagation," *Neural Computation*, vol. 21, no. 6, pp. 1589-1600, June, 2009. Article (CrossRef Link)
[30] I. E. Givoni, B. J. Frey, "Semi-Supervised Affinity Propagation with Instance-Level Constraints," in *Proc. of the international conference on Artificial Intelligence & Statistics*, pp. 161-168, 2009. Article (CrossRef Link)
[31] M. Leone, M. Weigt, "Clustering by soft-constraint affinity propagation: applications to gene-expression data," *Bioinformatics*, vol. 23, no. 20, pp. 2708-2715, October, 2007. Article (CrossRef Link)
[32] M. L. Sumedha, M. Weigt, "Unsupervised and semi-supervised clustering by message passing: soft-constraint affinity propagation," *European Physical Journal B*, vol. 66, no. 1, pp. 125-135, October, 2008. Article (CrossRef Link)
[33] C. Furtlehner, M. Sebag, X. Zhang, "Scaling analysis of affinity propagation," *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 81, no. 6 Pt 2, pp. 066102 , 2009. Article (CrossRef Link)
[34] C. Fu, J. Wang, X. Chen, Z. Qin, M. Zhao, "Flow Transformation of Anonymous Communication Based on Hierarchical Weighted Affinity Propagation Clustering," *Journal of Computational Information Systems*, vol. 7, no. 1, 2011. Article (CrossRef Link)
[35] A. Rodriguez and A. Laio, "Machine learning. Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492-1496, June, 2014. Article (CrossRef Link)
[36] G. Wang and Q. Song, "Automatic Clustering via Outward Statistical Testing on Density Metrics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 1971-1985, August, 2016. Article (CrossRef Link)
[37] F. R. Bach and M. I. Jordan, "Learning Spectral Clustering," *Advances in Neural Information Processing Systems*, vol. 16, no. 2, pp. 2006, June, 2003. Article (CrossRef Link)

Liuwei Feng is currently pursuing a master degree at Beijing Jiaotong University of Computer and Information Technology. Her current research interests include clustering algorithm.

Dongxia Chang received the B.S. and M.S. degree in applied mathematic from Xidian University, Xi'an, China, in July 2000 and April 2003, and the PhD degree from Tsinghua University, Beijing, China, in July 2009, respectively. From May 2010 to May 2012, she was a post-doctoral research at Beijing Jiaotong University, Beijing, China. Since May 2012, she has been an associate professor at the Institute of Information Science, Beijing Jiaotong University. Her current research interests include pattern recognition, clustering analysis, image analysis.

Yao Zhao received the BS degree from Fuzhou University, China, in 1989, and the ME degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an associate professor at BJTU in 1998 and became a professor in 2001. From 2001 to 2002, he was a senior research fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as associate editors of IEEE Transactions on Cybernetics, IEEE Signal Processing Letters, and an area editor of Signal Processing: Image Communication (Elsevier), etc. He was named a distinguished young scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a senior member of the IEEE.