An Innovative High-Dimensional Clustering Algorithm and Its Application to Urban Stratification and Planning

Zhuohao Du
School of Computer Engineering Jiangsu University of technology, Changzhou, China
2584808199@qq.com

Yanshan Qian
School of Computer Engineering Jiangsu University of technology, Changzhou, China
1016449508@qq.com

Chuanan Li*
School of Computer Engineering Jiangsu University of technology, Changzhou, China
*Corresponding author: 1594779583@qq.com

Yunqi Hu
School of Computer Engineering Jiangsu University of technology, Changzhou, China

Junda Qiu
School of Computer Engineering Jiangsu University of technology, Changzhou, China

***Abstract*—With the acceleration of urbanisation, efficient and scientific urban planning and policy making is particularly important. However, traditional clustering algorithms tend to segment data based on a single objective or parameter in the process of urban stratification, which is easy to ignore the complexity and dynamics within the city when dealing with multidimensional and dynamically changing urban data, resulting in less accurate and comprehensive clustering results. This paper introduces an innovative clustering algorithm based on the Pareto optimal theory, which effectively improves the accuracy of clustering and the ability of multi-dimensional data processing by selecting the Steiner point as the clustering centre. Applying this algorithm to urban stratification and planning can not only reveal the development situation of the city in a more comprehensive way, but also provide the government with more scientific and detailed data support, which can promote the formulation of more reasonable and sustainable urban development strategies.**

Keywords-Urban planning; Pareto optimal theory; multi-dimensional data; clustering centre

I. INTRODUCTION

A. Background of the study

Processing and understanding high-dimensional datasets is becoming increasingly important in modern data science, especially in urban planning and analysis. However, traditional clustering algorithms, such as K-means and DBSCAN, face many challenges in dealing with these complex datasets, including dimensionality catastrophe, inefficiency in recognising complex data structures, and sensitivity to noise. In this study, an innovative high-dimensional clustering algorithm is proposed to replace the traditional clustering centres by introducing Pareto-optimal Steiner points. This approach not only improves the clustering quality, but also enhances the robustness of the algorithm in dealing with high-dimensional datasets with noise and outliers. Particularly in the field of urban stratification, this new algorithm demonstrates its great potential in urban planning, resource allocation and traffic management, which helps to understand and manage complex urban systems more effectively.

B. Related Work

Wang[1]and others proposed a clustering algorithm based on GMM modeling to adjust the clustering centers with very sensitive initial parameters or states. If the initial settings are incorrect, it may cause the algorithm to fall into a local optimum, and the algorithm has relatively high computational complexity and cannot be applied to large data sets.Hu[2]and others proposed the F2CAN clustering algorithm, which has the advantage of fast convergence, but it may still be sensitive to the initial parameters, causing the algorithm to fall into a local optimum, and the computational complexity of the algorithm is high.Song[3]and others proposed a weighted bidirectional k-means (WBKM) clustering algorithm, which causes the distances between data points to become relatively close to each other as the dimensionality of the data increases, thus making clustering more difficult and the clustering results inaccurate.Huang[4]et al. proposed a robust deep k-means model, and although this clustering algorithm can overcome the drawbacks of low-level features, the clustering centers are randomly generated and thus have a significant impact on the final clustering results.Yang[5]and others proposed the ISBFK-means algorithm based on influence space, which reduces the sensitivity to the selection of initial clustering centers, although the algorithm is overly complex and requires a lot of resources when dealing with large datasets.Wu[6]et al. proposed a hierarchical clustering algorithm called HCNN, which can handle heterogeneous and non-spherical datasets, but the algorithm is sensitive to outliers and the clustering centers are stochastic, resulting in less accurate clustering results.Ma[7]et al. proposed a new hierarchical clustering algorithm (CTCEHC) based on Minimum Spanning Tree (MST), which reduces the computational complexity but is not applicable to high-dimensional data as the distances between the data points may become very close in high-dimensional spaces.Guang[8]et al. proposed a fast hierarchical clustering of localized density

979-8-3503-6144-5/24/$31.00 ©2024 IEEE

peaks (FHC-LDP) by means of a correlation transfer method. This method can handle large-scale data efficiently, but this clustering result is highly affected by outliers.Xu[9]et al. proposed a graph theory based graph adaptive density peak clustering algorithm (GADPC), which can handle complex datasets better, but its robustness to noise and outliers is limited, and thus can lead to less accurate clustering results.Chen[10]et al. proposed the NQ-DBSCAN clustering algorithm, which can reduce the time complexity to O(n*log(n)), making it more efficient in dealing with large-scale data. However, the algorithm is less accurate when dealing with high dimensional and non-spherical data clusters.

II. KEY METHODS

A. The optiomal rally point in high dimensional coordinate system

Inside a closed box in three-dimensional space, there are $n(n \geq 3)$ points with weights whose positive weights are $\xi_i \in [0,1](1 \leq i \leq n)$ and $\sum_{i=1}^{n} \xi_i = 1$, respectively.If there is a point P^*, the Euclidean distance between it and other specific points satisfies the following conditions:

$$D = \min \sum_{i=1}^{n} \left| P^* P_i \right|$$
$$= \min\left(\sqrt{(x^*-a_1)^2+(y^*-b_1)^2+(z^*-c_1)^2}+\ldots+\sqrt{(x^*-a_n)^2+(y^*-b_n)^2+(z^*-c_n)^2}\right) \quad (1)$$

then P^* can be defined as the optimal rally point.(see Fig. 1)

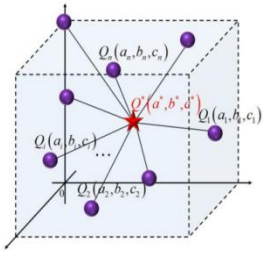

Figure 1. Optimal set point in three-dimensional space

B. Optimal marshalling algorithm

- Step 1: Determine an initial growth point, noted as $x^0 \in x$, and set step size λ (set to 1/1000 in this paper), where $x = (x_1, x_2, ..., x_m)$ is the set of vectors for box E. Set $X_{\min} = x^0$ and $F_{\min} = f(x^0)$, where $f(x^0)$ is the backlight function of x^0.

- Step 2: Using x^0 as the centre of your thinking, draw line segments parallel to the x-axis and y-axis, and extend them further by $a_1 \leq {x_1}^0 \leq b_1$, $a_2 \leq {x_2}^0 \leq b_2$,..., $a_m \leq {x_m}^0 \leq b_m$ to make a new branch.Find $S^0_{i_1 j_1} (1 \leq i_1 \leq m, 1 \leq j_1 \leq k_1)$ from the branch of λ, where $S^0_{i_1 j_1}$ is the growth point of j_1 from the branch of i_1.

- Step 3: Compare $f(S^0_{i_1 j_1})$ and $F_{\min}$. $X_{\min} = S^0_{i_1 j_1}$ holds if $f(S^0_{i_1 j_1}) \leq F_{\min}$ is satisfied; otherwise, $X_{\min}$ and $F_{\min}$ remain unchanged.

- Step 4: If $f(x^0) \leq f(S^0_{i_1 j_1})$ is satisfied, then the growth hormone concentration can be calculated as $C_{S^0_{i_1 j_1}} = 0$. Otherwise, bring in equation (1) to calculate $C_{S^0_{i_1 j_1}}$:

$$C_{S^0_{i_1 j_1}} = \frac{f(x^0) - f(S^0_{i_1 j_1})}{\sum_{i_1=1}^{m} \sum_{j_1=1}^{t_1-1} [f(x^0) - f(S^0_{i_1 j_1})]} \quad (2)$$

- Step 5: Calculate and record the growth hormone concentration at all growth points and plot the state of the morpholino concentration from 0 to 1. Choose a random number δ_0 in the interval,and if

$$\sum_{i_1=1}^{r_1} \sum_{j_1=1}^{t_1-1} C_{S^0_{i_1 j_1}} < \delta_0 < \sum_{i_1=1}^{r_1} \sum_{j_1=1}^{t_1} C_{S^0_{i_1 j_1}} \quad (3)$$

- Step 6: Repeat steps 2 through 5, stopping when the result is unchanged. At this point, $X^* = X_{\min}$ is the globally optimal solution; set it to $\dot{\mu} = x^*$ and end the calculation.

III. PRACTICAL EXAMPLE

A. Experimental data sources

Relevant data on the development of Chinese cities from 2015-2021 were collected from Wanfang database, and by reviewing relevant literature, this paper applies principal component analysis to downsize the multi-attribute features of cities, and the dataset is as shown in Table I (Selected cities in 2015 as an example):

TABLE I. RAW DATA FOR SELECTED CITIES IN 2015

City	primary industry	secondary industry	teriary industry	Total Retail Sales of Consumer Goods
Beijing	140.4	4419.8	20218.9	12271.86922
Tianjin	162.31	4489.59	6227.61	3963.2
Baoding	433.46	1645.67	1221.43	1509.3077
Cangzhou	321.3	1602.5	1316.8	1109.1272
Chengde	235.6	636.4	486.6	490.7935
Handan	402.8	1500.7	1241.9	1364.4978
Hengshui	168.9	563.1	488	609.0258

For a better study, we processed the raw data in the above table, firstly excluding the cities with large missing data, then deflating the data, and finally min-max normalising the data to the range of (0, 1), to obtain the experimental data as shown in the following Table II: (taking some cities as an example)

TABLE II. EXPERIMENTAL DATA FOR EACH YEAR OF TREATMENT

City	2018			
	primary industry	secondary industry	teriary industry	Total Retail Sales of Consumer Goods
Beijing	0.0922706	0.5285797	1	1
Tianjin	0.1292816	0.4774684	0.3105605	0.301106293
Baoding	0.3030157	0.1637303	0.0573056	0.096100599
Cangzhou	0.2034052	0.1838047	0.0625672	0.06551246
Chengde	0.1811958	0.0610837	0.0226316	0.028578097
City	2019			
	primary industry	secondary industry	teriary industry	Total Retail Sales of Consumer Goods
Beijing	0.0848315	0.5270540	1	0.969531305
Tianjin	0.1246221	0.4648694	0.3016243	0.283272088
Baoding	0.2530867	0.1279165	0.0535909	0.096365485
Cangzhou	0.1978018	0.1496007	0.0634892	0.065753977
Chengde	0.1916914	0.0486175	0.0217889	0.028683858
City	2020			
	primary industry	secondary industry	teriary industry	Total Retail Sales of Consumer Goods
Beijing	0.0709405	0.5431713	1	0.950437477
Tianjin	0.1168443	0.4736792	0.2987242	0.264725482
Baoding	0.2835843	0.1235386	0.0650060	0.097744601
Cangzhou	0.1861399	0.1343468	0.0600858	0.06660106
Chengde	0.1896180	0.0434742	0.0201698	0.028775195
City	2021			
	primary industry	secondary industry	teriary industry	Total Retail Sales of Consumer Goods
Beijing	0.0572258	0.5514906	1	0.860652333
Tianjin	0.1139733	0.4715465	0.2931090	0.223462973
Baoding	0.2701479	0.1246929	0.0683111	0.083794378
Cangzhou	0.1721507	0.1354696	0.0619873	0.064320576
Chengde	0.1840538	0.0449970	0.0208378	0.027982473

B. Experimental results and analysis

This study employed an optimised set-point model and a simulated plant growth algorithm to analyse key features of 297 Chinese cities from 2015 and 2017. We initialised predefined points as seed points and set 1000 iterations for adequate algorithm convergence. The first seed point was used as the starting value for both the best configuration point, best_node, and the objective function, best_obj_function, of the optimal set point.

During the experiment, we set a specific growth step step_length, which was determined after a series of experimental adjustments. In each iteration, we calculate the objective function value of each seed point and select the seed point with the smallest objective function value as the current optimal configuration point. In order to extend the theory to the fourth dimension, we adopt an innovative approach to demonstrate the characteristics of the fourth dimension by the size of the data points.

Through the comprehensive analysis of the 2015 and 2017 data, as shown in the results of Figure 2. A clustering center with coordinates (0.149352, 0.079257, 0.026955, 0.045817) was identified in this study.As shown in Figure 3, this clustering centre not only reflects the spatial distribution of the data, but also provides insights into the characteristics of the multidimensional data. The methodology and findings of this study provide new perspectives for understanding and analysing high-dimensional datasets and provide valuable references for research in related fields.

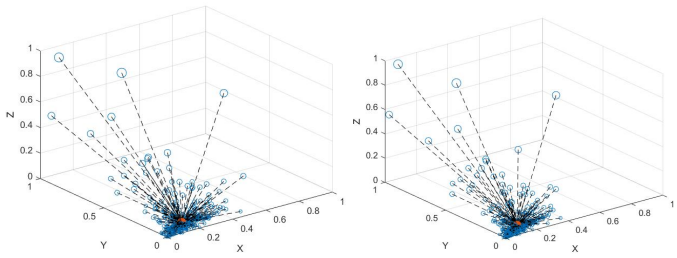

Figure 2. Optimal assembly point selection map

This study enhances the k-means clustering algorithm by substituting its random cluster centers with optimal nodes derived from 2015 and 2017 Chinese city data. We analyze data from 297 cities spanning 2016 to 2021, as shown in Figures 3 and 4, focusing on the annual distance of each city from the optimal clustering point. By aggregating past data, a refined threshold delineation strategy was developed, improving the precision of cluster analysis.

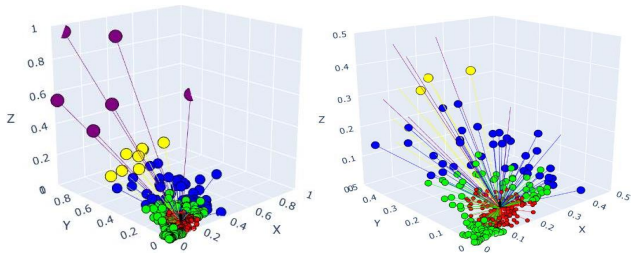

Figure 3. Overall (left) and localised (right) map of urban stratification (2016)

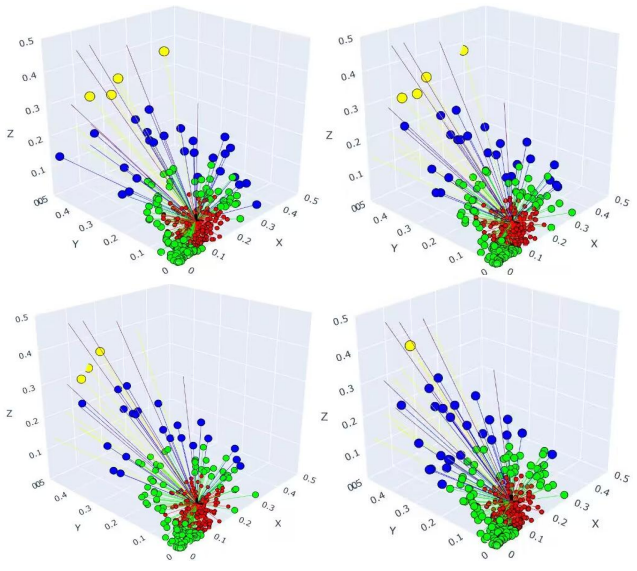

Figure 4. Urban stratification map, 2018-2021

Five thresholds of [0.1,0.2,0.45,0.7,1.5] were set by calculating the distance from each data point to the centre of

clustering in 2016. In order to improve the accuracy, this paper adjusts the parameters with the clustering from 2018 to 2021, and finally sets the thresholds to[0.0958,0.213,0.556,0.721,1.50] to better judge the current status of the city. In this paper, the threshold is applied to 2021, which can better reflect the current status of the city.

The following paper carries out the traditional clustering algorithm for comparison experiments, selecting 2018 and 2021 data for experiments, and the results are shown below:

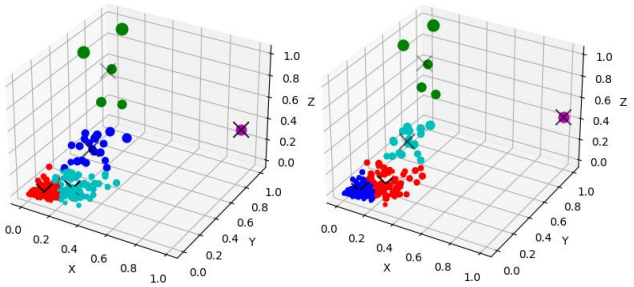

Figure 5. Traditional clustering results map, 2018 (left) 2021 (right)

According to the comparison experiments, it can be seen from Fig. 5 that the KNN clustering algorithm converges faster and more accurately than the traditional methods by randomly selecting the clustering centres and forming clusters, which are poorly clustered according to the comparison of the data over the years (in this paper, we only put the years 2018 and 2021). At the same time, in this paper, we expand the selection of the optimal set of nodes as well as clustering to high dimensions, which makes the final results more accurate.

IV. Discussion

A. Analyze

In this study, we extend the key clustering method from three-dimensional space to four-dimensional space for the first time, demonstrating a novel method based on a simulated plant growth algorithm for accurately determining the Pareto-optimal Steiner points as the optimal set points for clustering. The enhancement of this algorithm to four-dimensional space not only broadens the scope of its application, but also significantly improves the accuracy of the clustering results. Compared with the traditional algorithm of randomly selecting the clustering centre, the method in this study effectively overcomes the uncertainty and limitation caused by randomness by using the precisely determined optimal set point as the clustering centre. Our method not only changes the traditional clustering pattern based on cluster formation, but also provides a new perspective and technical path for accurate clustering of complex datasets through the precise selection of clustering centres and the fine adjustment of clustering thresholds.

The focus of the discussion in this paper is to explore the implications and significance of applying the Pareto-optimal Steiner points from plant growth algorithms to clustering centres, in particular the innovation of extending the method from its application in three dimensions to four dimensions. The method proposed in this study represents a shift towards a more controlled and accurate clustering technique than traditional clustering methods that rely on random selection of clustering centres and are based on cluster formation. By determining the clustering centres based on the optimal set of nodes, our method not only avoids the uncertainty associated with randomness, but also provides a more precise and purposeful solution for dealing with the clustering task in complex data analysis through fine parameter tuning and threshold setting

B. Correlate

1) Linkages between results and practice

The official annual city stratification for 2021 is as follows: Tier 1 cities are Shanghai, Beijing, Shenzhen, Guangzhou, etc. There are a total of 15 new Tier 1 cities, e.g. Chengdu, Hangzhou, Chongqing, etc., a total of 30 Tier 2 cities, e.g. Hefei, Kunming, Wuxi, etc., 70 Tier 3 cities, e.g. Langfang, Shantou, Baoding, etc., a total of 90 Tier 4 cities, and a total of 128 Tier 5 cities.

The results of this paper's analysis of the state of the city in 2021 are shown in Table III:

TABLE III. 2021 Table Reflecting Partial Results Of Urban Clustering

City	Beijing	Shanghai	Baoding	Wuxi	Chengdu
X	0.0551	0.0491	0.2124	0.0650	0.3011
Y	0.6338	1.0	0.1156	0.5849	0.5327
Z	1.0	0.9626	0.0568	0.2155	0.4002
W	0.8218	1.0	0.0912	0.1803	0.5102
Distance	1.1239	1.3165	0.0787	0.5462	0.6066
Category	5	5	1	3	4
City	Yingkou	Hefei	Shantou	Shenzhen	Chongqing
X	0.0583	0.1802	0.0622	0.0109	1.0
Y	0.0527	0.3625	0.1209	0.9903	0.9768
Z	0.0169	0.2072	0.0395	0.5856	0.4480
W	0.0217	0.2805	0.0803	0.5239	0.7718
Distance	0.0952	0.3371	0.0973	1.0776	1.3063
Category	1	3	2	5	5

The experimental results in 2021 are basically consistent with the actual results with high accuracy. Meanwhile, the accuracy and stability of our algorithm on high-dimensional datasets are better than that of traditional methods. By using the simulated plant growth algorithm to find Pareto-optimal Steiner points as clustering centres in a four-dimensional space, we effectively overcome the limitations of random centre selection in traditional algorithms. This approach improves the clustering quality and provides new insights for analysing high-dimensional data, especially for complex urban planning.

2) Comparison with traditional clustering algorithms

For the traditional k-means clustering algorithm, we did a comparison experiment, still applying the experimental data of 2021, and extended the k-means clustering algorithm to a four-dimensional space by the same approach, indicating the fourth dimension by the size of the data points, and the experimental results are shown in Fig. 6:

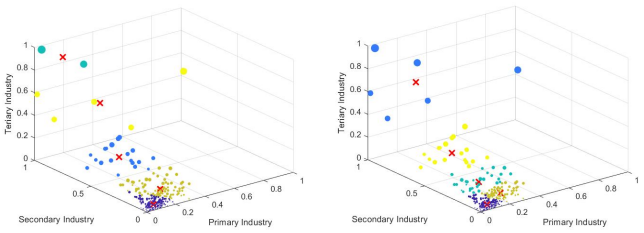

Figure 6. K-means clustering result graph

Also we calculated the profile factor of k-means and the results are:Average profile factor: 0.63367, Average profile factor: 0.60526 and the profile factor plot is shown in Fig. 7:

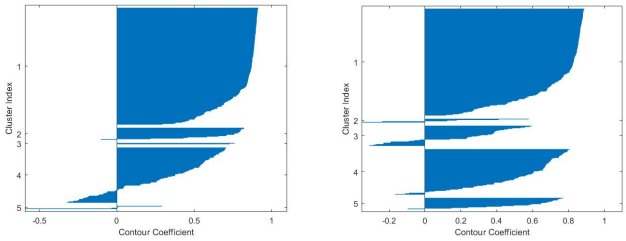

Figure 7. Contour Coefficient Chart

As can be seen from Figures 6 and 7: traditional K-means clustering relies heavily on random initialisation of the cluster centres, which can lead to variations in results between runs. However, our approach systematically determines the clustering centres through the optimal set points identified by the algorithm. This not only provides a more stable and repeatable clustering process, but also allows for more precise control of the clustering results through manual parameter tuning and threshold setting.

Furthermore, whereas traditional K-means clustering forms clusters based on the proximity of data points to these randomly selected centres, our approach introduces a more considered and theoretically informed approach. By clustering around these optimally identified nodes, our approach goes beyond grouping based solely on proximity and provides a more subtle and potentially more meaningful clustering mechanism that is guided by the intrinsic patterns implied by the principle of Pareto optimality and plant growth simulation algorithms.

C. Reflect and Expand

In this study, we have advanced the traditional K-means algorithm by identifying Pareto optimal Steiner points as clustering centers using a plant growth simulation algorithm, effectively enhancing the algorithm's extension into four-dimensional space. This enhancement has notably improved the accuracy and dimensional coverage of clustering. While precise parameter tuning and threshold setting allow for more detailed clustering control, they also present challenges in determining optimal parameter configurations, particularly in applications involving high-dimensional data. Future work will focus on streamlining the parameter adjustment process to enhance the operability and applicability of the method.

V. CONCLUSIONS

This paper enhances the clustering algorithm for high-dimensional spaces by visualising the first three dimensions and representing the fourth dimension as the data point size. We compare the optimal set of point clustering centres with the traditional clustering results. As shown in Figs. 5 and 6, the random clustering centres of KNN and K-means produced unsatisfactory results. Our improved algorithm not only handles high-dimensional data, but also solves the problem of random centre selection with faster convergence and higher accuracy, which is confirmed by years of comparative data analysis.

REFERENCES

[1] Wang,J.H.,Jiang,J.M.(2021)Unsupervised deep clustering via adaptive GMM modeling and optimization, Neurocomputing 433 199–211.https://doi.org/10.1016/j.neucom.2020.12.082.

[2] Hu,L.,Pan,X.Y,Tang,Z.H.,Luo,X.(2022)A fast fuzzy clustering algorithm for complex networks via a generalized momentum method, IEEE Trans.Fuzzy Syst.30 (9) 3473–3485.https://doi.org/10.1109/TFUZZ.2021.3117442.

[3] Song,K.,Yao,X.W.,Nie,F.P.,Li,X.L.,Xu,M.L.(2021)Weighted bilateral k-means algorithm for fast co-clustering and fast spectral clustering, Pattern Recognit. 109 107560. https://doi.org/10.1016/j.patcog.2020.107560.

[4] Huang,S.D.,Kang,Z.,Xu,Z.L.,Liu,Q.H.(2021)Robust deep k-means: An effective and simple method for data clustering, Pattern Recognit. 117 107996.https://doi.org/10.1016/j.patcog.2021.107996.

[5] Yang,Y.Q.,Cai,J.H.,Yang,H.F.,Li,Y.T.,Zhao,X.J.(2022)Isbfk-means:A new clustering algorithm based on influence space, Expert Syst. Appl. 201 117018.https://doi.org/10.1016/j.eswa.2022.117018.

[6] Wu,C.R.,Peng,Q.L.,Lee,J.,Leibnitz,K.J.,Xia,Y.N.(2021)Effective hierarchical clustering based on structural similarities in nearest neighbor graphs, Knowl.-Based Syst. 228 107295.https://doi.org/10.1016/j.knosys.2021.107295.

[7] Ma,Y.,Lin,H.R.,Wang,Y.,Huang,H.,He,X.F.(2021)A multi-stage hierarchical clustering algorithm based on centroid of tree and cut edge constraint, Inform. Sci. 557 194–219.https://doi.org/10.1016/j.ins.2020.12.016.

[8] Guan,J.Y.,Li,S.,He,X.X.,Zhu,J.H.,Chen,J.J.(2021)Fast hierarchical clustering of local density peaks via an association degree transfer method, Neurocomputing 455 401–418.https://doi.org/10.1016/j.neucom.2021.05.071.

[9] Xu,T.,Jiang,J.(2022)A graph adaptive density peaks clustering algorithm for automatic centroid selection and effective aggregation, Expert Syst. Appl. 195 116539.https://doi.org/10.1016/j.eswa.2022.116539.

[10] Chen,Y.,Tang,S.,Bouguila,N.,Wang,C.,Du,J.,Li,H.(2018),A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data, Pattern Recognit.83 375–387.https://doi.org/10.1016/j.patcog.2018.05.030.