A novel bidirectional clustering algorithm based on local density

BAICHENG LV
Dalian University of Technology
WENHUA WU (lxyuhua@dlut.edu.cn)
Dalian University of Technology
ZHIQIANG HU
Newcastle University

Research Article

Keywords: cluster analysis, bidirectional clustering algorithm based on local density (BCALoD), cluster numbers

Posted Date: February 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-141525/v1

Version of Record: A version of this preprint was published at Scientific Reports on July 9th, 2021. See the published version at https://doi.org/10.1038/s41598-021-93244-2.

Abstract

With the widely application of cluster analysis, the number of clusters is gradually increasing, as is the difficulty in selecting the judgment indicators of cluster numbers. Also, small clusters are crucial to discovering the extreme characteristics of data samples, but current clustering algorithms focus mainly on analyzing large clusters. In this paper, a bidirectional clustering algorithm based on local density (BCALoD) is proposed. BCALoD establishes the connection between data points based on local density, can automatically determine the number of clusters, is more sensitive to small clusters, and can reduce the adjusted parameters to a minimum. On the basis of the robustness of cluster number to noise, a denoising method suitable for BCALoD is proposed. Different cutoff distance and cutoff density are assigned to each data cluster, which results in improved clustering performance. Clustering ability of BCALoD is verified by randomly generated datasets and city light satellite images.

Introduction

Cluster analysis is a data processing algorithm using unsupervised learning that has been widely used in machine learning and information recognition [1]. The main clustering methods used so far include K-means clustering, the Gaussian mixture model (GMM), hierarchical clustering, mean shift clustering, and density-based spatial clustering of applications with noise (DBSCAN) [2]. K-means clustering can achieve good clustering performance for spherical datasets and is one of the most widely used clustering methods [3]. However, K-means clustering has poor clustering ability in dealing with aspheric data, and it must determine the number of clusters in advance [4, 5]. GMM can calculate the expectation and variance of Gaussian datasets [6], but it is difficult to use for effective cluster analysis of data with poor Gaussian characteristics. Mean shift clustering [7] can automatically determine the number of clusters, but for a small-cluster dataset the selection of the sliding window radius may affect the clustering results. DBSCAN does not need to determine the number of clusters and can do cluster analysis for data with arbitrary shapes [8]. However, if the density of the sample set is not uniform, and the differences in cluster spacing vary somewhat, the clustering quality is poor and too many parameters must be adjusted. Rodriguez et al. [9] proposed a density-based fast searching method that overcomes the drawbacks of conventional data clustering methods and can quickly cluster aspheric sets. Unlike the mean shift method, their procedure does not require embedding the data in a vector space; however, it has the problem of requiring manual work to select the number of clusters and easily ignores small clusters, making clustering results depend on manual experience and subjective judgment.

As cluster analysis is applied in more and more fields, the amount of sample data increases, and the cluster number also increases. The increase in the cluster number increases the difficulty in judging the number of clusters for various clustering methods. Small clusters denote points in the data sample that are far away from large clusters and have aggregation phenomenon. Small clusters seldom appear, but in the field of engineering structural safety analysis [10, 11], small clusters usually indicate extreme or unconsidered working conditions. In some cases, accurately identifying small-cluster information is more important than identifying large clusters. Unfortunately, existing methods focus mainly on forming large

clusters, and there is little research on small clusters. Also, existing data clustering algorithms for processing noise rely mainly on an artificial threshold. In cluster analysis, using a unified standard denoising indicator increases the discarding of small clusters because the clustering characteristics of each cluster are different. These problems impose challenges on traditional clustering methods.

In this paper, a bidirectional clustering algorithm is proposed that works by combining the advantages of mean shift clustering and clustering by fast search and find of density peaks. The proposed algorithm is divided into an up process and a down process. In the up process, the local density of different data points is calculated to find high-local-density points nearest to data points, and then data chains are formed from data points ranging from low local density to high local density. In the down process, the highest local-density data points are treated as clustering centers, and then the data chains are merged, all data points are traversed, and finally the clustering operation is completed from high-local-density data to low-local-density data. By comparing K-means clustering, GMM, and BCALoD, it was found that the proposed clustering algorithm can quickly calculate the number of clusters and has good recognition performance for small clusters. In the clustering process, the number of parameters requiring adjustment is reduced to a minimum.

The noise problem was also addressed in this study in real-world data measurement [12] by using the characteristics of a large range and the non-evident aggregation property of noise. A noise reduction decision indicator is proposed that is suitable for BCALoD and assigns a different cutoff distance and cutoff density for each cluster to retain important data as much as possible. By clustering random data examples and city light photographs, it was found that the proposed algorithm has good denoising performance.

Bidirectional Clustering Algorithm Based On Local Density (Bcalod)

Assume that the clustering center is data points with maximum local density, and then calculate the local density (ρ_i) of each point in the dataset:

$$\rho_i = \sum_{j \in I \setminus \{i\}} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \qquad (1)$$

where d_{ij} is the Euclidean distance between point j and point i, and d_c is the cutoff distance selected in the proportion of 1% to 2% of the total distance of data points. Because a Gaussian kernel is used to represent the local distance ρ [7,13], the probability that the local density is the same at all points in the area is low. Assuming that the local density ρ of each point is different, data points are sorted according to the value of the local density ρ. Fig.1 shows the schematic diagram of BCALoD.

In the up process, the condition for judging point k as a cluster center is given by within the cutoff distance, no point has a larger local density than point k. For all points in the dataset, each point

establishes a connection with only its upper layer, and there is no relation with the next layer; that is, the tops of the selected data chains have nothing to do with the starting point. The up process is started by selecting from the point with the smallest local density ρ. The addressing sequence reflects only the path from bottom to top and does not change the local density of the cluster centers. For small clusters, if no point within the cutoff distance has a higher local density, those clusters are formed as one category respectively. Overall, the up process can be regarded as the inverse operation of clustering data points in a density peak clustering algorithm.

In clustering by fast search and find of density peaks, when the data sizes of large clusters are much greater than those of small clusters, the information of small clusters is easily overwhelmed, resulting in subjectivity in determining the number of clusters. Bidirectional clustering can properly identify the clustering of small clusters. The mean shift algorithm must embed random data points and continuously iterate the sliding window, which leads to a high data calculation cost and usually hampers analyzing high-dimensional clusters. The proposed method of constructing data clusters by data chains can reduce the operational cost and ensure that the selection of clustering centers is unrelated to the initial selection. BCALoD establishes the data chains, is not required to embed the data into the vector space, can automatically determine the number of clusters, is more sensitive than other methods to small clusters, and reduces the adjusted parameters to the lowest. In summary, this method combines the advantages of clustering by fast search and find of density peaks and mean shift clustering.

We used a Gaussian mixture distribution to generate a 2D dataset (Case 1), as shown in Fig.2a. Fig.2b shows the results of the BCALoD clustering algorithm. Table 1 shows the size of each cluster.

Table. 1. Cluster sizes of 2D Gaussian mixture distribution datasets

No.	1-7	8	9	10	11	12	13	14
Set size	0	0	50	400	800	1000	2000	5000
Cluster Result	1	3	50	394	778	996	2019	5003

Table 1. Cluster size. The resulting clusters 1 through 7 contained one data point each, while cluster 8 contained three data points. These data points were distant from the other data points, which are referred to as discrete points. In the BCALoD cluster analysis, such data points are independently grouped into one category. The size of cluster 8 is 50, which is much smaller than the amount of data contained in other clusters; this type of cluster is referred to as a small cluster. The results demonstrate the excellent clustering performance of small clusters by the BCALoD algorithm. The accuracy rate is over 98.5%.

Noise Recognition And Cutoff Distance Optimization

The influence of noise is usually unavoidable in the real world [14]. How to deal with noise reasonably is an important problem in cluster analysis [15]. The noise in cluster analysis has the characteristics of large range and small aggregation. Cluster analysis reveals that empirical values can quickly determine a relatively reasonable range. When the distance between two clusters is artificially increased or noise is added, the values of the cutoff distances also change, whereas the clustering properties in the clusters do not change. Meanwhile, because the aggregation degree of each cluster is different, it is unreasonable to use a uniform cutoff distance for denoising.

We introduced noise points to Case 1, as shown in Fig.3a. The BCALoD algorithm was used to do cluster analysis for all data. Fig.3b shows the clustering results.

If the number of noise points is much lower than the number of real clusters, denoising can be done simply by comparing cluster sizes. However, when there are real small clusters or noise clusters, it is easy to discard real small clusters by denoising using only cluster size. Because of the characteristics of noise, the local density of noise is much lower than that of actual clustering centers. Therefore, when clustering density is gradually filtered from small to large, the low clustering density data points in noise clusters and real clusters are deleted first, while the clustering centers of real clusters are not deleted (Fig.4a). In other words, the number of clustering centers of real clusters has strong robustness against changes in clustering density.

Set the range of the local density stabilized in i clusters as h_i^1. The maximum local density corresponding to the i cluster is h_i^2. Set the decision indicator D_i as

$$D_i = \frac{h_i^1}{h_i^2} \tag{2}$$

If h_i^1 is high, the number of clusters selected as the i cluster is stable. If h_i^2 is low, the number of clusters retained is high, which means that more data points are retained. Fig.4b shows the decision indicators of the noise-containing 2D Gaussian mixture distribution dataset. When the cluster number of the dataset is determined, the cutoff local densities and cutoff distances of each cluster can be calculated through data filtering (Fig.4c), and then the clustering results can be obtained (Fig.4d).

BCALoD algorithm could retain more sensitivity to small clusters and save non-noise data to the greatest extent possible. Also, the algorithm used a decision indicator to scientifically assign a different cutoff local density and cutoff distance for each cluster. Because the denoising process does not require repeated calculations on data points, this algorithm can reduce calculation costs.

Results

We used a combination dataset to verify the BCALoD algorithm by comparing the results of different clustering algorithms, as shown in Fig.5. By comparing the BCALoD algorithm, K-means clustering, and mixed Gaussian clustering, the relative performance of the BCALoD algorithm was determined. In K-means clustering, the silhouette coefficient [16] is used to calculate the clustering and separation degrees, and the maximum value of the silhouette coefficient is selected as the number of clusters, which completes K-means clustering. In mixed Gaussian clustering, the Bayesian information criterion (BIC) [17] is used to select the number of clusters, and the minimum value of the BIC is selected as the data number of clusters.

For image clustering, pixel values of different data points are different. Because the pixel value of L_i is an integer, it can easily yield the same local density. Therefore, taking the pixel value as the main factor, the local density was defined as:

$$P_i = \frac{\rho_i}{\max(\rho)} + L_i \qquad (3)$$

where ρ_i is the local density calculated by Eq. (1). Because L_i is an integer, the value of ρ_i of each point was normalized to make the local density of different data points different. Under the condition of the same pixel, data points were sorted by comparing ρ_i.

Fig.6 shows a city lighting satellite image and the clustering results by BCALoD cluster algorithm.

As shown in cases above, for nonspherical clusters, spirals datasets, noise-containing datasets, and city light image data, the proposed algorithm achieved good clustering performance. The only parameter that required adjustment in the clustering calculation was cutoff density, which reduced the difficulty of parameter adjustment in the clustering process as much as possible. The number of clusters was automatically determined by the method based on BCALoD, which was more sensitive to the information of small clusters, and assigned a different cutoff distance and cutoff local density for each data cluster.

Conclusions

A BCALoD algorithm is proposed in this paper that is based on local density that combines the merits of clustering by fast search and find of density peaks and mean shift clustering. The algorithm forms data chains from low-local-density data points to high-local-density data points, treats the latter as clustering centers, and then integrates the data chains and completes the clustering operations. The number of clusters is automatically determined by the BCALoD algorithm, which is more sensitive to small clusters, and reduced the number of parameters requiring adjustment to the lowest.

Using the characteristics of noise, a denoising method is proposed based on local density, which ensures a denoising effect and retains the sensitivity of the BCALoD algorithm to small clusters. The denoising

method assigns a cutoff local density and cutoff distance for each cluster, retaining useful data as much as possible. The proposed algorithm can also reduce calculation cost.

The BCALoD algorithm can do effective cluster analyses on aspheric clusters, spirals datasets, noise-containing datasets, and image data. The proposed algorithm can also achieve ideal clustering performance for clusters of various shapes. When the number of clusters is large, the BCALoD algorithm can quickly determine the number of clusters, is more sensitive to small clusters, and removes noise effectively.

References

1. Shirkhorshidi, A. S. *et al.* Big data clustering:A review. 14th International Conference on Computational Science and Its Applications, Guimaraes, Portugal. 2014: 707–720.
2. Chetan, D. & Meenakshi, B. An improvement of DBSCAN Algorithm to analyze cluster for large datasets. 2013 IEEE International Conference in MOOC, Innovation and Technology in Education, MITE 2013, 2013:42–46.
3. Liu, G. *et al.* Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. *Journal of Cleaner Production.* **183**, 304–314 (2018).
4. Michael, L. & Sumitra, M. A genetic algorithm that exchanges neighboring centers for k-means clustering. *Pattern Recognit. Lett.* **28** (16), 2359–2366 (2007).
5. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31** (8), 651–666 (2010).
6. Yang, M-S., Lai, C-Y. & Lin, C-Y. A robust em clustering algorithm for Gaussian mixture models. *Pattern Recogn.* **45** (11), 3950–3961 (2012).
7. Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **17** (8), 790–799 (1995).
8. Yang, K. *et al.* DBSCAN-MS: Distributed density-based clustering in metric spaces. 35th IEEE International Conference on Data Engineering, ICDE 2019, 2019: 1346–1357.
9. Rodriguez, A. & Laio, A. *Machine learning. Clustering by fast search and find of density peaks* 3441492–6(Science, (New York, N.Y.), 2014).
10. Kurasova, O. *et al.* Strategies for Big Data Clustering. 26th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2014, 2014: 740–747.
11. Wu, W. *et al.* Design, implementation and analysis of full coupled monitoring system of FPSO with soft yoke mooring system. *Ocean Eng.* **113**, 255–263 (2016).
12. Herranen Henrik, K. *et al.* Juri. Acceleration data acquisition and processing system for structural health monitoring. 2014 IEEE International Workshop on Metrology for Aerospace, MetroAeroSpace 2014,2014: 244–248.
13. Hinneburg, A. & Gabriel, H-H. DENCLUE 2.0: Fast clustering based on kernel density estimation. 7th International Symposium on Intelligent Data Analysis, IDA 2007, 2007: 70–80.

14. Yu, D., Wen-Hua, W. & Qian-Jin, Y. Prototype measurement for deep water floating platforms based on monitoring technology. ASME 2013 32nd International Conference on Ocean, Offshore and Arctic Engineering, OMAE 2013, 2013.
15. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science.* **344** (6191), 1492 (2014).
16. Errecalde, M. L., Cagnina, L. C. & Rosso, P. Silhouette + attraction: A simple and effective method for text clustering. *Natural Language Engineering.* **22** (05), 687–726 (2016).
17. Nishida, M. & Kawahara, T. Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing. *IEEE Transactions on Speech & Audio Processing.* **13** (4), 583–592 (2005).
18. It's Valley Fog Season (Online picture) .https://www.nasa.gov/image-feature/it-s-valley-fog-season

Figures

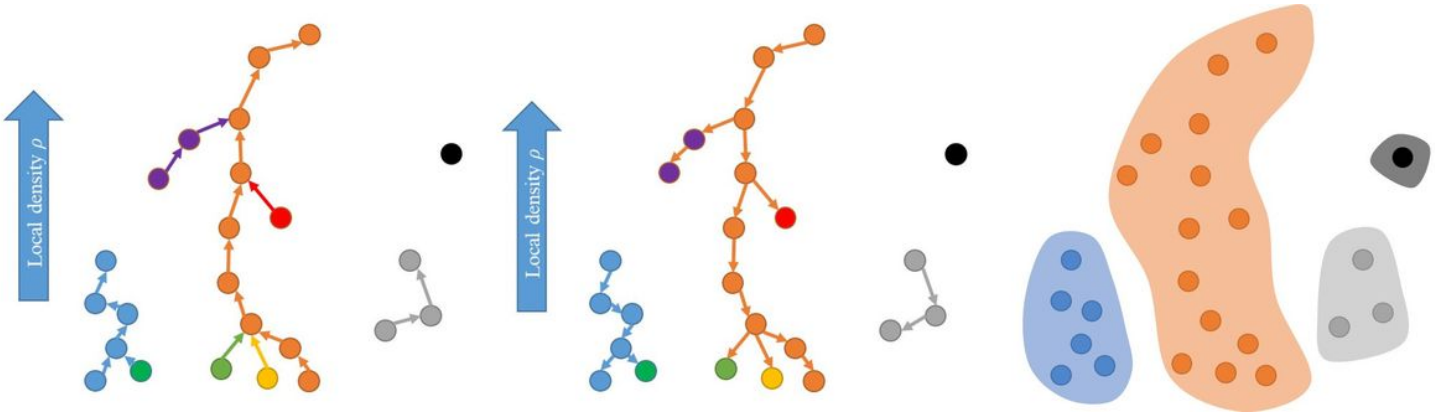

Figure 1

Bidirectional clustering algorithm based on local density. (Left) (a) Up process. Starting from the data point with minimum local density (qi), the data points qj whose local density is greater than ρi and closest to the qi point must be searched for. The search operation is repeated until there is no data point with a larger local density within the cutoff distance dc to form a data chain from q1 to qm. When the same calculation is done for the remaining data, the data chains of all data points are finally obtained. (Middle) (b) Down process. The clustering operation is finally completed by sorting out the data chains formed in the up process, classifying those data chains by their top data points, merging them, and then traversing all the points in the dataset, as shown in (Right) (c) clustering result.

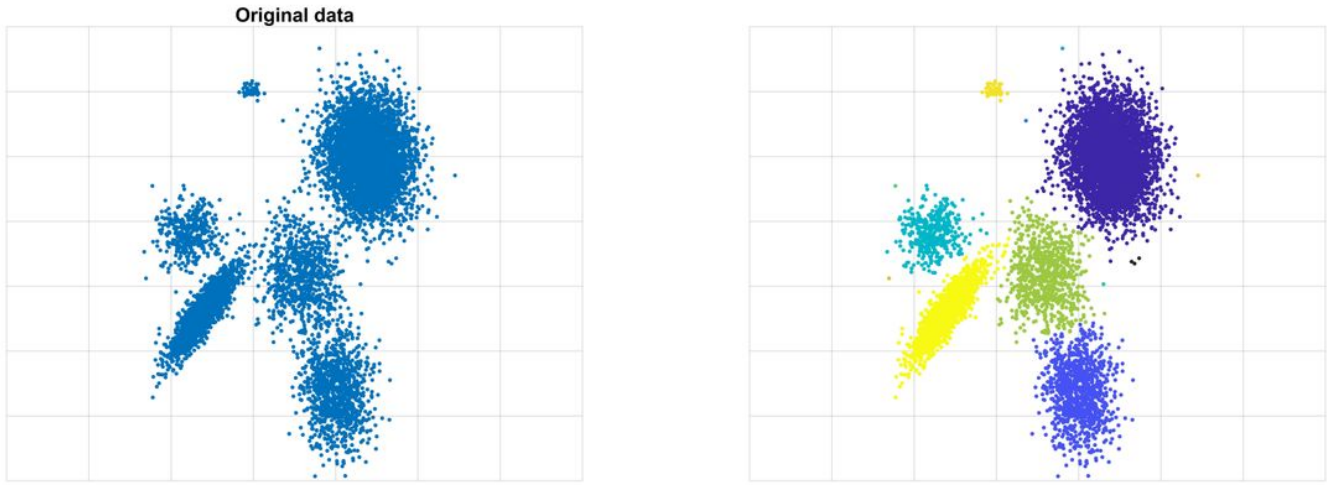

Figure 2

Cluster analysis of 2D Gaussian mixture distribution datasets. (Left)(a) Case 1 dataset. 2D Gaussian mixture distribution datasets. (Right) (b) Results of the BCALoD clustering. Clustering results of the BCALoD algorithm.

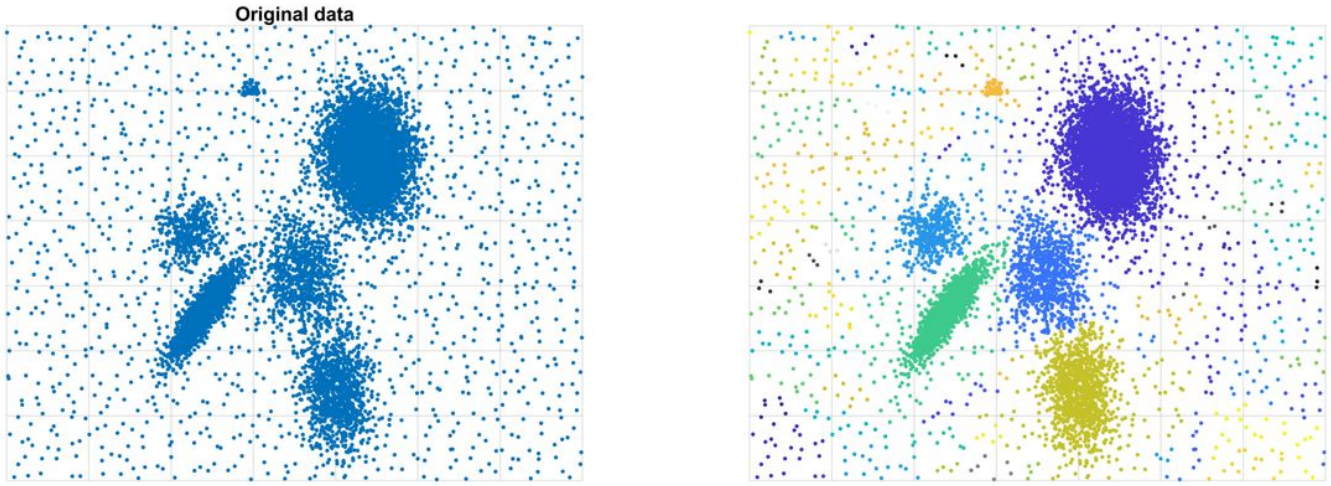

Figure 3

Clustering of a noise-containing 2D Gaussian mixture distribution dataset. (Left) (a) Noise points were introduced to Case 1. The BCALoD algorithm was used to do cluster analysis for all data, and the dc value in Case 1 was 1.18. (Right) (b) Fig.3b shows that some noise points were clustered into noise clusters, and the other noise points were clustered into real clusters.

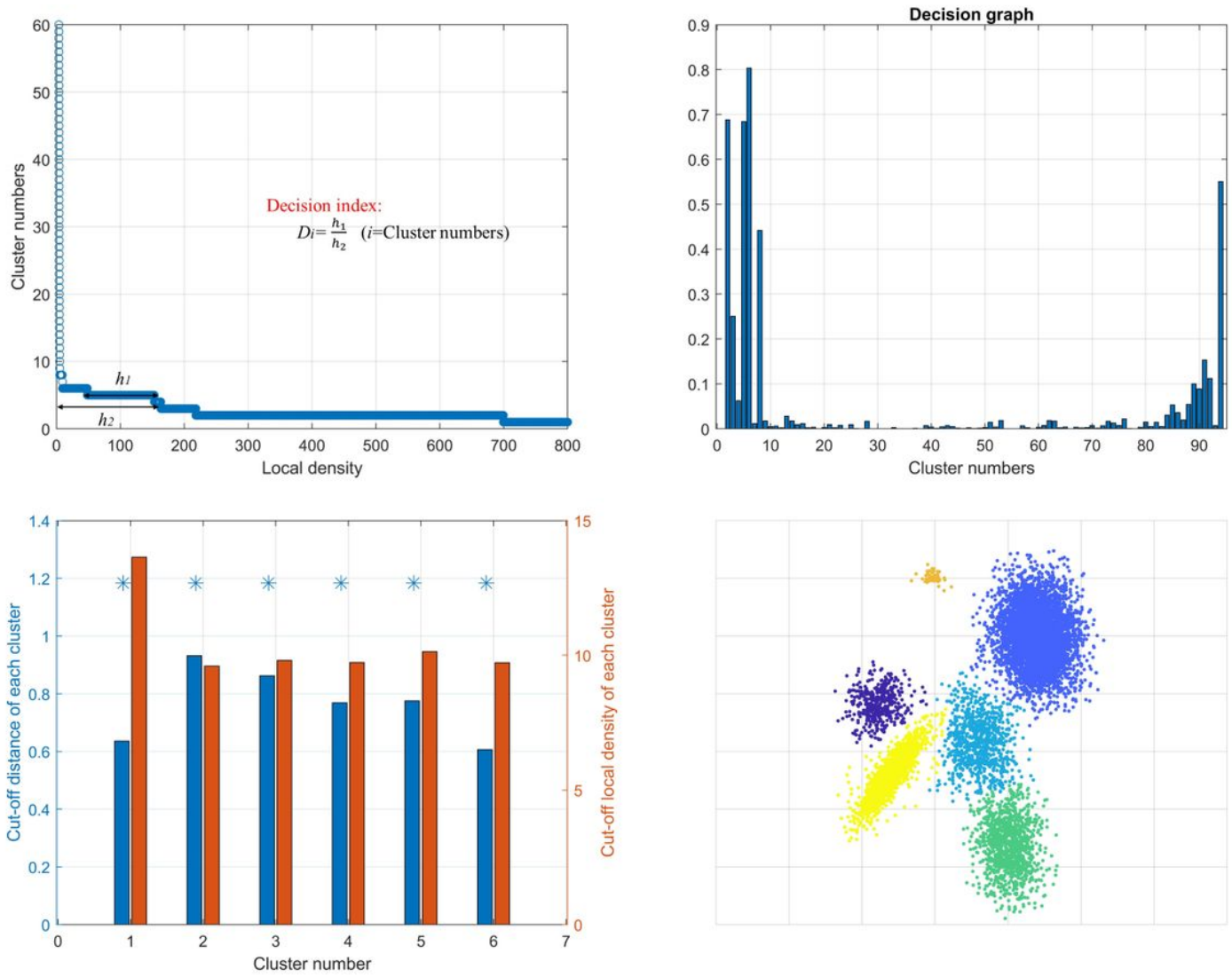

Figure 4

BCALoD noise recognition. (Top Left) (a) Fig. 4a shows the number of clusters filtered from small to large local density. When the local density gradually increased, it was observed that the number of clusters gradually decreased. When the local density was small, the number of clusters decreased sharply, and the data points deleted during this period were noise points. When the cutoff local density increased to a certain value, the number of clusters became 1. (Top Right) (b) The clustering indicator of the noise-containing 2D Gaussian mixture distribution dataset was calculated. The highest value of the decision indicator Di was used as the best clustering indicator. To save data points as much as possible, the smallest local density when the decision indicator reached the maximum was selected as the cutoff local density to denoise the dataset. Six clusters were calculated in the dataset. (Bottom Left) (c) Fig. 4c shows cutoff local densities and cutoff distances of each cluster respectively. There, the blue point is the initial cutoff distance. (Bottom Right) (d) Fig. 4d shows the final denoising clustering result calculated by BCALoD.

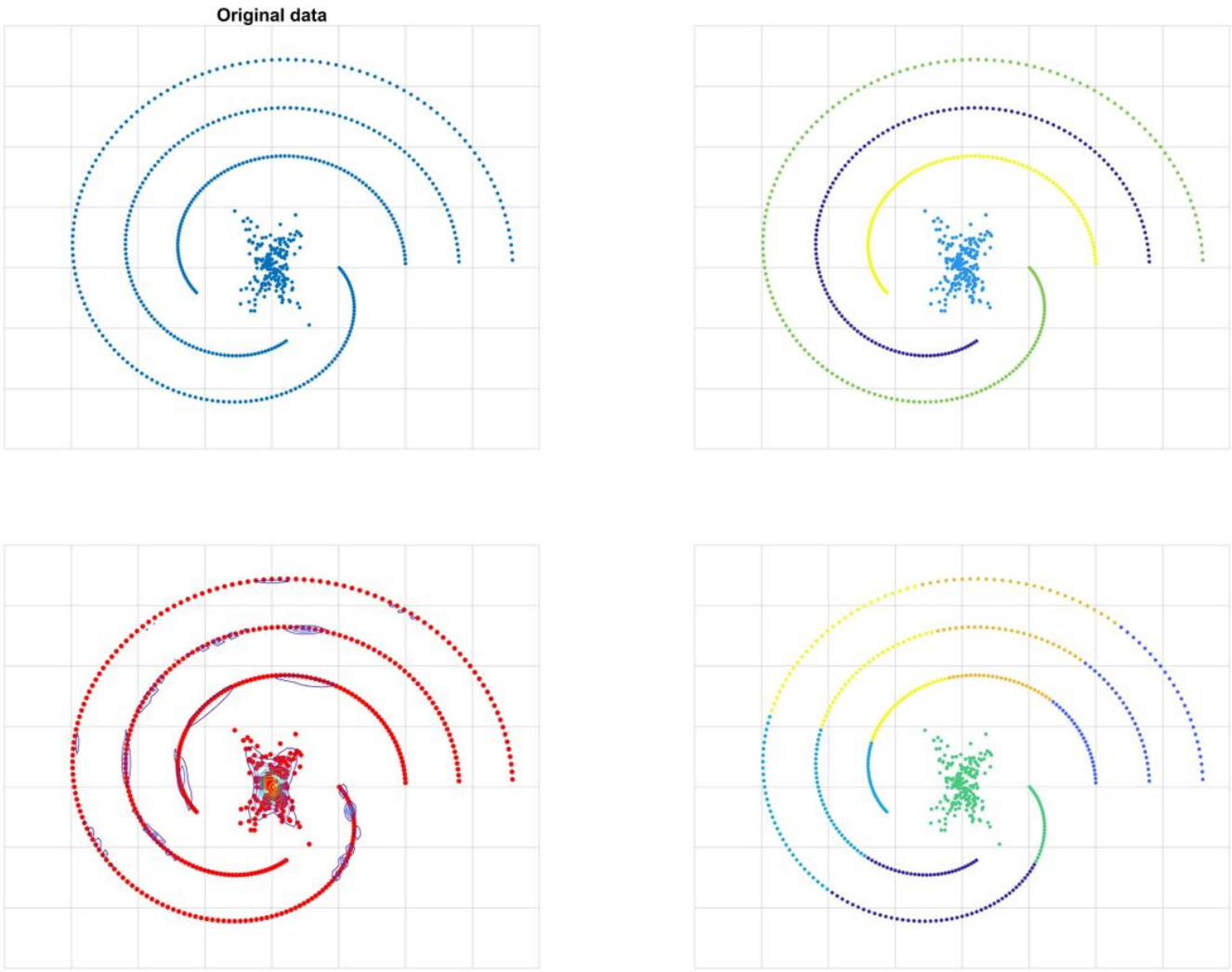

Figure 5

Combination dataset cluster analysis. (Top Left) (a) Data points comprised a Gaussian mixture distribution and spirals. (Top Right) (b) Accurate numbers of data points in different clusters were obtained by using the BCALoD algorithm. (Bottom Left and Right)(c-d) GMM and K-Mean were used to do cluster analysis for the combination dataset. For non-Gaussian spirals GMM could not calculate the expectation and variance of the dataset. Also, the K-mean method has poor clustering result for spirals data points. In this case, neither the silhouette coefficients nor the BIC could accurately provide the optimal solution of the cluster number, and determining the cluster number was difficult. However BCALoD cluster algorithm could effectively cluster the dataset.

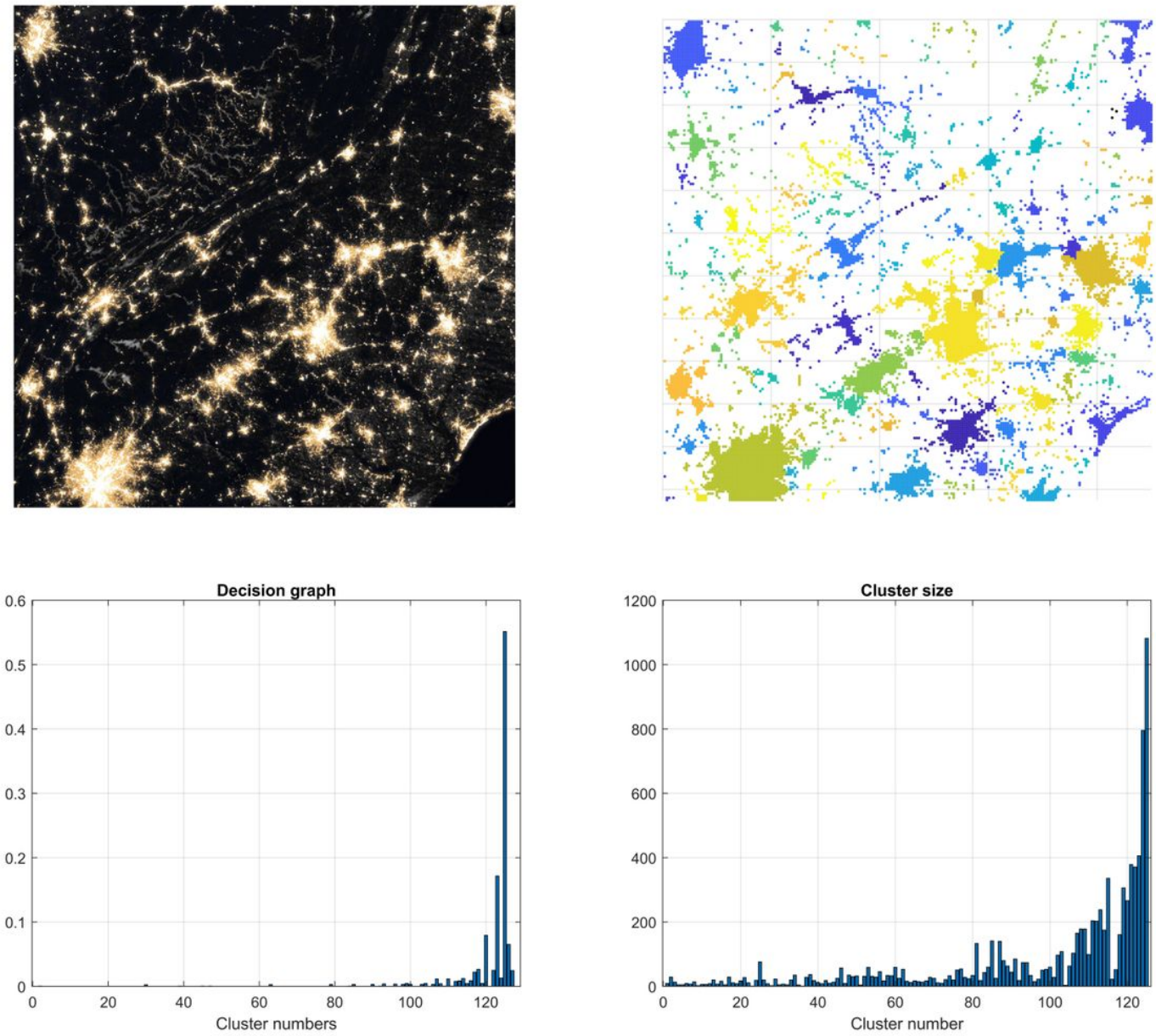

Figure 6

Clustering results of city lighting satellite image. (Top Left) (a) A NASA city light satellite image was selected as the dataset. This was also done to verify the BCALoD algorithm's ability to cluster datasets with multiple cluster centers and small clusters. The brightness of each point in the area was different, and the lights were widely distributed. There were many distribution centers in the image, and the distribution shapes were irregular. It would be very difficult to execute effective clustering calculations for these images using traditional methods. (Top Right)(b) The BCALoD algorithm was used to do cluster analysis on the images, and the clustering results obtained. (Bottom Left) (c) Decision diagram of satellite image BCALoD clustering. (Bottom Right)(d) Cluster size. Using a clustering calculation, the total value of light brightness in each region was obtained. The number of city light clusters reached 130, indicating that the BCALoD algorithm has good clustering capability for datasets of multiple clustering centers. The results also show that the BCALoD algorithm has accurate clustering performance for data

clusters of various shapes, and that small-cluster datasets can be clustered separately using this algorithm.