Environmental Engineering and Management Journal *November 2021, Vol. 20, No. 11, 1853-1864*
http://www.eemj.icpm.tuiasi.ro/; http://www.eemj.eu

"Gheorghe Asachi" Technical University of Iasi, Romania

AN ANALYSIS OF AIR QUALITY WITH DENSITY PEAK CLUSTERING AND COULOMB FORCE THEORY

Limin Wang[1], Honghuan Wang[2], Wei Zhou[3], Xuming Han[4*]

[1]School of Information, Guangdong University of Finance & Economics, Guangzhou 510320, China
[2]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
[3]School of Computer Science and Engineering, Changchun University of Science and Technology, Changchun 130022, China
[4]College of Information Science and Technology, Jinan University, Guangzhou 510632, China

Abstract

The variation of air quality has strong nonlinear characteristics, so it is difficult to obtain accurate analysis results. In order to overcome this defect, this paper introduces the Coulomb force theory into the density peak clustering (CDPC), and builds a new method for air quality analysis which is characterized by data similarity detection. We test the proposed algorithm in the experimental dataset and compare it with other most advanced algorithms. The simulation results show that the proposed algorithm has better clustering performance. We use Changchun air quality data as an example to verify the practicality of the method and to provide an effective tool for air quality analysis. We apply the artificial intelligence theory to the air pollution research field, and realize the interdisciplinary integration to provide a new method for the air quality analysis field. This method provides a new technology and solution for the construction and planning of smart city, and it has certain practical value to provide effective reference basis and intellectual support for the innovation and management of smart city.

Key words: air quality, clustering analysis, Coulomb force theory, density peak clustering, smart city

Received: November, 2020; Revised final: May, 2021; Accepted: October, 2021; Published in final edited form: November, 2021

1. Introduction

The quality of atmospheric environment is deeply involved in human health and lives. Air quality analysis has gradually become the focus of pollution prevention. Air quality is an essential element of a smart environment (Boncescu and Robescu, 2017; Graiia and Gago, 2018). Changes in air quality are one of the most basic guidelines for properly evaluating intelligent cities. By observing the change of the daily air quality index of the city, we can clearly understand the present situation of the development of the smart city (Pop et al., 2018). Cluster analysis technology is a technique used in statistical data analysis, mainly in machine learning, data mining, pattern recognition, image analysis, and bioanalysis (Abirami and Chitra, 2019; Amruthnath and Gupta, 2018; Bradley et al, 2019; Hajarolasvadi and Demirel, 2019). This paper use cluster analysis technology to cluster the air quality index of Changchun of Chinese. So as to analyze the construction of Changchun in the past five years smart city processes, the environment changes and put forward rationalization proposals (Dai et al., 2019).

The goal of smart city is maximize optimize the city function, promote economic growth and improve the urban residents quality life by using intelligent technology and data analysis (Gaur et al., 2015). In the process of constructing smart city, the improvement of smart environment is also crucial. Smart cities can leverage resources and sensible technologies to perceive, analyse and aggregate critical information in

*Author to whom all correspondence should be addressed: E-mail: hanxvming@163.com; Phone: +86 13134463221

the core processes that run cities (Wenge et al., 2014). So that the intelligent response to people welfare, pollution, social safety, urban management, commercial activities and other needs, to create a better city life for humans (Afify et al., 2020; Parmar et al., 2018). Consequently, it is important to use effective cluster analysis technique to analyze air quality (Dragomir et al., 2019; Florea et al., 2019).

Cluster analysis, as an essential data mining skill, is used in the classification process where one does not need to give any pre-determined criteria for classification and it can be done automatically from the specimen datum. The varying techniques used in cluster testing often bring about diverse conclusions (Afrasiabi et al., 2019, 2020; Altintas and Ber, 2001; Chattopadhyay et al., 1991; Wegener et al., 2019). In clustering analysis technology, many kinds of algorithms are often used, such as DPC (density peak clustering) (Rodriguez and Laio, 2014), AP (affinity propagation clusering) (Frey and Dueck, 2007), K-means (Hartigan and Wong, 1979) algorithm and so on. DPC as a new algorithm proposed in 2014, can effectively cluster the data by calculating and analyzing the distance δ and local-density ρ to determine the cluster center point. Data points belonging to the same cluster have higher similarity (Zhang et al., 2018).

DPC algorithm has some limitations in the process of clustering data, among which unclear identification of clustering centers and outlier detection are the two main limitations (Chen et al., 2016; Shi et al., 2018). The DPC algorithm clustering process usually needs to combine decision graph and γ graph to judge the cluster center points (Parmar et al., 2019). In most cases, the class centers can be judged by the decision figure. Even when the class centers in the figure are not clearly displayed, it can be well judged by γ graph (Buczak and Guven, 2015; Mehmood et al., 2016). This may lead to errors in the classification of data in the datasets (Cai et al., 2015; Role and Nadif, 2014; Wang et al., 2019). Therefore, it is necessary to strengthen the recognition ability of clustering centers . Outlier discovery process is the act of finding outliers in the clustering process. In practice, outliers can destroy the clustering effect of data and affect the prediction (Palm et al., 2018; Wei et al., 2018). Therefore, it is necessary to strengthen the outlier detection of the algorithm.

In recent years, intelligent cities have received more and more intense focus because of its highly intelligent, information and other characteristics. With the processing of smart city construction, smart environment is an essential part of smart environment construction (Kilic et al., 2018; Lei et al., 2017; Labianca et al., 2018; Pejic et al., 2017). The quality of an urban environment is an important evaluation criterion for the success of smart city construction (Năstase and Șerban, 2019). The assessment of environmental quality is usually measured by the quality of air (Dutta and Banerjee, 2019; Fan et al., 2016; Li et al., 2019; Sanchez-Fernandez et al., 2016). If there is a good clustering algorithm to provide technical help for air quality analysis, it will be of great help to smart environment analysis.

From the above issues, a novel density peak clustering algorithm based on Coulomb force theory (CDPC) is proposed in this paper. We have tested multiple classical datasets using CDPC algorithm to prove the feasibility and correctness of the new algorithm. Compared with DPC and other algorithms, CDPC overcome the problem of unclear cluster centers recognition and limitation of outliers detection. The main relevancy of this paper are as follows. Firstly, the Coulomb force theory is introduced into DPC algorithm to optimize the γ graph to enhance the recognition ability of the cluster center. Secondly, improved ability to detect outliers. Finally, the new algorithm provides more reliable technical assistance for the building of intelligent environment in Changchun during the past five years.

2. Materials and methods

This paper introduces Coulomb force theory into DPC algorithm, and then forms CDPC algorithm. This algorithm perfectly incorporates the Coulomb force theory into the local density metric of the DPC algorithm. The similarity between the data is measured by modeling the data as static charge in the magnetic field and calculating the Coulomb force. Coulomb forces can be widely applied in many fields. CDPC algorithm have the following assumptions: (1) the total value of similarity between the cluster center and the data points except itself is larger than that corresponding to the other data points in the cluster which it belongs; (2) data points can be assigned to clusters of data points with large total value of similarity. CDPC algorithm includes three main steps in execution. Firstly, calculate density and δ values for all points in the dataset. Secondly, generate Coulomb force decision graph and γ graph. Thirdly, form data clusters with Coulomb force. In Table 1, we show the list of full-text nomenclature.

2.1. Calculate the local density of all data points in the dataset

In the same way as DPC, we calculate their local-density ρ adapting a typical intercept d_c and rank ρ values in descent order, as follows Eq. (2):

$$d_{ij} = dis\tan ce(node_i, node_j) \tag{1}$$

$$\rho_i = \sum\nolimits_j \chi(x) \times (d_{ij} - d_c) \tag{2}$$

If $x>0$, $\chi(x)=0$, if not, $\chi(x)=1$, and d_c is the cut-off distance defined manually by the user. Another local density $node_i$ presented as follows Eq. (3):

$$\rho_i = \sum\nolimits_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{3}$$

The distance δ can be defined as follows Eq. (4):

$$\delta_i = \begin{cases} \min\{d_{ij}\}, & if\ \exists j, \rho_j > \rho_i \\ \max\{d_{ij}\} & otherwise \end{cases} \quad (4)$$

Table 1. Nomenclature list

Abbreviation	*Terminology full name*
AP	Affinity propagation clusering
CDPC	Coulomb force density peak clustering
DPC	Density peak clustering
d_{ij}	Euclidean distance
d_c	Cut-off distance
F	Coulomb force
Fm	F-measure indicator
K-Means	K-means clustering
k	Coulomb constant
F_{ij}	The similarity between two data points
F_i	The sum of similarity for every point in the database
K_{ij}	The weight of similarity
q_i	The charge amount of the *i*-th charge
q_j	The charge amount of the *j*-th charge
R	The distance between two charges
Sil	Silhouette indicator
γ_i	Cluster center point judgment parameter
δ_i	Distance of data points
ρ_i	Local-density
ρ_{max}	Maximum local-density
ρ_{min}	Minimum local-density

2.2. Generate Coulomb force decision diagram and γ diagram

Coulomb law states that the interaction between the charges of two stationary points in a vacuum is reversely correlated to the square of the length and proportional to the product of the electric quantity. The direction of the force is on their line. The coulomb law is shown in Eq. (5):

$$F = k \times \frac{(q_i \times q_j)}{r^2} \quad (5)$$

F is the Coulomb force between two stationary charges; and k is the coulomb constant; q_i is the charge amount of i-th charge; q_j is the charge amount of the j-th charge; and r is the distance between two charges.

Coulomb law inspired the idea that the points of a dataset could be imagined as static charges in a vacuum. So that the similarity between data points can be measured by calculating the Coulomb force. The greater the coulomb force between two data points, the greater the similarity between them and the greater the possibility of belonging to the same cluster class. A local density can be replaced by a Coulomb force F. From Eq. (7) and Eq. (8), $F_i > \rho_i$, so the points of the data move to the right in the decision diagram as a whole. By $\gamma_i = (F_i \times \delta_i)$ leads to an increasing value of γ_i, which highlights the clustering centroids and outliers near the y-axis even more, thus making the clustering effect more accurate. Hence, there is a more precise way to measure cluster centers and outliers in this paper. A mapping between Coulomb force theory and the parameters of DPC are shown in Table 2. In the Table 2, we can use Eq. (6) to calculate the Coulomb force between two data points, that is, the similarity between two data points.

$$F_{ij} = \frac{(q_i \times q_j)}{d_{ij}^2} \quad (6)$$

Further, find the sum of similarity for every point in the database Eq. (7):

$$F_i = \sum\nolimits_j K_{ij} \times \frac{(q_i \times q_j)}{d_{ij}^2} \quad (7)$$

In Eq. (7), the K_{ij} is the weight of similarity, the value of K_{ij} is defined as Eq. (8):

$$K_{ij} = \frac{(\rho_{\max} - \rho_{\min})}{(\rho_i - \rho_j)^2} \quad (8)$$

Table 2. Mapping between Coulomb force theory and the parameters of DPC algorithm

Parameters from Coulomb force	*Parameters from DPC*	*Equations*
q_i	ρ_i	Eq. (2) or Eq. (3)
q_j	ρ_j	Eq. (2) or Eq. (3)
r	d_{ij}	Eq. (1)

From the road network model and its application (Cai et al., 2015), we can know that the local-density of area where the points are located closely affects the similarity between the points. The closer the local-density of two different data points, the higher the similarity. Therefore, the similarity weights k is used to represent the local connectivity of spatial datasets. In Eq. (8), $(\rho_i\text{-}\rho_i)^2$ is used to represent the square of the local density difference value of different data points, which is inversely proportional to its corresponding similarity, so it is used as the denominator. The $\rho_{max}\text{-}\rho_{min}$ is introduced to provide a uniform measure of the local density difference values of different data. The similarity weights are inversely proportional to the difference in density size between different data points.

At the beginning of clustering, we do not know that the whole dataset can be divided into several classes numbers. So at first we default that the data in the entire dataset is the same class. Then we stipulate that the similarity weights are calculated by Eq. (8). In fact, for the data belonging to the same cluster, the local density difference satisfies the inverse ratio condition. While for different clusters, the local density difference does not satisfy the inverse ratio condition. However, because the distance between the points in the same class is small and the distance between the points in the diverse class is large. According to the Eq. (6), the similarity about two points of same cluster is much larger than that between the two data points of different clusters. In a word, after adding the similarity weight, the similarity of points of the same class is still larger than that of points of different classes. Therefore, depending on

the criterion of Eq. (8), we can still distinguish whether two points pertains to the same class. Finally, the total of the force of each point and other nodes except itself through Eq. (7), also called the sum of similarity. The greater F_i of the total similarity index of the point, the better the mutuality between the point and these nodes in its dataset. Therefore, the better the value of similarity sum, the more likely the point is the cluster center.

In Fig. 1, we show how the calculation of the local density of data points can be improved by introducing Coulomb force theory. First, we hypothesize each data point as an electric charge. Then, we calculate the similarity between data points x_i and x_j, i.e., F_{ij} in Fig. 1, by Eq. (6). Further, the local density of data point x_i, i.e. F_i in Fig. 1, is calculated by Eq. (7).

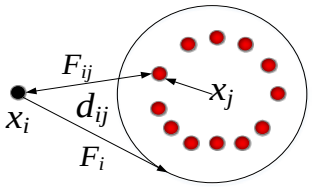


Fig. 1. Calculate the local density of data points by using coulomb force theory

In the DPC algorithm, for those cases in which the cluster center can not be judged in the decision graph, a formula is given to combine ρ value with δ value as follows Eq. (9):

$$\gamma_i = (\rho_i \times \delta_i) \quad (9)$$

In the CDPC algorithm, we changed γ calculation method to the follow Eq. (10):

$$\gamma_i = (F_i \times \delta_i) \quad (10)$$

where $\delta_i = min_{Fj>Fi}(d_{ij})$ or $\delta_i = max_{Fmax}(d_{ij})$.

Similar to the DPC algorithm, the greater γ_i value, the more possibly it is to be the cluster center point. In summary, we use F_i as transverse axis and δ_i as longitudinal axis to form the decision diagram. Using the newly defined γ_i parameters we can draw the γ graph.

2.3. Form data clusters with Coulomb force

Clustering based on Coulomb force theory is done by detecting clustering centers based on Coulomb force and distance. First, the cluster center points of a class are detected by its reasonable high coulomb force and distance. Second, any data points can be assigned to a corresponding cluster. Finally, the outliers can be identified by their relatively high distance, relatively low Coulomb force.

For the CDPC algorithm, we classify the outliers into a separate category, and the cluster classes to which the outliers belong also have their "cluster centers". The γ values of the clustering centroids of the outlier points are much smaller than those of the normal clustering centroids.

3. Results and discussion

3.1. Experimental data

In order to maintain the integrity of the simulation experiment, the test environment is Windows10 system, Intel Core i7 processor and MATLAB2016a programming language environment. In this paper, six groups of synthetic and real-world datasets are selected, and the datasets are shown in Table 4.

3.2. Evaluating indicator

For the purpose of test the effectiveness, *Silhouette* (*Sil*) and *F-measure* (*Fm*) index, are utilized in the test to analyze the clustering performance of the updated approach.

Table 3. The details of CDPC algorithm flow

Algorithm 1	***The CDPC algorithm.***
Require:	Initial data points x_i, d_c
Ensure:	Accurately identify all clusters and outliers
Step 1:	Compute local-density ρ and distance δ values for all data points in the dataset
	1.1 Compute d_{ij} by distance equation
	1.2 Sort d_{ij} in the ascending order
	1.3 Determine the cut-off distance d_c by using the DPC principle.
	1.4 Compute ρ_i by Eq. (2) or Eq. (3)
	1.5 Put all local density ρ_i in the descending order
	1.6 Calculated distance δ_i values based on (Eq. (4)
Step 2:	Generate coulomb force decision diagram and γ diagram
	2.1 Compute coulomb force by Eq. (7)
	2.2 Compute γ values by Eq. (10)
	2.3 Form the decision graph with distance δ and coulomb force F; form γ chart with newly defined γ
Step 3:	Form data clusters with coulomb force
	3.1 Assign the value of the coulomb force to all data points
	3.2 Identify normal cluster centers and outlier clusters centers
	3.3 Iterate until all data points are clustered

Table 4. Experimental datasets

Datasets	*Points*	*Dimensions*	*Clusters*
Flame	240	2	2
Spiral	312	2	3
Aggregation	788	2	7
D1	87	2	3
Iris	150	4	3
Wine	178	13	3

Suppose there is a sample set $D=\{x_1, x_2,..., x_n\}$, which includes n sample points, and these sample points are divided into k clusters C_i (i=1, 2,..., k). $a(i)$ is the average distance between a sample and other samples in its cluster, reflecting the degree of cohesion. $b(i)$ is the average distance between a sample and other cluster samples, reflecting the degree of separation. Then the contour coefficient of the i-th object is in Eq. (11):

$$Sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{11}$$

The mean value of $Sil(i)$ of all samples is called the contour coefficient of the clustering result, which is a measure of whether the clustering is reasonable and effective. The value of the contour coefficient of the clustering result is between [-1, 1]. The larger the $Sil(t)$ value, the more compact the cluster and the more scattered between the clusters, the better the clustering effect. The *Fm* value is an exterior index combining the exactitude P and the recall rate R. For real cluster P_j and clustered cluster C_i, the formulas for calculating the exactitude P and recall rate R are shown in Eq. (12) and Eq. (13):

$$P(P_j, C_i) = \frac{|P_j \cap C_i|}{|C_i|} \tag{12}$$

$$R(P_j, C_i) = \frac{|P_j \cap C_i|}{|P_j|} \tag{13}$$

The formula for the *F-measure* indicator is as shown in Eq. (14):

$$F(P_j, C_i) = \frac{2 \cdot P(P_j, C_i) \cdot R(P_j, C_i)}{P(P_j, C_i) + R(P_j, C_i)} \tag{14}$$

The value of *Fm* indicator is in the range [0,1]. The greater the *Fm* index, the higher the clustering performance. A weighted average of the *F-measure* values of all clusters gives the *F-measure* value of the entire clustering result (Eq. 15):

$$F = \sum_j \frac{|P_j|}{N} \max_i F(P_j, C_i) \tag{15}$$

As shown in Table 5, the *Sil* index value obtained by CDPC algorithm are almost the same as that obtained by DPC algorithm. Compared with K-Means and AP algorithms, except for the big difference in the *Sil* index values which obtained from the data clustering of *Flame* and *Spiral* datasets, the index values obtained from the other four datasets are basically flat. The *Sil* index of *D1* dataset shows NaN, indicating that AP algorithm cannot obtain effective clustering results for *D1* dataset. However, other algorithms get good results in this dataset. This shows that the CDPC algorithm is robust to the processing of different datasets.

As shown in Table 6, the *F-measure* index value obtained by CDPC algorithm are almost the same as that obtained by DPC algorithm. Also, the CDPC algorithm yields overall higher *F-measure* values than K-Means and AP. Therefore, it can be shown that the CDPC algorithm has a high clustering accuracy.

3.3. Identifying the cluster centers accurately and clearly

Compared with the DPC algorithm, the CDPC algorithm can identify the cluster centers more accurately and clearly through decision graph and γ graph, and then can cluster more accurately.

As shown in Fig. 2, for the *Iris* dataset, the decision graph of both DPC and CDPC is difficult to identify the cluster centers. The specific number of cluster centers can not be determined. But when we combine the γ graph to judge, we can find that the CDPC algorithm is very clear to judge that there are three center points, while the DPC algorithm determines that there are only two cluster center points. The final result graph shows that the CDPC algorithm achieves correct clustering of *Iris* dataset.

3.4. Identifying outliers accurately

Compared with DPC algorithm, K-means algorithm and AP algorithm, CDPC algorithm can accurately identify the outliers in some datasets. The CDPC algorithm can identify the outliers in the upper left corner of the *Flame* dataset as shown in Fig. 3. As shown in Fig. 4, the DPC and CDPC algorithms are able to cluster both *Spiral* and *Aggregation* datasets correctly. Therefore, CDPC can achieve the same effect as DPC to achieve correct clustering for different types of datasets that do not contain anomalies. Based on Coulomb force theory, CDPC algorithm is proposed. The algorithm identifies clustering centers and outliers by Coulomb force.

We have evaluated the new algorithm through experiments on different datasets. Simulation tests show that the CDPC algorithm has the following functions:

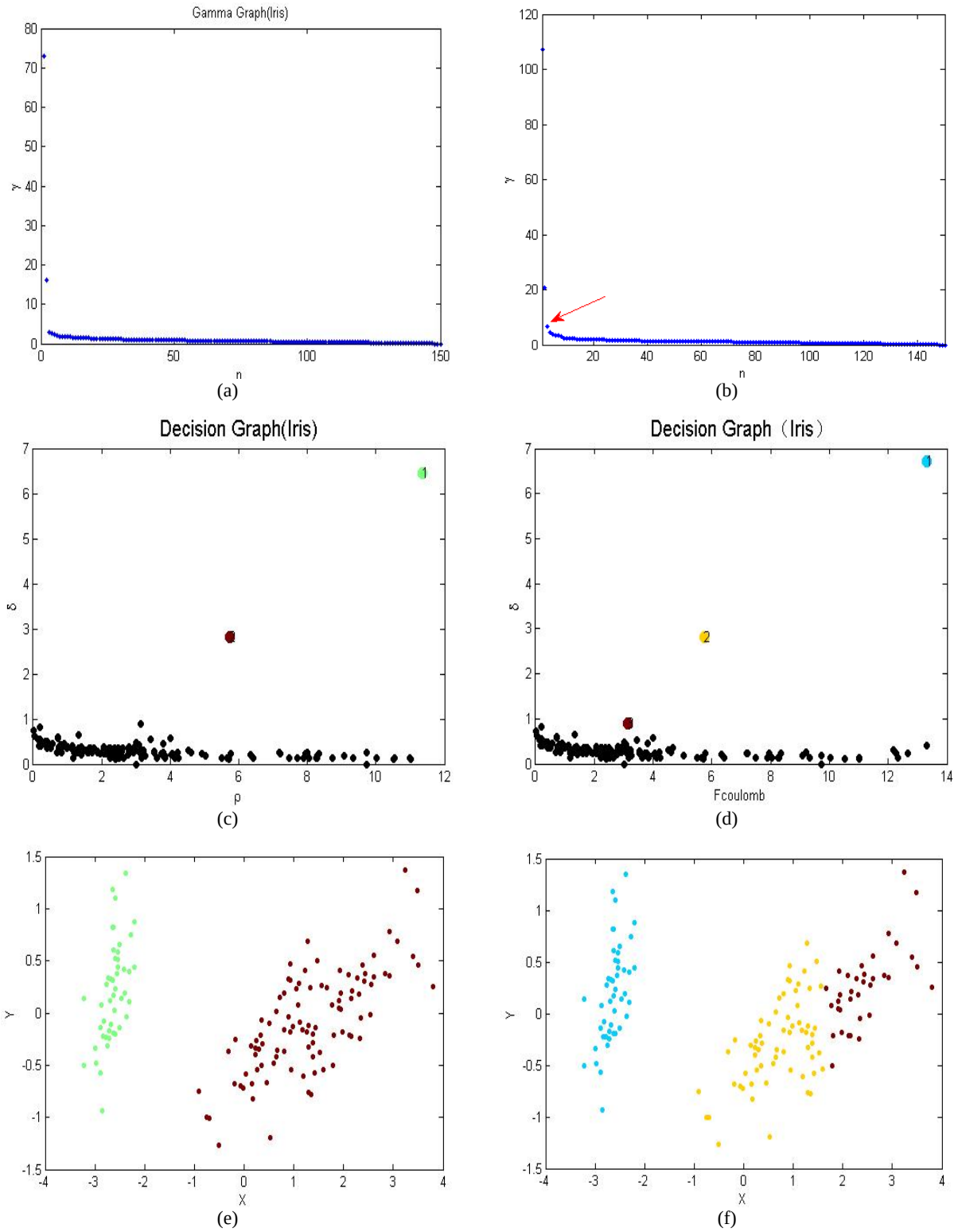


Fig. 2. Identify the cluster centers accurately and clearly by CDPC in *Iris* dataset: (a) γ graph of DPC, (b) γ graph of CDPC, (c) decision graph of DPC, (d) decision graph of CDPC, (e) clustering result of DPC (cluster1 is green,cluster 2 is brown), (f) clustering result of CDPC (cluster1 is green,cluster 2 is brown, cluster 3 is yellow)

Firstly, generate γ graph that make it easier to accurately identify the quantity of clustering centers; secondly, identifying outliers accurately; finally, processing different types of datasets effectively. Based on the original DPC algorithm, this paper use Coulomb force theory to enhance the performance of the original method. We change the transverse axis of the decision graph of the DPC algorithm from local density to Coulomb force and change the calculation method of parameter γ to the one shown in Eq. (10). So as to identify clustering centers and outliers more effectively.

Table 5. *Sil* index in different datasets

Clustering algorithms	*Sil*					
	Flame	*Spiral*	*Aggregation*	*D1*	*Iris*	*Wine*
K-Means	0.5815	0.5317	0.7068	0.7331	0.7355	0.8214
AP	0.5217	0.5256	0.6768	NaN	0.8492	0.6973
DPC	0.4250	-0.0867	0.6419	0.6988	0.8462	0.6943
CDPC	0.3496	-0.0867	0.6419	0.6988	0.8462	0.6943

Table 6. *F-measure* index in different datasets

Clustering algorithms	*F-measure*					
	Flame	*Spiral*	*Aggregation*	*D1*	*Iris*	*Wine*
K-Means	0.7006	0.3276	0.7829	0.9745	0.8207	0.5835
AP	0.7645	0.3335	0.7637	0.5874	0.7642	0.6029
DPC	1	1	1	1	0.7715	0.5892
CDPC	0.9945	1	1	1	0.7715	0.5892

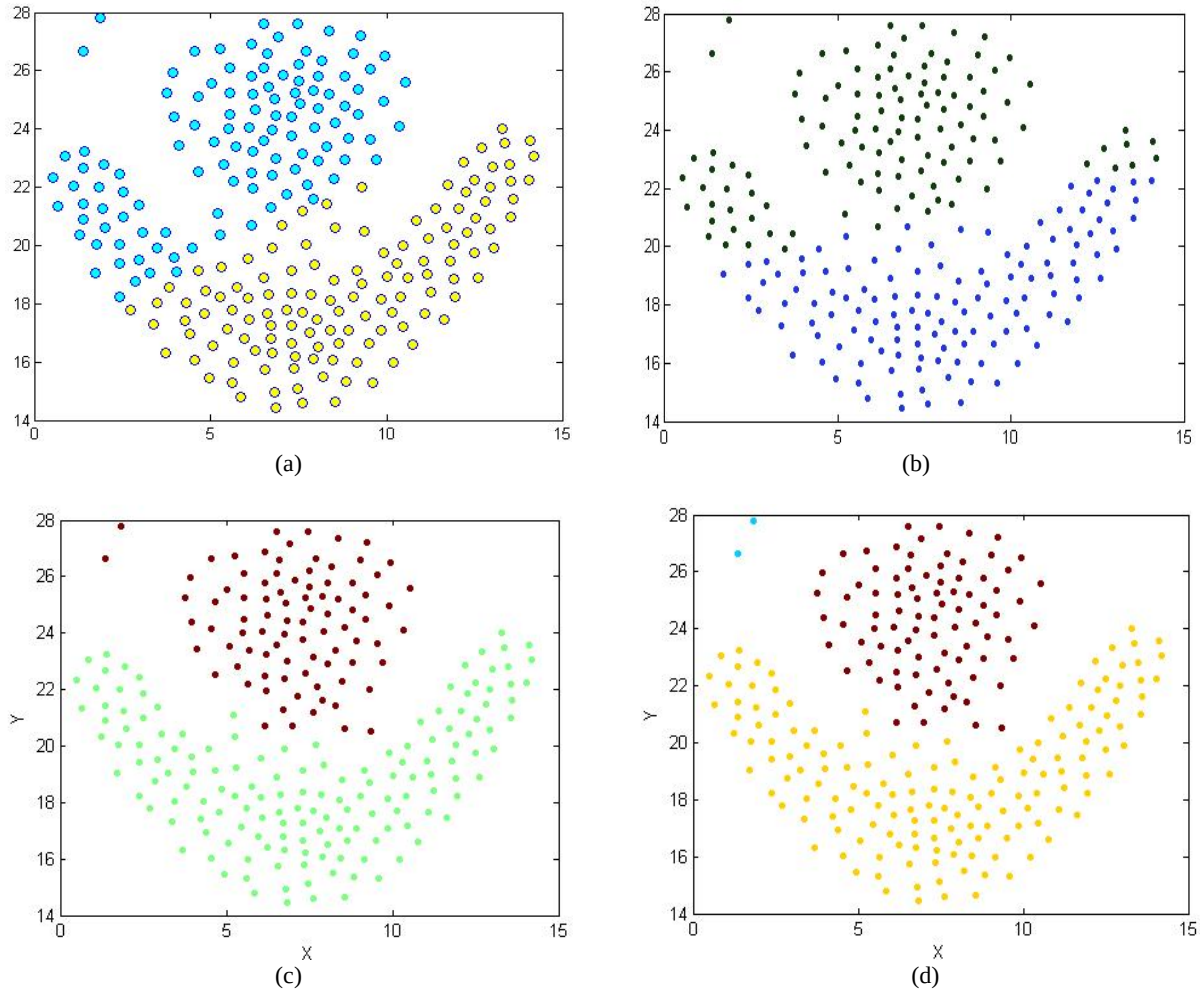


Fig. 3. Identify outliers accurately in *Flame* dataset by CDPC: (a) K-Means (cluster1 is blue, cluster 2 is yellow), (b) AP (cluster1 is blue, cluster 2 is black), (c) DPC (cluster1 is green, cluster 2 is brown), (d) CDPC (cluster1 is yellow, cluster 2 is brown, *outliers* is blue)

CDPC is superior to the DPC in the γ graph. Regarding the γ figure, in Fig. 2, we can easily find that the number of cluster center points can not be clearly judged in the decision figure of DPC algorithm. The CDPC algorithm can get the correct number of cluster center points very clearly by γ graph. CDPC algorithm has the ability to accurately identify outliers compared with DPC、K-Means and AP algorithms. It can divide the outliers in the dataset into one class, namely the outliers cluster class, as shown in Fig. 3. For the CDPC algorithm, the cluster center points of outlier cluster can be successfully identified. In DPC algorithm, the outliers usually have larger δ values and smaller local densities. The γ values corresponding to the outliers are similar to the γ values of the points in other non-clustered centers,

which leads to the outliers being automatically divided into the normal clusters during clustering. But in the CDPC algorithm, the γ value of the cluster center of the abnormal point cluster class is amplified. The cluster center of the corresponding abnormal point cluster class will be recognized, so as to achieve the purpose of identifying the abnormal point. The details are shown in Fig. 5.

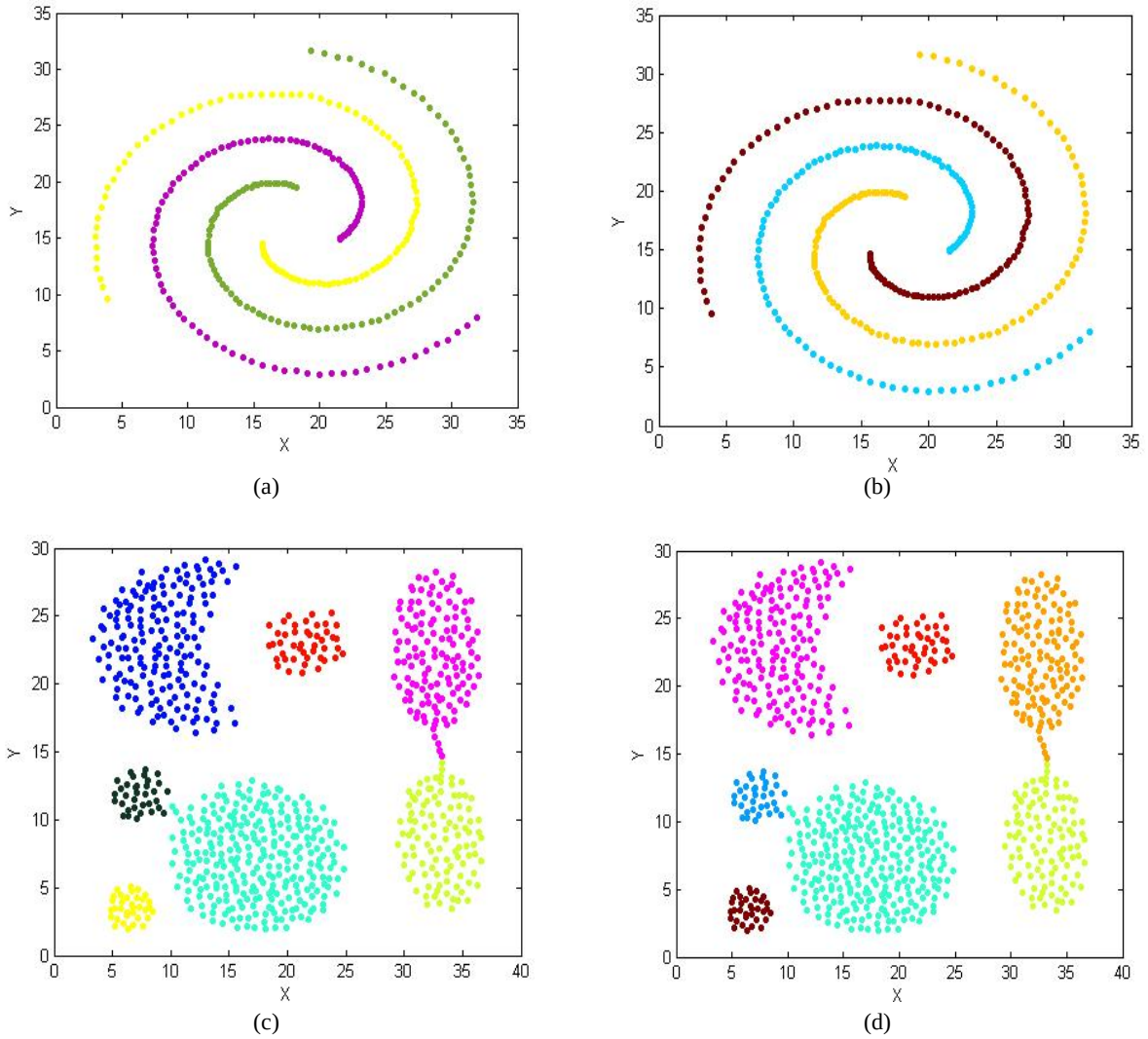
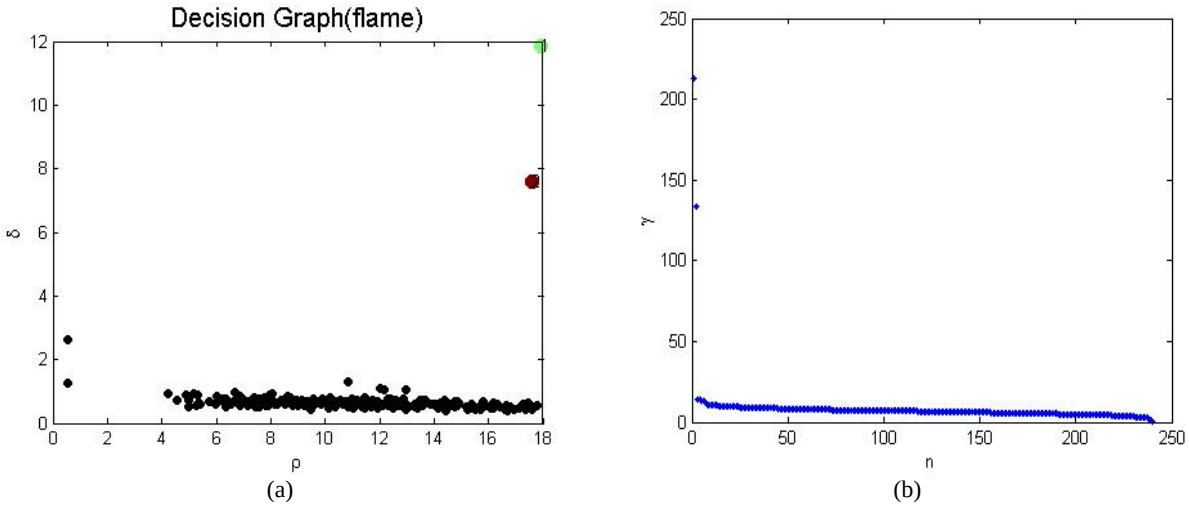
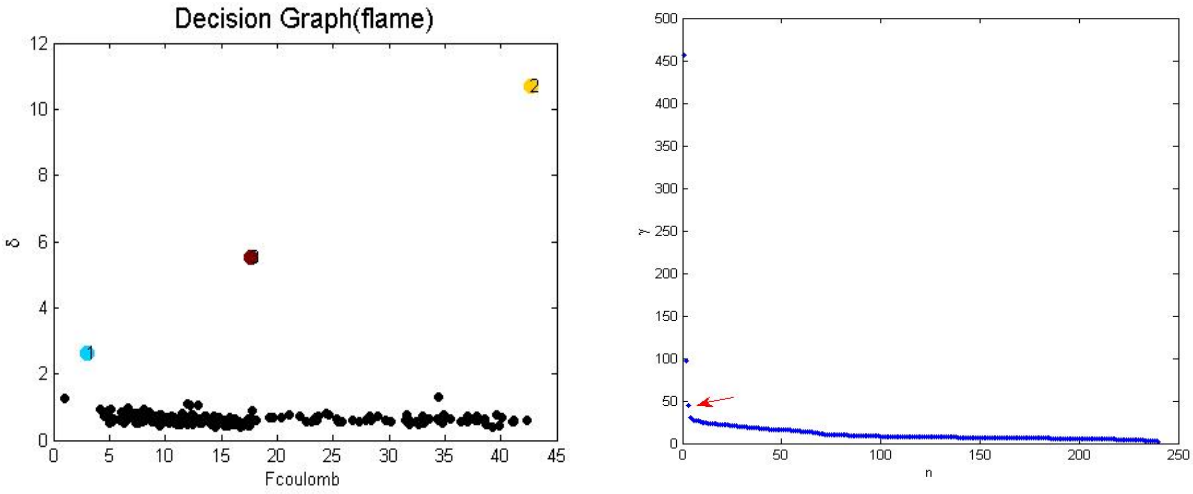


Fig. 4. Clustering results of the DPC and CDPC: (a) clustering result on *Spiral* by DPC (cluster1 is green,cluster 2 is red, cluster 3 is purple), (b) clustering result on *Spiral* by CDPC (cluster1 is blue,cluster 2 is yellow, cluster 3 is brown), (c) clustering result on *Aggregation* by DPC (cluster1 is blue,cluster 2 is yellow, cluste r 3 is purple, cluster 4 is red,cluster 5 is black, cluster 6 is light yellow, cluster 7 is light blue), (d) clustering result on *Aggregation* by CDPC (cluster1 is blue,cluster 2 is yellow, cluster 3 is purple, cluster 4 is red,cluster 5 is brown, cluster 6 is light yellow, cluster 7 is light blue)





(c) (d)

Fig. 5. Outliers identification on *Flame* dataset: (a) decision graph of DPC, (b) γ graph of DPC, (c) decision graph of CDPC, (d) γ graph of CDPC

3.5. Application of the analysis to air quality of Changchun

For analyzing the construction of the smart environment in Changchun during the development and construction of the smart city. And then evaluate the construction of the intelligent city in Changchun. In this paper, air quality data from 2015 to 2019 were obtained from Changchun air quality monitoring bureau. Air quality was analyzed by AQI, PM2.5, PM10, SO_2, NO_2 and O_3.

In this paper, air quality is divided into five grades of "excellent", "good", "light pollution", "moderate pollution" and "heavy pollution", and marked as "1", "2", "3", "4" and "5" respectively. Through data preprocessing, data with values of 0 and larger values in the data were eliminated to avoid affecting the overall clustering results. In this way, the relationship between air quality and the construction of a smart city in Changchun in the past five years can be better analyzed. The dataset information after pretreatment in this paper is shown in Table 7.

3.6. Air quality clustering results analysis

By clustering the air quality data of Changchun from 2015 to 2019 by CDPC algorithm, the clustering results shown in Table 8 are obtained. The air quality data of Changchun from 2015 to 2019 were clustered by the CDPC algorithm, and the clustering results are shown in Table 8. We classify the clustering results into five levels, namely "excellent", "good", "light pollution", "moderate pollutin" and "heavy pollution". The data in the table are respectively the number of days of air quality level in each year.

As can be seen from the Table 8, except for the large proportion of light, moderate and heavy air pollution days in 2015, the remaining years are relatively small. This shows that the air quality in the last five years is in a good state on the whole. It can be seen from Fig. 6, the days of air quality in "light pollution", "moderate pollution" and "heavy pollution" are in the overall downward trend. Air quality in the "good" days are on the rise. Air quality in the "excellent" days between 2015 and 2017 showed an upward trend, but there was a downward trend between 2017 and 2019 years. Even in 2019, the number of days in which air quality was "excellent" fell to almost the same level as in 2015. It can be seen from this that in the process of developing and building a smart city, Changchun has better control over air pollution and good air quality, but it is not strict to control the excellent air quality.

On the whole, the construction of the smart environment is in good condition, so we can know that the construction of the smart city in Changchun is more successful. Through the analysis of the above results, at the macro level, we will put forward the following suggestions for the construction of smart environment in Changchun in the process of building smart city: It is hoped that the Changchun municipal government can strengthen the control of the situation that the air quality is "excellent". Meanwhile, maintain the growth trend of the air quality is "good", increase the proportion of the excellent air quality, strive to eliminate the air pollution, and then accelerate the construction speed of the smart city in Changchun.

At the micro level, through the clustering analysis of Changchun city air quality in the past five years, this paper gives some reasonable suggestions for Changchun to build a smart city: (1) reducing the development of heavy industry and strengthening environmental governance; (2) increasing the investment and development of Changchun tertiary industry and reducing the mining and use of coal; (3) Changchun city is a cold area, so it should invest more in the research and development of new energy and make use of new energy to heat the city; (4) increasing the urban green building area, reduce the burning of straw, etc.

Table 7. Air quality data of Changchun from 2015 to 2019

Date	*AQI*	*PM2.5*	*PM10*	*SO_2*	*NO_2*	*O_3*
2015.1.01	82	60	83	97	34	43
2015.1.02	94	70	91	83	36	48
2015.1.03	58	42	64	86	37	47
2015.1.04	204	154	187	99	68	41
2015.1.05	99	59	149	70	38	49
2015.1.06	61	30	72	69	31	55
2015.1.07	103	77	115	85	49	44
2015.1.08	132	101	142	82	54	46
2015.1.09	159	121	160	118	66	39
2015.1.10	185	140	186	146	71	36
2015.1.11	174	132	167	104	59	48
2015.1.12	155	118	146	82	52	54
2015.1.13	121	92	111	93	52	55
2015.1.14	253	203	249	120	67	27
2015.1.15	140	107	139	107	53	50
2015.1.16	85	63	86	91	41	55
2015.1.17	66	47	82	67	38	57
2015.1.18	76	56	77	81	42	53
2015.1.19	143	110	133	98	57	62
2015.1.20	124	94	118	86	55	65
...	...	...	...	...	...	...
2019.12.12	52	36	47	16	34	48
2019.12.13	60	43	50	19	41	38
2019.12.14	87	64	73	17	48	31
2019.12.15	44	30	44	16	33	39
2019.12.16	46	32	36	15	33	42
2019.12.17	60	43	42	14	33	46
2019.12.18	73	53	58	19	47	38
2019.12.19	104	78	76	20	50	38
2019.12.20	58	41	47	21	42	45
2019.12.21	67	48	52	19	41	48
2019.12.22	105	79	79	27	58	33
2019.12.23	269	219	207	34	80	30
2019.12.24	193	145	150	26	63	44
2019.12.25	195	146	141	28	61	51
2019.12.26	279	229	221	30	81	30
2019.12.27	168	127	138	26	57	56
2019.12.28	195	146	155	38	72	27
2019.12.29	145	111	129	29	48	36
2019.12.30	69	50	59	23	34	49
2019.12.31	55	39	50	24	34	47

Table 8. Clustering results of air quality in 2015-2019

Year	*Air quality indicators*				
	Excellent	*Good*	*Light pollution*	*Moderate pollution*	*Heavy pollution*
2015	52	147	93	22	50
2016	60	215	61	22	9
2017	65	236	35	21	6
2018	125	208	12	0	11
2019	127	150	38	29	21

4. Conclusions

In this paper, we propose an analysis of air quality with density peak clustering algorithm and coulomb force theory (CDPC). We can use the proposed cluster model to perform analysis on air quality in the process of building a smart city in Changchun. This algorithm uses Coulomb force theory to calculate the similarity between data points, further calculate the weighted sum of each data point to the similarity of other data points in the dataset and then update the decision diagram.

Meanwhile, the calculation method of parameter γ value is further improved to update the γ graph with the identification of low density data points. Therefore, CDPC algorithm can be used to analyze the air quality data of changchun in recent five years, so as to produce better decision graph and γ graph to better identify the clustering center and abnormal points of data.

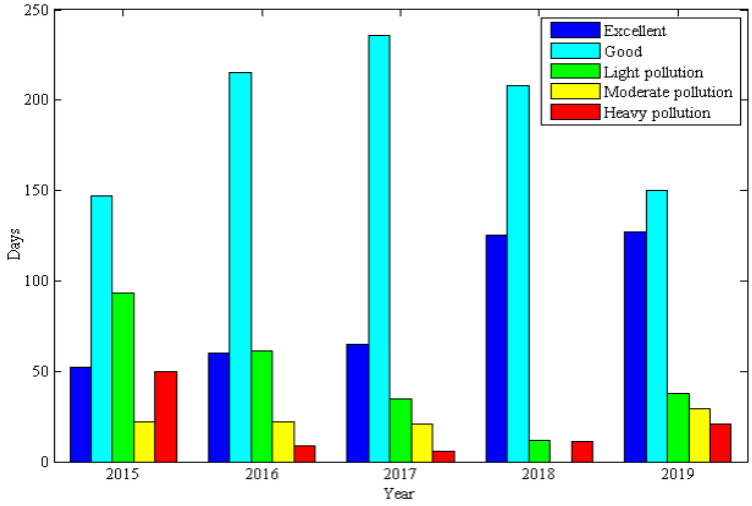


Fig. 6. The bar chart of air quality clustering results in 2015-2019

The experimental results on synthetic and real-world datasets show that the CDPC algorithm performs better than DPC and other algorithms. In the process of testing algorithm performance, we only compared the three most advanced algorithms. If the comparison can be added to more algorithms and data sets, the clustering performance of the algorithm can be better illustrated. The CDPC algorithm provides a new technology support for the construction of air quality in smart cities, and has a certain use value to provide reliable intellectual support for the innovation drive of smart cities.

We can also have a good conclusion that the air quality construction of smart city of Changchun in the past five years, more and more days have good or excellent air quality. In the future, we will use the optimization algorithm to optimize the parameters of the CDPC algorithm to establish a more reasonable and effective analysis model, so as to quickly and accurately obtain the optimal clustering results of urban air indicators and urban water pollution and other data.

Acknowledgements

The research was supported by the National Science Foundation of China under Grant No. 61572225 and 61472049, the foundation of JiLin Province Education Department under Grant No. JJKH20190724KJ, the Jilin Provincial Science & Technology Department Foundation under Grant No. 20190302071GX and 20200201164JC, the Development and Reform Commission Foundation of Jilin Province under Grant No. 2019C05311.

References

Abirami S., Chitra P., (2019), *Real Time Twitter-Based Disaster Response System for Indian Scenarios*, 26th Int. Conf. on High Performance Computing, Data and Analytics Workshop (HiPCW), 82-86.

Afify H.M., Mohammed K.K., Hassanien A.E., (2020), Multi-images recognition of breast cancer histopathological via probabilistic neural network approach, *Journal of System and Management Sciences*, **1**, 53-68.

Afrasiabi M., Roethlin M., Klippel H., Wegener K., (2019), Meshfree simulation of metal cutting: an updated Lagrangian approach with dynamic refinement, *International Journal of Mechanical Sciences*, **160**, 451-466.

Afrasiabi M., Meier L., Röthlin M. Klippel H. Wegener K., (2020), GPU-accelerated meshfree simulations for parameter identification of a friction model in metal machining, *International Journal of Mechanical Sciences*, **176**, 105571, https://doi.org/10.1016/j.ijmecsci.2020.105571.

Altintas Y., Ber A., (2001), Manufacturing automation: metal cutting mechanics, machine tool vibrations, and CNC design, *Applied Mechanics Reviews*, **54**, B84, https://doi.org/10.1115/1.1399383.

Amruthnath N., Gupta T., (2018), *Fault Class Prediction in Unsupervised Learning using Model Based Clustering Approach*, In 2018 Int. Conf. on Information and Computer Technologies (ICICT), 5-12.

Boncescu C., Robescu L.D., (2017), Air dispersion modelling and simulation in aeration tanks, *Environmental Engineering and Management Journal*, **16**, 1049-1054.

Bradley P., (2019), Methodology for the sequence analysis of building stocks, *Building Research and Information*, **47**, 141-155.

Buczak A., Guven E., (2015), A survey of data mining and machine learning methods for cyber security intrusion detection, *IEEE Communications Surveys & Tutorials*, **18**, 1153-1176.

Cai Q., Gong M., Ma L., Ruan S., Yuan F., Jiao L., (2015), Greedy discrete particle swarm optimization for large-scale social network clustering, *Information Sciences*, **316**, 503-516.

Chattopadhyay A.K., Chollet L., Hintermann H.E., (1991), On performance of brazed bonded monolayer diamond grinding wheel, *CIRP Annals - Manufacturing Technology*, **40**, 347-350.

Chen M., Li L., Wang B., Cheng, J., Pan L., Chen X., (2016), Effectively clustering by finding density backbone based-on knn, *Pattern Recognition*, **60**, 486-498.

Dai C., Li Y., Alavi M., Cai Y., Sun W., Huang G., (2019), A support vector regression and monte carlo simulation-based interval two-stage programming for environmental systems planning in Beijing, *Environmental Engineering and Management Journal*, **18**, 329-348.

Dragomir E., (2019), A multi-agent system for monitoring and analyzing the air quality index, *Environmental Engineering and Management Journal*, **18**, 147-157.

Dutta P.K., Banerjee S., (2019), Monitoring of aerosol and other particulate matter in air using aerial monitored sensors and real time data monitoring and processing, *Journal of System and Management Sciences*, **9**, 104-113.

Fan Y., Zhao M.Y., Ma L., Zhao L.Y., (2016), Research on the accessibility of urban green space based on road network- a case study of the park green space in city proper of Nanjing, *Journal of Forest & Environmental Science*, **32**, 1-9.

Florea A., Lorint C., Danciu C., (2019), Particulate matters generated by caprisoara tailing pond and their impact on air quality, *Environmental Engineering and Management Journal*, **18**, 803-810.

Frey B., Dueck D., (2007), Clustering by passing messages between data points, *Science*, **315**, 972-976.

Gaur A., Scotney B., Parr G., Mcclean S., (2015), Smart city architecture and its applications based on iot, *Procedia Computer Science*, **52**, 1089-1094.

Graiia A., Gago G., (2018), Environmental analysisof flood risk in urban planning: a case study in las quemadillas, Cordoba, span, *Environmental Engineering and Management Journal*, **17**, 2527-2536.

Hajarolasvadi N., Demirel H., (2019), 3D CNN-based speech emotion recognition using K-means clustering and spectrograms, *Entropy*, **21**, 479, https://doi.org/10.3390/e21050479.

Hartigan J.A., Wong M.A., (1979), A k-means clustering algorithm, *Applied Statistics*, **28**, 100-108.

Kilic B., Ipek O., Sahin A., (2018), A comparative computational intelligence approach for heat transfer analysis of corrugated plate heat exchangers, *Environmental Engineering and Management Journal*, **17**, 1831-1840.

Labianca G., Gisi S., Notarnicola M., (2018), Assessing the correlation between contamination sources and environmental quality of marine sediments using multivariate analysis, *Environmental Engineering and Management Journal*, **17**, 2391-2399.

Lei N., Huang Y., Yan L., Wang Z., (2017), Evolution pattern of engineered road turfy soil examined by inversion and analytic hierarchy process, *Environmental Engineering and Management Journal*, **16**, 2173-2180.

Li Q., Han J., Marusic S., Lu L., (2019), Development of a hybrid app-based survey methodology for evaluating the real-time indoor environmental quality in buildings, *Journal of System and Management Sciences*, **9**, 81-103.

Mehmood R., Zhang G., Bie R., Dawood H., Ahmad H., (2016), Clustering by fast search and find of density peaks via heat diffusion, *Neurocomputing*, **208**, 210-217.

Năstase G., Serban A., (2019), Experimental study on CO_2 capture in a residential space, *Environmental Engineering and Management Journal*, **18**, 1001-1011.

Palm J., Ellegard K., Hellgren M., (2018), A cluster analysis of energy-consuming activities in everyday life, *Building Research and Information*, **46**, 96-113.

Parmar M., Wang D., Zhang X., Tan A. H., Miao C., Jiang J., Zhou Y., (2019), FREDPC: A feasible residual error-based density peak clustering algorithm with the fragment merging strategy, *IEEE Access*, **7**, 89789-89804.

Parmar M., Wang D., Zhang X., Tan A.H., Miao C., Jiang J., Zhou Y., (2018), REDPC: a residual error-based density peak clustering algorithm, *Neurocomputing*, **348**, 82-96.

Pejic V., Cedilnik M., Lisec A., (2017), Impact on the environment of industrial packaging waste transport, *Environmental Engineering and Management Journal*, **16**, 1155-1160.

Pop D., Micle V., Sur L., (2018), Optimizing the process of depollution through thermal absorption of soils contaminated with crude oil, *Environmental Engineering and Management Journal*, **17**, 2619-2626.

Rodriguez A., Laio A., (2014), Clustering by fast search and find of density peaks, *Science*, **344**, 1492-1496.

Role F., Nadif M., (2014), Beyond cluster labeling: semantic interpretation of clusters' contents using a graph representation, *Knowledge-Based Systems*, **56**, 141-155.

Sanchez-Fernandez R., Iniesta-Bonillo M.A., Cervera-Taulet A., (2016), Environmental sustainability in the mediterranean destinations: a latent class segmentation analysis, *Environmental Engineering and Management Journal*, **15**, 1501-1510.

Shi B., Han L., Yan H., (2018), Adaptive clustering algorithm based on knn and density, *Pattern Recognition Letters*, **104**, 37-44.

Wang L., Zhou W., Wang H., Parmar M., Han X., (2019), A novel density peaks clustering halo node assignment method based on k-nearest neighbor theory, *IEEE Access*, **7**, 174380-174390.

Wegener K., Afrasiabi M., Klippel H., Rthlin M., (2019), Meshless single grain cutting simulations on the GPU, *International Journal of Mechatronics and Manufacturing Systems*, **12**, 272-279.

Wei W., Jian L., Hong R., Wang H., Ming X., (2018), Efficient k-nearest neighbor classification over semantically secure hybrid encrypted cloud database, *IEEE Access*, **6**, 41771-41784.

Wenge R., Zhang X., Dave C., Chao L., Hao S., (2014), Smart city architecture: a technology guide for implementation and design challenges, *Communications China*, **11**, 56-69.

Zhang C., Ni M., Yin H., Qiu K., (2018), Developed density peaks clustering with support vector data description for access network intrusion detection, *IEEE Access*, **6**, 46356-46362.