ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20171121052315
Original scientific paper

# Improved Density Peak Clustering Algorithm Based on Choosing Strategy Automatically for Cut-off Distance and Cluster Centre

Limin WANG, Mingyang LI, Xuming HAN, Ruihong ZHOU, Kaiyue ZHENG, Meihan LIU

**Abstract:** Due to the defect of quick search density peak clustering algorithm required an artificial attempt to determine the cut-off distance and circle the clustering centres, density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster center (CSA-DP) is proposed. The algorithm introduces the improved idea of determining cut-off distance and clustering centres, according to the approximate distance that maximum density sample point to minimum density sample point and the variation of similarity between the points which may be clustering centres. First, obtaining the sample point density according to the k-nearest neighbour samples and tapping the sample sorting of the distance to the maximum density point; then finding the turning position of density trends and determining the cut-off distance on the basis of the turning position; finally, in view of the density peak clustering algorithm, finding the data points which may be the centres of the cluster, comparing the similarity between them and determining the final clustering centres. The simulation results show that the improved algorithm proposed in this paper can automatically determine the cut-off distance, circle the centres, and make the clustering results become more accurate. In the end, this paper makes an empirical analysis on the stock of 147 bio pharmaceutical listed companies by using the improved algorithm, which provides a reliable basis for the classification and evaluation of listed companies. It has a wide range of applicability.

**Keywords:** clustering center; cut-off distance; Density Peak Clustering Algorithm; maximum density; similarity

## 1 INTRODUCTION

With the maturity of big data era and artificial intelligence, human civilization has entered a new era of data-intensive computing, where data has become an extremely important asset. Due to constant proliferation of data recently, effective analyzing and utilizing vast amounts of raw data and retrieving valuable information is becoming the focus of many researchers and common subjects [1]. The nature of data mining is to identify potential rules and valuable information from massive raw data, through a series of scientific analyses and processing [2]. As an important branch of data mining technology, cluster analysis, without any prior information provided, extracts valuable information by exploring similar relationship between internal data structure information and data points from huge volumes of data [3]. With the rapid development of Chinese stock market and the explosive growth of data, a large amount of valuable information is hidden behind the massive data, but the information is often difficult to detect with the naked eye and experience [4]. Returns and risk in the stock market has been one of the hot issues that the researchers are concerned about and how to avoid the risk to get the greatest benefit has come to the foreground. The stock market is a complex system which has a wide variety of complex structures and various factors affect each other [5]. For a large number of stocks, how to judge whether a stock has investment value is the key to solve the problem [6]. In this complex financial situation, the application of data mining methods in the analysis of stock market becomes more and more significant. It uses the relevant data and data analysis tools to discover the relationship between data. The results can be used to predict and make reasonable decisions [7, 8]. Clustering analysis is an important branch of data mining, the purpose is to study the similarity between data and divide similar data into the same class [9]. Therefore, cluster analysis can be used to measure the degree of similarity between stocks according to the value of each stock by clustering similar value of the stock into a class, so as to grasp the overall trend of the stock, to determine the potential value of the stock.

Based on the clustering "Things of one kind come together" principle, the data set is divided into several groups or clusters, so that the similarities of the same classes of objects are higher than those of different classes [10]. Because of the complexity and diversity of data, many scholars put forward a large number of outstanding clustering algorithms based on the specific problems in different areas, but all have the disadvantage of non-conformance [11]. Selection of clustering algorithm, combined with its own characteristics and advantages of the algorithm [12], needs for comprehensive analysis of the size and structural data characteristics. Usually, clustering algorithms can be divided into the following categories: partition-based method, hierarchical-based method, density-based method, grid-based method, model-based method [13, 14]. Partition-based method is one of the most widely used clustering algorithms whose core idea is: the set of data containing $n$ samples is a predefined cluster number $k$, through continuous optimization of some target classification criterions while the error value convergence of the objective function, the end of iteration, the entire set of data is divided into $k$ clusters [15]. Unlike partition-based method, the main idea of hierarchical-based method is a clustering tree including data points [16]. Clustering results meet certain requirements through repeated decomposition or aggregation operations. Since partition-based method and hierarchical-based method often only found the convex clusters, in order to compensate for this deficiency, density-based method was developed to find all kinds of arbitrary shape clusters [17]. The core idea of such algorithm is: each object cluster in the whole sample is a group composed of dense sample points. These points are divided by low-density area. The tenacity of the algorithm is to filter the low density region and find dense sample points [18, 19].

In June 2014, Alex and Alessandro offered a reference of density-based methods, called clustering by fast search and finding density peaks, DP. The algorithm can quickly find the density peak of any shape data (namely cluster center), efficiently perform sample point distribution and eliminate outliers [20]. Nevertheless, DP is the ideal clustering algorithm, but there are also shortcomings, such as the sample density measurement. There is no uniform density measurement criterion. Different scale data sets use different sample density measurement criteria. When the sample scale is small, the cut-off distance $d_c$ which is selected subjectively has greater impact through sample density measure on clustering results [21]. After drawing a decision graph based on density and distance, it is needed to circle the relatively large density and distance of data points as an artificial cluster center [22]. Aiming at the deficiency of DP algorithm, this paper puts forward the optimization scheme: the method treats the distance from the maximum density point to the minimum density point as cut-off approximate value, regards the data points that the product of density and distance is bigger and their similarity between them is lower as clustering centers, called density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster center (CSA-DP). The specific steps are as follows: firstly, obtaining the density of sample point according to the k-nearest neighbor samples and tapping the samples sort of the distance to density maximum point; secondly, finding the turning point of density trend and, determine the cut-off distance according to the position of the turning point; thirdly, to calculate the product of the density and distance of the data points, and sort the product from large to small, find out the data points with the product greater than the average product, and calculate the similarity between these points. Finally, circle clustering centers according to the similarity of the changes, assign other data points, and get the clustering results. This algorithm avoids the disadvantage, which needs to determine the cut-off distance and circle the clustering centers manually, it also makes the results more accurate than the original clustering algorithm.

In the part 2 of this paper, we introduce the density peak clustering algorithm in detail, and point out the deficiencies of the algorithm. In the part 3, we analyze the rationality of the improved algorithm in this paper, and illustrate the algorithm with the detailed procedure. In the part 4， through UCI data sets, we validate the clustering results of the improved algorithm and compare with other original algorithms. In the part 5, we apply the improved algorithm to the practical problems, analyze and illustrate the stock data. In the last part, we make the corresponding summary and clarify the direction of the future research.

## 2 DENSITY PEAK CLUSTERING ALGORITHM

Density peak clustering algorithm can discover automatically the cluster centers of sample data sets, and realizes high clustering of arbitrary shape sample data sets. The core idea of the clustering algorithm is to describe cluster centers on the features [23]. The thought of the algorithm about cluster centers has two characteristics: 1) The density of itself is very big, that it is surrounded by its neighbors whose density are not more than itself; 2) The "distance" between other data points with greater density are relatively large, in other words, distance of the two class centers is relatively far [24]. Density peak clustering algorithm steps as follows: first of all, determine the cluster centers according to the characteristics of the class center; then, classify the other data points to the classification of the nearest point, whose density is higher than oneself. DP algorithm introduces the local density of sample *i* and the distance $\delta_i$ of sample *j* with bigger local density than *i*. Its definition can be shown by Eqs. (1) and (3).

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c) \tag{1}$$

Where $d_{ij}$ is the Euclidean distance of sample *i* and *j*, $d_c$>0 is the cut-off distance. Among them, function $\chi(x)$ is

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{2}$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

For the sample with the biggest local density $\rho_i$,

$$\delta_i = \max{}_j \, d_{ij} \tag{4}$$

When the local density of $x_i$ is maximum, $\delta_i$ represents the distance between the point $x_i$ and the furthest point from $x_i$. Otherwise, $\delta_i$ is indicated that the distance between the data point and the $x_i$, which is the nearest of all the local points with bigger local density [25].

The Eq. (1) shows that the sample local density for DP is influenced by the cut-off distance. At the same time, this algorithm points out that when the number of data set is large, the clustering result of DP algorithm is less influenced by truncation distance, otherwise it is greatly influenced by truncation distance [26]. In order to avoid the influence of truncation distance for the local density of sample and even clustering result, when the data set samples are smaller, the DP algorithm uses the kernel algorithm to calculated sample density [27]. Under this situation, the expression can be expressed as follows:

$$\rho_i = \sum_{j \neq i} \exp\left[ -\left( \frac{d_{ij}}{d_c} \right)^2 \right] \tag{5}$$

Literature [20] points out, according to the above formulas, decision graph can be drawn correspondingly by the local density $\rho_i$ and distance $\delta_i$. According to the characteristics of cluster centers, circle the points with the relatively large local density and distance as the centers in the decision graph. However, in the above identified cluster centers, it uses qualitative analysis rather than quantitative analysis, and contains the subjective factors. Different people may get different results in the same decision graph, for example, some people may think that these are the cluster centers, and some people think that those are the cluster centers. As shown in the following Fig. 1.

Select the larger samples $\rho$ and $\delta$ as the class cluster center. For the rest of the sample *j*, it is classified by DP algorithm as the class cluster of samples, whose density is larger than $j$ and the distance is the nearest to the $j$ [28]. It needs one step to distribute the remaining sample. One-step distribution strategy makes the DP algorithm effective. By constructing the decision graph of sample distance and sample density, DP algorithm makes any dimensional data set class cluster center can be display in 2D plane, realizing the clustering analysis of arbitrary dimensional data [29]. However, when the scale of every cluster is small, which make the "ideal" class cluster and data set have equivalent scale, the sparse of samples will lead the density peak point dim. When it is hard to find the density peak points using the density estimation method, the decision graph constructed by $\gamma$ ($\gamma_i = \rho_i{\cdot}\delta_i$) curve is usually adopted to choose the class cluster center (density peak point), and its specific schematic diagram is shown in Fig. 2.

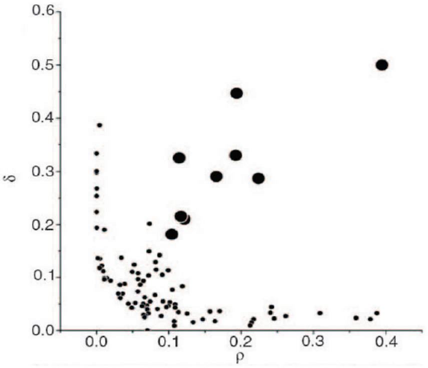


**Figure 1** Decision graph

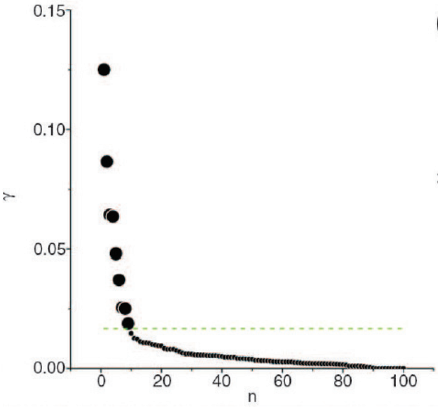


**Figure 2** The value of $\gamma$ in decreasing order

Among them, the horizontal axis is the index set and the vertical axis is the value $\gamma$. The bigger $\gamma$ is, the more likely $x_i$ is to be the cluster center. Therefore, only need to do $\gamma$ in descending order, take some data points as cluster centers from front to back. $\gamma$ of non-cluster center is relatively smooth, from the non-cluster centers to the cluster centers, the value $\gamma$ has obvious jump phenomenon.

## 3 DENSITY PEAK CLUSTERING ALGORITHM BASED ON CHOOSING STRATEGY AUTOMATICALLY FOR CUT-OFF DISTANCE AND CLUSTER CENTER

Literature [20] gave no selection methods of the cut-off distance, but only reference suggestions: choose a cut-off distance, which makes the average neighbor number of each data point is 1-2 % of the total number, and the "neighbor" here refers to that the distance between the point and the neighbor is not more than the cut-off distance in sense, obviously this statement is subjective and one-sided.

Through a lot of experiments, for data sets with relatively large amounts of data, the accuracy of the parameters between 1 % and 2 % is relatively high, but for small data sets, a big value of the cut-off distance can probably get better results. The selection of parameter $d_c$ determines the success or failure of clustering algorithm in a sense, which can't get too big or too small. If the cut-off distance is too large, that will make the local density value of each data point too large, and the differentiation is not high. If $d_c$ is too small, the same cluster could be split into multiple clusters. If select the cut-off distance directly, it will depend on the specific issues, even for different problems, reasonable magnitude of the cut-off distance is likely to have very big difference.

Based on the investigation to the DP clustering algorithm, according to the applicability of density clustering algorithm has the characteristics of density changes, looking for the high density area which is separated by low-density areas in data set, this paper puts forward an improved algorithm with reference for the cut-off distance from maximum density of sample points to minimum sample points. Let densest sample points as a starting point to investigate the density distribution of the data points.

In clustering that is based on density, the data points with high density, density of the data points between each cluster is very low, and the density of the biggest data point must be within a cluster, and the distance between other data points with cluster is less than the distance to data points in other cluster. Let the maximum density point as $M$, according to the distance between all data points and $M$ too reorder the data sets, we can get

$$order = \{m_1, m_2, ..., m_n \mid dist(M, m_1) \leq \\ \leq dist(M, m_2) \leq ... \leq dist(M, m_n)\} \quad (6)$$

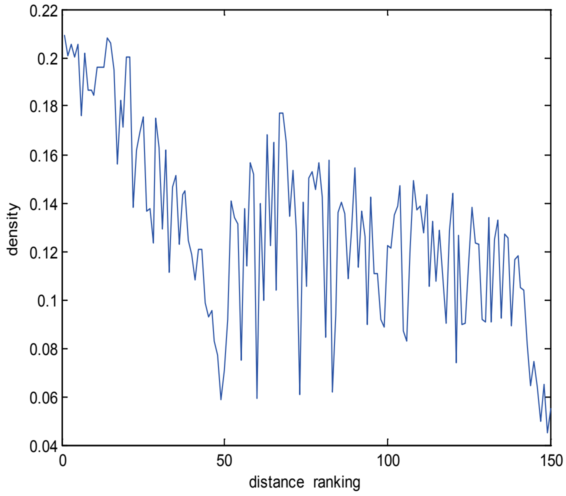


**Figure 3** The distance to the maximum density of points in increasing order

To avoid the influence of cutting distance to the local density calculation results, we firstly use the Euclidean distance between sample point and its $K$ nearest neighbor points and the reciprocal of density to calculate density of each data point. Using Iris data set as an example, we can get the order and density of data points drawing as follows in Fig. 3.

We can see from the Fig. 3, with the increase of distance between data point and $M$, density curve has an obvious trough at position 50, this is because the points in the same cluster of $M$ is close to $M$, so the density in the first half of the curve is higher. The data points located in the 50 or so between is in the area between each cluster and the density is very low. And data points which larger than 50 are in other cluster area, the density begin to rise again [30]. For small data sets, get the integer of percentage which is the ratio of trough location of total data points as a reference value of truncation distance $d_c$.

After determining the truncation distance $d_c$, the density $\rho$ and the distance $\delta$ are calculated, the DP clustering algorithm requires to circle the points with larger density and distance as the clustering centers artificially. Through the investigation of DP clustering algorithm, the bigger product $\gamma$ of the density and distance is, the more likely it is to be the clustering center. The similarity between clustering centers is always lower, however, the similarity between the clustering center and the same cluster is higher. In order to avoid the trouble of artificial marks, this paper presents a method of determining the clustering centers, according to the change of the similarity between the data points with larger $\gamma$. First off, find out the data points with the product $\gamma$ of density and distance greater than the average product, then sort these points by the product $\gamma$ from large to small, and get the similarity between them. With the decrease of product $\gamma$, there will always be a sudden change of similarity, take the data points with larger $\gamma$ and smaller similarity ahead as the clustering centers. Using Compound data set as an example, the results of the data are shown in Tab. 1, the first and third lines represent the comparison of points between the top $i$ and the top $i$+1 of $\gamma$, the second and fourth lines represent the Euclidean distance between the two points.

**Table 1** Experimental results on compound data set

| Data point | $i$ = 1 | $i$ = 2 | $i$ = 3 | $i$ = 4 | $i$ = 5 |
|---|---|---|---|---|---|
| Euclidean distance | 16.987 | 5.7881 | 11.871 | 8.2123 | 14.026 |
| Data point | $i$ = 6 | $i$ = 7 | $i$ = 8 | $i$ = 9 | $i$ = 10 |
| Euclidean distance | 3.7165 | 18.916 | 2.2192 | 0.8062 | 1.1402 |

As can be seen from the Tab. 1, when $i$ is less than 6, with the decrease of $\gamma$, the Euclidean distance between data points is greater than a certain value, namely, similarity is less than a certain value, Euclidean distance is always maintained a larger value, and similarity is smaller. When $i$ = 6, the Euclidean distance between data points with the sixth and seventh $\gamma$ is 3.7165, smaller than all the previous values. When $i$ = 8, 9, 10, Euclidean distances become smaller, namely, similarity is relatively high. Accordingly, for Compound data set, select the data points with $\gamma$ of the first six largest as the clustering centers. This is exactly the same as the number of classes in the data set.

The process of the proposed algorithm is shown below.

| **Input:** Data set<br>**Output:** Cluster sets $C = \{C_1, C_2,..., C_k\}$, evaluating indicator | |
|---|---|
| **Step 1** | Get the Euclidean distance between point and point. |
| **Step 2** | Find out the total distance from the each data point to $K$ nearest neighbors, and regard the reciprocal of the total distance as the density. |
| **Step 3** | Get the maximum density point, and find the distances from all other points to the maximum density, do the ascending order. |
| **Step 4** | Draw coordinate graph of distance sorting as abscissa and density as ordinate, find out the trough point. |
| **Step 5** | Make the trough point position ratio of the total number as an approximate value of cut-off distance. |
| **Step 6** | Calculate the true density of the original DP algorithm by using the cut-off distance. |
| **Step 7** | Calculate the product $\gamma$ of density and distance, and sort the data points according to $\gamma$ from large too small. |
| **Step 8** | Find out the data points with the product greater than the average product, and calculate the Euclidean distance between the points with the top $i$ and the top $i$+1 of $\gamma$. ($i$=1, 2, 3…) |
| **Step 9** | According to the position of the European distance suddenly reduced, take data points with the first $K$ largest $\gamma$ for the clustering centers. ($K$ is the number of clusters) |
| **Step 10** | Output clustering results, and get the graph of $\gamma$ in decreasing order. |

## 4 THE ANALYSIS OF SIMULATION EXPERIMENT RESULTS

### 4.1 The Evaluation Index of Cluster Effectiveness

According to the clustering number and the correct clustering result of the known data, the method will be the result of the $Q$ clustering algorithm comparing with the known structure of $P$ in advance, known as the external evaluation method. This is a kind of based on the known structure of $P$ in advance the clustering result of evaluation method [31].

For the data set two entities of $p$ and $q$, existing in p and $q$ of the following four relationships:

(1) $p$ and $q$ belong to the same class in the C, and belong to the same partition in the $p$;

(2) $p$ and $q$ belong to the same class in the C, but do not belong to the same partition in the $p$;

(3) $p$ and $q$ do not belong to the same class in the C, but belong to the same partition in the $p$;

(4) $p$ and $q$ do not belong to the same class in the C, and do not belong to the same partition in the p.

Set $a$, $b$, $c$, $d$ entity logarithm to meet the 4 kinds of circumstances respectively, M is the entity logarithmic sum of the data set, there is the following:

$$M = a + b + c + d = \frac{N(N-1)}{2} \tag{7}$$

Among them, N is the number of entities in the data.

On the basis of the above definition, the similarity of $C$ and $P$ can use the following effectiveness evaluation method definition:

- Rand index

$$R = \frac{a+b}{M} \tag{8}$$

- Jaccard coefficient

$$J = \frac{a}{(a+b+c)} \tag{9}$$

The two effective evaluation methods above belong to [0, 1], the greater the value, the higher the similarity between $C$ and $P$.

- F-measure index

F-measure index, which is also called F-Score, is a commonly comprehensive evaluation index in information retrieval field. It groups the recall and precision in information retrieval field to comment the cluster [32]. The definition of the recall and precision for cluster $j$ and its related classify $i$ are as following expressions:

$$P = precision\ (i, j) = \frac{N_{ij}}{N_i} \tag{10}$$

$$R = recall\ (i, j) = \frac{N_{ij}}{N_j} \tag{11}$$

Where $N_{ij}$ represents the number of cluster $j$ that belongs to the classify $i$, $N_j$ represents the number in cluster $j$, $N_i$ represents the number of classify $i$. The definition of F-measure for the classify $i$ is as follows:

$$F\left(i\right) = (\partial^2 + 1)PR/\partial^2\left(P+R\right) \tag{12}$$

Where $\partial$ is a parameter, and in general $\partial = 1$.

- Silhouette index

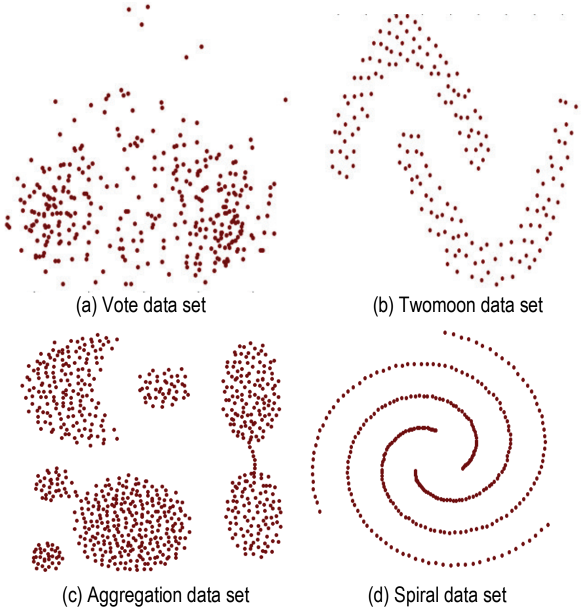
Assume a data set with $n$ samples be divided into $k$ clusters $C_i(i = 1, 2, \ldots, k)$, $a(t)$ is the average dissimilarity of sample $t$ in $C_j$ to all other samples in $C_j$, $d(t, C_i)$ is the average dissimilarity of sample $t$ in $C_j$ to all samples in another cluster $C_i$, then $b(t) = \min\{d(t, C_i)\}$, $i = 1, 2, \ldots, k$, $i \neq j$. The formula to calculate the Silhouette index Sil of sample $t$ is:

$$Sil\left(t\right) = \frac{\left[b\left(t\right) - a\left(t\right)\right]}{\max\left\{a\left(t\right), b\left(t\right)\right\}} \tag{13}$$

The average Sil value of all the samples in a cluster reflects the clustering quality, where the largest average Sil value represents the best clustering quality and the optimal number of clusters [33]. With a series of Sil values corresponding to clustering solutions under different numbers of clusters calculated, the optimal clustering solution is found with the largest Sil [34].

## 4.2 The Analysis of Experimental Results

The algorithm using MATLAB simulation, to verify the advantage of the proposed CSA-DP clustering algorithm, it uses the four representative data sets: the data set Vote, Twomoon, Aggregation and Spiral. Vote data set contains 435 points, its distribution is relatively dense, the noise is more, as shown in Fig. 4 (a); Twomoon data set contains 180 points, its distribution is relatively sparse, as shown in Fig. 4 (b). Aggregation and Spiral are shown in Fig. 4 (c) and (d).



(a) Vote data set

(b) Twomoon data set

(c) Aggregation data set

(d) Spiral data set

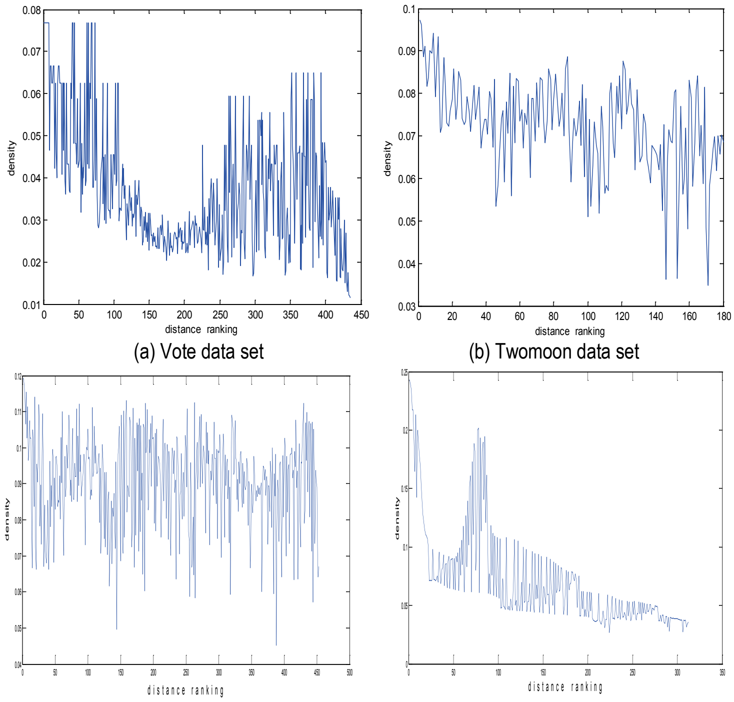
**Figure 4** Data sets

According to the actual circumstances of the data sets, using the total distance of the data point to the $K$ neighbor points, take the reciprocal as the data density, find a maximum density point of the data set, draw on the collating sequence to a maximum density point distance denoted abscissa of curve, and data points density represented as ordinate of curve. The graphs of Vote data set, Twomoon, Aggregation and Spiral are shown in Fig. 5.

Find trough point location, integer percentage according to proportion of the trough point abscissa to all points and regard as truncated distance. Then, that decision diagrams of the data set Vote, Twomoon, Aggregation and Spiral are shown in Fig. 6.

Conversely, if the parameter of truncation distance is set to 2 %, the decision graphs, respectively, are shown in Fig. 7. By comparing the Fig. 6, it can be seen that the decision graph of the proposed CSA-DP clustering algorithm is more intuitive, it can more accurately determine the clustering center.

The original DP algorithm needs to circle clustering centers artificially in decision graph, but there is no reliable basis about how many clustering centers and which data points should be circled in decision graph. For most decision graphs, there is no accurate boundary between the data points, with small density large distance and large density small distance. According to the idea proposed in this paper, on the basis of the product $\gamma$ of density and distance from large to small sort, it can be concluded that the Euclidean distance between them will change from big too small. In other words, the similarity will generally be higher than a certain value finally. According to these thought, the larger $\gamma$ is, the more likely it is to be the center of clustering, the similarity between clustering centers is lower, the similarity between clustering center and the points of the same cluster is higher, the clustering centers are the two points with the first two largest $\gamma$ for Vote data set, the clustering centers are the four points with the first four largest $\gamma$ for Twomoon data set, the eight points with the first eight largest $\gamma$ for Aggregation data set, the three points with the first three largest $\gamma$ for Spiral data set. The Euclidean distance in the experiment is shown in Tabs. 2-5.



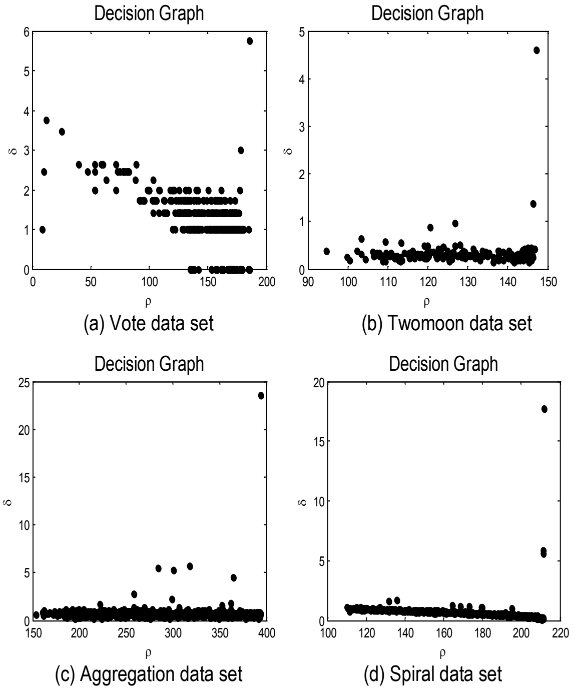
(a) Vote data set

(b) Twomoon data set

(c) Aggregation data set

(d) Spiral data set

**Figure 5** The distance to the maximum density of points in increasing order



(a) Vote data set

(b) Twomoon data set

(c) Aggregation data set

(d) Spiral data set

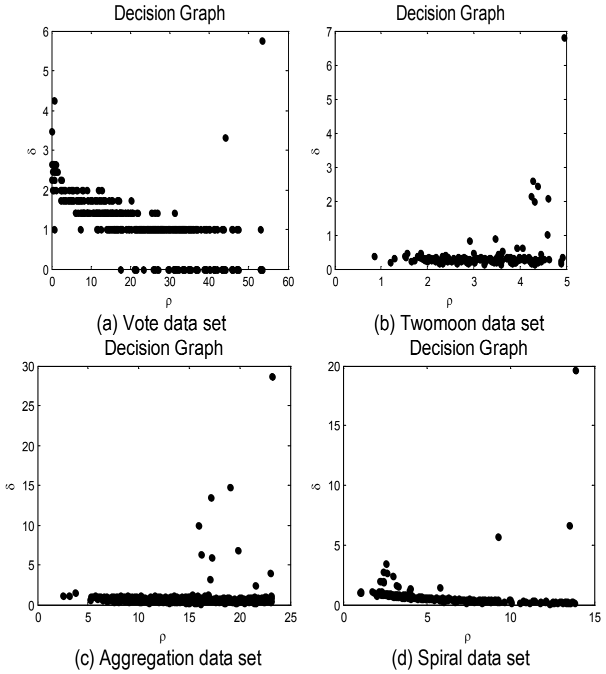
**Figure 6** Decision graph

**Table 2** Experimental results on vote data set

| Data point | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ |
|---|---|---|---|---|
| Euclidean distance | 3.4641 | 2.0000 | 2.2361 | 2.4495 |

**Table 3** Experimental results on twomoon data set

| Data point | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ |
|---|---|---|---|---|
| Euclidean distance | 1.3740 | 2.3332 | 6.1889 | 1.6697 |



(a) Vote data set

(b) Twomoon data set

(c) Aggregation data set

(d) Spiral data set

**Figure 7** Decision graph

**Table 4** Experimental results on aggregation data set

| Data point | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ | $i = 5$ |
|---|---|---|---|---|---|
| Euclidean distance | 9.9459 | 5.9737 | 15.9437 | 8.1844 | 26.3508 |
| Data point | i = 6 | i = 7 | i = 8 | i = 9 | i = 10 |
| Euclidean distance | 23.7803 | 17.2605 | 1.6125 | 11.5039 | 14.9624 |

**Table 5** Experimental results on spiral data set

| Data point | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ | $i = 5$ |
|---|---|---|---|---|---|
| Euclidean distance | 5.8577 | 5.7020 | 16.3500 | 22.7706 | 6.5487 |

**Table 6** The evaluation index contrast of vote data set for four algorithms

| Algorithm | Rand index | Jaccard coefficient | F-measure | Sil |
|---|---|---|---|---|
| K-means | 0.7335 | 0.5470 | 0.7157 | 0.2730 |
| AP | 0.6699 | 0.4492 | 0.6339 | 0.3416 |
| DP | 0.7820 | 0.6491 | 0.7874 | 0.4313 |
| CSA-DP | 0.7890 | 0.6586 | 0.7944 | 0.4277 |

**Table 7** The evaluation index contrast of twomoon data set for four algorithms

| Algorithm | Rand index | Jaccard coefficient | F-measure | Sil |
|---|---|---|---|---|
| K-means | 0.6801 | 0.4486 | 0.6300 | 0.6015 |
| AP | 0.6696 | 0.3957 | 0.5950 | 0.5881 |
| DP | 0.6744 | 0.3452 | 0.5875 | 0.6554 |
| CSA-DP | 0.8345 | 0.6672 | 0.8168 | 0.3364 |

**Table 8** The evaluation index contrast of aggregation data set for four algorithms

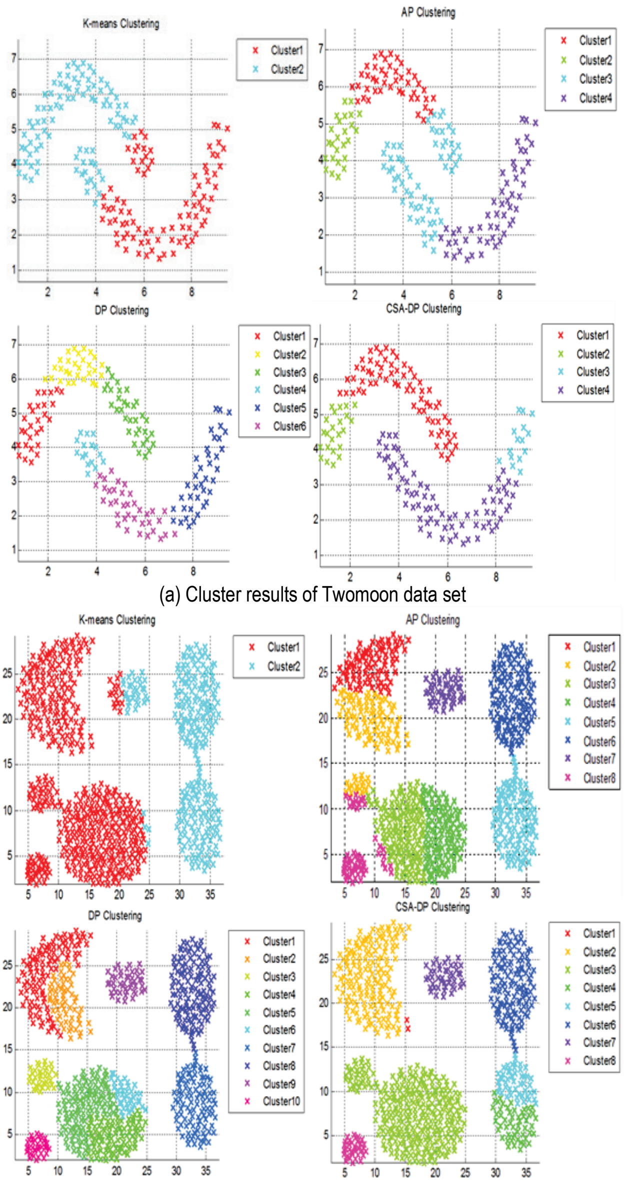
| Algorithm | Rand index | Jaccard coefficient | F-measure | Sil |
|---|---|---|---|---|
| K-means | 0.6545 | 0.3798 | 0.6120 | 0.6167 |
| AP | 0.9006 | 0.5618 | 0.7378 | 0.6065 |
| DP | 0.9051 | 0.5625 | 0.7496 | 0.5315 |
| CSA-DP | 0.9606 | 0.8402 | 0.9141 | 0.4324 |

**Table 9** The evaluation index contrast of spiral data set for four algorithms

| Algorithm | Rand index | Jaccard coefficient | F-measure | Sil |
|---|---|---|---|---|
| K-means | 0.4989 | 0.2470 | 0.4045 | 0.5041 |
| AP | 0.5961 | 0.1447 | 0.2595 | 0.4789 |
| DP | 1.0000 | 1.0000 | 1.0000 | -0.0867 |
| CSA-DP | 1.0000 | 1.0000 | 1.0000 | -0.0867 |

It can be seen from Fig. 6 and Fig. 7, to circle how many clustering centers in the end and which points are not clear. According to the algorithm presented in this paper, sort the data points from large $\gamma$ to small $\gamma$. For Vote data set, the Euclidean distance gets smaller at the second comparison points, and has been at a lower value later, namely, the similarity between the data points with the second largest $\gamma$ and the third largest $\gamma$ is larger. Therefore, take data points with the first two largest $\gamma$ as the clustering centers; For Twomoon data set, the Euclidean distance gets smaller at the fourth comparison points, namely, the similarity between the data points with the fourth largest $\gamma$ and the fifth largest $\gamma$ is larger. Therefore, take data points with the first four largest $\gamma$ as the clustering centers. For Aggregation data set, the Euclidean distance gets smaller at the eighth comparison points. Therefore, take data points with the first eight largest $\gamma$ as the clustering centeRS. For Spiral data set, when $i$ = 3, 4, the Euclidean distance is larger, and the Euclidean distances in the following are not small, which indicates the third and fourth compared points may be boundary cluster points or outliers. Therefore, the first three largest gamma points are used as the clustering centers of the Spiral data set.

The experiment was carried out on the four sets of Vote, Twomoon, Aggregation, Spiral, and compared with K-means algorithm, affinity propagation clustering algorithm (AP), density peaks clustering algorithm (DP) and the improved algorithm in this paper, it can be seen that the algorithm proposed here is not only better than the original DP algorithm, but also better than K-means algorithm and AP algorithm, its evaluation index is the highest. At the same time, it can also be seen that the original DP algorithm based on density clustering is better than K-means and AP that are both classical algorithms. It can be observed from the four tables above that the clustering accuracy of the density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster center is significantly better than that of the density peak clustering algorithm, it is especially obvious in Twomoon and Aggregation data set.



(a) Cluster results of Twomoon data set

(b) Cluster results of Aggregation data set

**Figure 8** Comparison of cluster results

Because Twomoon data set and Aggregation data set are two-dimensional, and therefore take these two data sets for example, the clustering results of the two data sets on K-means, AP, DP and CSA-DP are displayed in the two-dimensional plane later.

By clustering the two data sets, from the above two groups of clustering results can be clearly seen, K-means algorithm and AP algorithm can't get reasonable results for some data sets. But the clustering results obtained by the CSA-DP algorithm on the tested data sets can better reflect the real structure of the data.

From the comparison of the clustering accuracy, it is easy to find that the clustering accuracy of the CSA-DP algorithm proposed in this paper is obviously better than other algorithms. The clustering accuracy of CSA-DP algorithm, Rand and F-measure, is the highest in all test data sets, especially in the Spiral data set. This shows that the CSA-DP algorithm does avoid the problem of artificially setting up $d_c$, and improves the clustering performance of the algorithm.

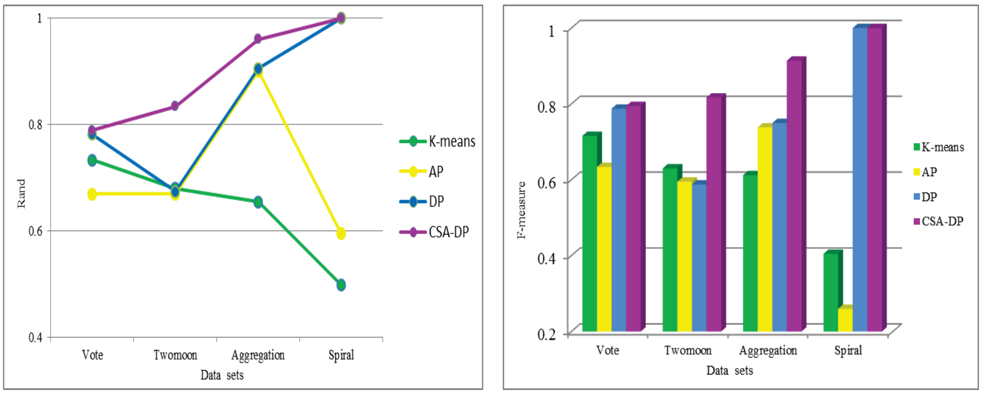


**Figure 9** Clustering accuracy comparison

**Table 10** Profitability of 147 bio pharmaceutical companies

| Stock ticker | Return rate of net assets (%) | Net profit margin (%) | Gross profit margin (%) | Net profit (Million Yuan) | Earnings per share (Yuan) | Operating income (Million Yuan) | Earnings per share (Yuan) |
|---|---|---|---|---|---|---|---|
| 002737 | 19.79 | 13.13 | 63.3092 | 164.4255 | 1.5016 | 1251.4335 | 11.4286 |
| 600535 | 17.66 | 12.18 | 37.5369 | 744.4962 | 0.7208 | 6108.6314 | 5.9143 |
| 002219 | 17.38 | 43.64 | 65.0937 | 159.3301 | 0.2585 | 365.0946 | 0.5923 |
| 002019 | 17.38 | 24.47 | 46.9743 | 107.1298 | 0.486 | 437.6843 | 1.9856 |
| 300406 | 17.22 | 42.82 | 74.1472 | 99.1379 | 0.9913 | 231.4856 | 2.3148 |
| 002294 | 15.32 | 37.98 | 76.6122 | 495.2557 | 0.7575 | 1303.7587 | 1.9942 |
| … | … | … | … | … | … | … | … |
| 300142 | -2.31 | -15.39 | 47.7081 | -61.8611 | -0.2643 | 401.7103 | 1.7167 |
| 000606 | -2.48 | -12.48 | 5.707 | -22.002 | -0.0466 | 176.2655 | 0.3733 |
| 300006 | -2.51 | -7.04 | 30.0923 | -24.9588 | -0.1236 | 354.0444 | 1.7544 |
| 000952 | -3.32 | -10 | 13.1804 | -23.0559 | -0.0915 | 230.4416 | 0.9155 |
| 600789 | -3.38 | -4.32 | 16.5175 | -50.0187 | -0.086 | 1157.803 | 1.9908 |
| 000004 | -6.56 | -17.15 | 58.7176 | -4.7063 | -0.056 | 27.4301 | 0.3266 |

## 5 APPLICATION OF THE IMPROVED ALGORITHM IN STOCK ANALYSIS

### 5.1 Data Selection

The data in this paper are from 147 bio pharmaceutical listed companies in 2014, select seven indicators reflecting the profitability of listed companies as the main object of study， return on net assets, net profit margin, gross profit margin, net profit, earnings per share, operating income and earnings per share. The clustering algorithm proposed in this paper is used for the clustering analysis, and the specific information is shown in Tab. 10.

### 5.2 The Analysis of Clustering Result

According to the above clustering results, combined with the analysis of the characteristics of financial data, this paper analyzes and summarizes the 147 listed companies in the bio pharmaceutical industry. (1) Class of blue chip, operating income and earnings per share are the most important financial indicators to measure the profitability of enterprises, are also the main basis for investors to assess the blue chip stocks. In the first category of enterprises, operating income and earnings per share are significantly higher than other categories of enterprises. (2) Stationary class, the overall financial characteristics of the second types of enterprises are similar, both the rate of return on net assets, earnings per share and net interest rates are at intermediate levels of the industry, the enterprise must have the core competitiveness, has a stable market share. They do not have a high level of profitability and growth momentum, but their steady development. (3) Growth class, compared with the former two categories of enterprises, although the profitability of the third types of enterprises have been reduced, the growth ability is very prominent. Most of these enterprises are young, new enterprises, or enterprises with reform and innovation. Although they do not occupy a stable market position, they have a strong momentum of development. (4) Downturn class, fourth types of enterprises do not have a higher level of profitability, the net interest rate is also very low, some companies even negative growth, the development of enterprises showed a decadent situation.

In this paper, the improved algorithm is applied to the cluster analysis of listed companies in the bio pharmaceutical industry, and the result is satisfactory. Therefore, the density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster center can provide an effective reference for investors to invest rationally and reduce investment risk, has a good application prospect.

**Table 11** Clustering results of 147 listed companies in bio pharmaceutical industry

| Category | Stock ticker |
|---|---|
| First species | 600535, 002019, 000963, 002022, 002728, 000538, 000739, 600276, 603998, 603456, 300267, 600332 |
| Second species | 002737, 002219, 300406, 002294, 600201, 002262, 600252, 600566, 002653, 002038, 600572, 300039 |
| Third species | 002365, 600200, 600488, 002399, 600222, 300239, 002099, 600671, 000606, 300006, 000952, 600789 |
| Fourth species | 000790, 300110, 300086, 000590, 002118, 300142, 000004 |

## 6 CONCLUSION

Density clustering algorithm uses data set density degree in a spatial region as the base to find clusters, can automatically discover the number of clusters and clusters of arbitrary shape. It handles the skew data sets well, therefore it is suitable to cluster unknown episodes [35]. This paper bases on the insufficient of the DP algorithm: need to manually enter the $d_c$ value, proposes an improved density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster center. The algorithm gives a method to determine $d_c$ and the cluster centers, they are based on the trough point position ratio of the total number as an approximate value of $d_c$ and the variation of similarity between the points which may be clustering centers. Find the distances between all the points and the point with the maximum density, next make the distances in ascending order and find the densities of the corresponding points. Then draw the coordinate diagram and find the wave trough. Through these steps, the automatic determination of the cut-off distance is realized last. As for cluster centers, find the data points which may be the centers of the cluster, compare the similarity between them and determine the final clustering centers. By comparing the simulation results, two representative data sets for clustering comparison, CSA-DP is indeed more accurate than DP. The next step is to continue to study the DP algorithm, and to optimize the density peak clustering algorithm using swarm intelligence optimization algorithm.

### Acknowledgments

The research was supported by the national science foundation of China under grant no. 61202306, 61472049, 61402193, 61572225 and 61202309; the national society science foundation of china under grant no. 15bgl090; the foundation of Jilin provincial science & technology department under grant no. 201215119, 20130522177jh, 20130101072jc, 20150312001zg and 20160101259jc; the society science research foundation of Jilin province under grant no. 2014b166; Jilin province education department under grant no. 2014159 and 2016 [78]; the open fund project of logistics industry economy and the intelligent logistics laboratory of Jilin province under grant no. 02060311 and 02060312; the fund of Jilin province economic structure strategic adjustment under grant no. 2015y040 and 2015y037.

## 7 REFERENCES

[1] Han, J. W. & Kamber, M. (2006). *Data Mining Concepts and Techniques*. 2$^{nd}$ ed. New York: Elsevier Inc, 383-424.

[2] Jin, W., Tung, A. K. H., Han, J. et al. (2006). Ranking outliers using symmetric neighborhood relationship. Wee-Keong N., ed. *Advances in Knowledge Discovery and Data Mining*. Berlin: Springer, 577-593. https://doi.org/10.1007/11731139_68

[3] Zhang, X., Furtlehner, C., & Sebag, M. (2008). Data streaming with affinity propagation. *Proceedings of the 2008 Machine Learning and Knowledge Discovery in Databases*, Berlin Heidelberg, Springer, 628-643. https://doi.org/10.1007/978-3-540-87481-2_41

[4] Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data sets. *SIGMOD Conference*, 427-438. https://doi.org/10.1145/335191.335437

[5] Xie, P., Zou, C. W., & Liu, Haier. (2012). Research on Internet financial model. *New financial review*, (1), 11-22.

[6] Liu, T. T. (2014). *Empirical analysis of the impact of financial position on stock prices in Chinese listed companies*. Southwestern University of Finance and Economics, 5-6.

[7] Liu, X. J., Chen, Min. et al. (2013). The present situation of Chinese stock market. *Technology and market*, (1), 125-125.

[8] Huang, X. Z. (2006). China stock market value of the investment and speculative thinking -- "securities investment" the combination of theory and practice. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition)*, (3), 91-95.

[9] Hong, C. & Yeung, D. Y. (2008). Robust path-based spectral clustering. *Pattern Recognition, 41*(1), 191-203. https://doi.org/10.1016/j.patcog.2007.04.010

[10] Bai, L. (2012). Theoretical Analysis and Effective Algorithms of Cluster Learning. *Dissertation*, Taiyuan: Shanxi University, 133-139.

[11] Xu, R. (2005). Survey of clustering algorithm. *IEEE Tran on Neural Networks, 16*(3), 645-678. https://doi.org/10.1109/TNN.2005.845141

[12] Clara, P. & Domenico, T. (2003). P-Auto Class: scalable parallel clustering for mining large data sets. *IEEE Trans on Knowledge and Data Engineering, 15*(3), 629-641. https://doi.org/10.1109/TKDE.2003.1198395

[13] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recogn Lett*, 31, 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

[14] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global K-means clustering algorithm. *Pattern Recogn*, 36, 451-464. https://doi.org/10.1016/S0031-3203(02)00060-2

[15] Xie, J. Y., Jiang, S., Xie, W. et al. (2011). An efficient global K-means clustering algorithm. *J Comput, 6*, 271-279. https://doi.org/10.4304/jcp.6.2.271-279

[16] Duan, M. X. (2009). Research and application of hierarchical clustering algorithm. *Dissertation*, Central South University, 6-7.

[17] Ji, C. & Lei, Y. (2017). Parallel clustering by fast search and find of density peaks. *International Conference on Audio, Language and Image Processing, IEEE*, 563-567.

[18] Cai, Y. & Yuan, J. S. (2011). Text Clustering Based on Improved DBSCAN Algorithm. *Computer Engineering, 37*(12), 49-50.

[19] Wang, G. Z. & Wang, G. L. (2009). Improved fast DBSCAN algorithm. *Journal of Computer Applications, 29*(9), 2505-2508. https://doi.org/10.3724/SP.J.1087.2009.02505

[20] Rodriguez, A. & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492-1496. https://doi.org/10.1126/science.1242072

[21] Wang, S., Wang, D., Li, C. et al. (2016). Clustering by Fast Search and Find of Density Peaks with Data Field. *Chinese Journal of Electronics, 25*(3), 397-402. https://doi.org/10.1049/cje.2016.05.001

[22] Mehmood, R., Bie, R., Dawood, H. et al. (2015). Fuzzy Clustering by Fast Search and Find of Density Peaks. *International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), IEEE*, 258-261. https://doi.org/10.1109/IIKI.2015.62

[23] Wang, S., Wang, D., Li, C. et al. (2015). Comment on "Clustering by fast search and find of density peaks". arXiv preprint arXiv:1501.04267.

[24] Zhang, W. & Li, J. (2015). Extended fast search clustering algorithm: widely density clusters, no density peaks. arXiv preprint arXiv:1505.05610.

[25] Huang, L., Wang, G., Wang, Y. et al. (2016). A link density clustering algorithm based on automatically selecting density peaks for overlapping community detection. *International Journal of Modern Physics B, 30*(24), 1650167. https://doi.org/10.1142/S0217979216501678

[26] Liu, W. & Hou, J. (2017). Study on a density peak based clustering algorithm. *International Conference on Intelligent Control & Information Processing, IEEE*, 60-67.

[27] Liu, S., Zhu, L., Sheong, F. K. et al. (2017). Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories. *Journal of Computational Chemistry, 38*(3), 152. https://doi.org/10.1002/jcc.24664

[28] Du, M., Ding, S., Xu, X. et al. (2017). Density peaks clustering using geodesic distances. *International Journal of Machine Learning & Cybernetics*, 1-15. https://doi.org/10.1007/s13042-017-0648-x

[29] Xie, J., Gao, H. & Xie, W. (2016). K-nearest neighbors optimized clustering algorithm by fast search and ending the density peaks of a data set. *Scientia Sinica Informationis, 46*(2), 258-280. https://doi.org/ 10.1360/N112015-00135-20

[30] Wang, J., Xia, L. N., & Jing, J. W. (2009). Maximum density clustering algorithm. *Journal of the Graduate School of the Chinese Academy of Sciences, 26*(4), 539-548.

[31] Mauceri, C. & Ho, Diem. (2007). Clustering by kernel density. *Computational Economics, 29*(2), 199-212. https://doi.org/10.1007/s10614-006-9078-7

[32] Wang, L., Wu, L. L., & Fu, D. M. (2014). A density-based fuzzy adaptive clustering algorithm. *Beijing Keji Daxue Xuebao/Journal of University of Science & Technology Beijing, 36*(11), 1560-1566.

[33] Tan, Y., Rui-Fei, H. U., & Yin, G. F. (2008). Adapted DBSCAN with multi-threshold. *Journal of Computer Applications, 28*(3), 745-748. https://doi.org/10.3724/SP.J.1087.2008.00745

[34] Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems, 120*(15), 92-96. https://doi.org/10.1016/j.chemolab.2012.11.006

[35] Zhou, R., Zhang, S., Chen, C. et al. (2016). A Distance and Density-Based Clustering Algorithm Using Automatic Peak Detection. *IEEE International Conference on Smart Cloud*, 176-183. https://doi.org/10.1109/SmartCloud.2016.39

**Contact information:**

**Limin WANG,** Professor
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech, Changchun 130117, China
wlm_new@163.com

**Mingyang LI,** Postgraduate
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech, Changchun 130117, China
724306390@qq.com

**Xuming HAN,** Professor
Corresponding author
School of Computer Science and Engineering
Changchun University of Technology, Changchun 130012, China
hanxvming@163.com

**Ruihong ZHOU,** Doctor
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech, Changchun130117, China
zrh@jlufe.edu.cn

**Kaiyue ZHENG,** Postgraduate
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech, Changchun 130117, China
495195583@qq.com

**Meihan LIU,** Postgraduate
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech, Changchun 130117, China
liumeihan0327@vip.qq.com