TMsDP: two-stage density peak clustering based on multi-strategy optimization

Jie Ma
School of Business and Management, Jilin University, Changchun, China and Information Resource Research Center, Jilin University, Changchun, China
Zhiyuan Hao
School of Business and Management, Jilin University, Changchun, China, and
Mo Hu
Department of Network and New Media, School of Journalism and Communication, Nanjing Normal University, Nanjing, China

Received 25 August 2021
Revised 15 December 2021
16 February 2022
18 March 2022
26 April 2022
30 May 2022
Accepted 6 June 2022

Abstract

Purpose – The density peak clustering algorithm (DP) is proposed to identify cluster centers by two parameters, i.e. ρ value (local density) and δ value (the distance between a point and another point with a higher ρ value). According to the center-identifying principle of the DP, the potential cluster centers should have a higher ρ value and a higher δ value than other points. However, this principle may limit the DP from identifying some categories with multi-centers or the centers in lower-density regions. In addition, the improper assignment strategy of the DP could cause a wrong assignment result for the non-center points. This paper aims to address the aforementioned issues and improve the clustering performance of the DP.

Design/methodology/approach – First, to identify as many potential cluster centers as possible, the authors construct a point-domain by introducing the pinhole imaging strategy to extend the searching range of the potential cluster centers. Second, they design different novel calculation methods for calculating the domain distance, point-domain density and domain similarity. Third, they adopt domain similarity to achieve the domain merging process and optimize the final clustering results.

Findings – The experimental results on analyzing 12 synthetic data sets and 12 real-world data sets show that two-stage density peak clustering based on multi-strategy optimization (TMsDP) outperforms the DP and other state-of-the-art algorithms.

Originality/value – The authors propose a novel DP-based clustering method, i.e. TMsDP, and transform the relationship between points into that between domains to ultimately further optimize the clustering performance of the DP.

Keywords Data clustering, Density peak clustering algorithm, Merging strategy, Pinhole imaging strategy, Point-domain, Point-domain similarity

Paper type Research paper

1. Introduction

As a powerful machine learning method in the data mining field, the clustering strategy has a broad research prospect in effectively identifying the internal structure of data samples, such as mining spatiotemporal co-location events in trajectory data sets (Ansari *et al.*, 2021),

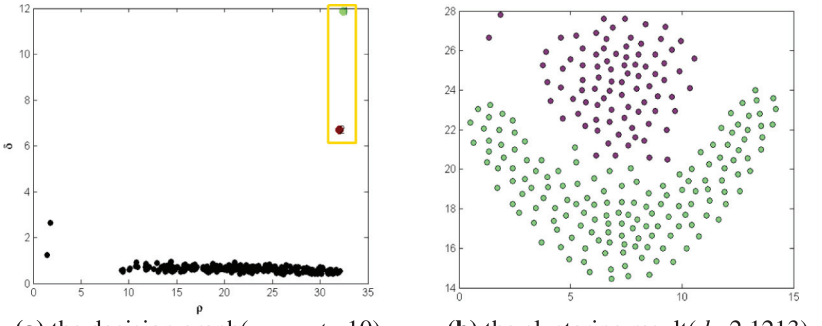
Funding: This research was funded by the Major Project of the National Social Science Foundation of China, grant number 20&ZD125, and the National Natural Science Foundation of Jilin Province, grant number 20210101480JC.

Data Technologies and Applications
Emerald Publishing Limited
2514-9288
DOI 10.1108/DTA-08-2021-0222

conducting customer segmentation (Li *et al.*, 2021) and detecting CT scan images (Singh and Bose, 2021). In addition, as an important branch in clustering algorithms, density-based clustering has been concerned and studied by a large number of researchers. Various density-based clustering methods have been proposed and widely utilized in different fields to date, such as fault recognition in wind turbines with a density-based clustering algorithm (Luo *et al.*, 2021) and risk assessment on railway investment with an improved density-based approach (Guo *et al.*, 2021). In 2014, the density peak clustering algorithm (DP) was proposed by American scholars in *Science* (Rodriguez and Laio, 2014). Since its establishment, the DP has been studied and applied by a large number of investigators in various fields, such as text clustering (Jo, 2020), medical analysis (Medeghri and Sabeur, 2021) and image recognition (Wang *et al.*, 2019, 2020; He *et al.*, 2021). Specifically, there are three significant parameters, i.e. the d_c value (cutoff distance), the ρ value (local density) and the δ value (the distance between a point and another point with a high ρ value), and an important principle in the original DP, i.e. the cluster centers should have a higher ρ value and a higher δ value than other points (Abbas *et al.*, 2021; Wang *et al.*, 2021). Although the DP has better clustering performance than other traditional density-based clustering algorithms, it still contains a critical limitation, i.e. the higher ρ value and the higher δ value could not accurately reflect whether a point is a cluster center.

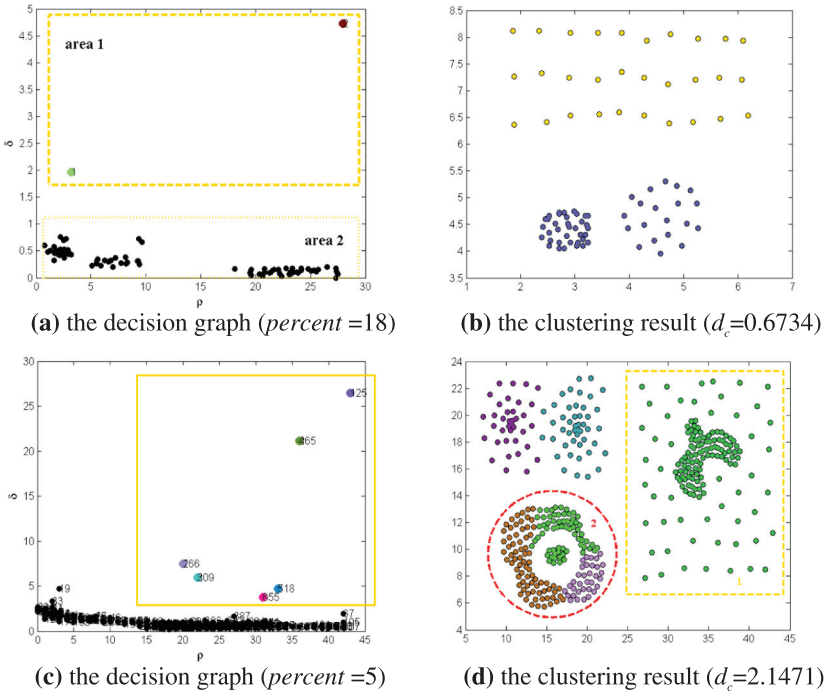
To give a concrete example, two different situations are discussed in this paper; Figures 1 and 2 show situation 1 and situation 2, respectively. For situation 1, it is clearly shown in Figure 1(a) that the data set *flame* should have two different categories, and the two potential cluster centers both have a higher ρ value and a higher δ value than other points. Actually, Figure 1(b) shows that the DP could indeed obtain a clustering result which is close to the natural category. The combination results of Figure 1 seem to demonstrate that the aforementioned principle about the ρ value, the δ value and the cluster centers is reasonable. However, situation 2 illustrates that the principle is unreasonable yet. As shown in Figure 2, the DP could just obtain the inferior clustering results when analyzing the data sets *D1* and *compound*, which are not consistent with the principle mentioned above.

Obviously, the DP could identify only two potential cluster centers for data set *D1* (it has three different natural categories), while it could just identify six wrong clusters for the data set *compound* (it has six different natural categories). The difference between situation 1 and situation 2 reflects the following deficiencies of the DP: (1)



(a) the decision graph(*percent* =10) **(b)** the clustering result(d_c=2.1213)

Figure 1.
The selection of cluster center points (they are the data points in the oblong) and the clustering result of *flame*



(a) the decision graph (*percent* =18) **(b)** the clustering result (d_c=0.6734)

(c) the decision graph (*percent* =5) **(d)** the clustering result (d_c=2.1471)

Figure 2.
The selection of cluster center points (they are the data points in the oblong) and the clustering result of *D1* and *compound*

the DP could not detect the accurate density peak points when analyzing some data sample with multi-density or variable density; (2) the DP is challenging to identify some data samples with non-single cluster center accurately and (3) the drawback of the original density calculation method and the improper assignment of the non-central points ultimately affect the overall clustering performance.

To address the aforementioned issues, the authors develop an enhanced DP-based clustering method, i.e. two-stage density peak clustering based on multi-strategy optimization (TMsDP), to further optimize the clustering performance of the DP. The main contributions and innovations of the TMsDP are as follows:

(1) Point-domain is constructed by introducing the pinhole imaging strategy to confirm the search scope of potential centers. The point-domain improves the clustering efficiency by transforming the relationship between points into that between domains.

(2) Point-domain density is determined to measure the distribution of points in a point-domain, while the domain distance is calculated by introducing the Hausdorff distance to improve the clustering accuracy.

(3) Domain similarity is proposed to achieve the domain merging process. In a data space, the domain similarity between point-domains is higher, and it is more likely to merge with each other.

The details of TMsDP are discussed in this study. Specifically, Section 2 presents a brief introduction of the DP, Section 3 describes the specific technical details of the proposed TMsDP, Section 4 analyzes the experimental results with different data sets to verify the clustering performance of the TMsDP and Section 5 summarizes this study by discussing the results and future areas for potential investigations.

2. Density peak clustering

2.1 Preparation

In the original DP (Rodriguez and Laio, 2014), d_c is set as the manual parameter, which denotes the appropriate position in an ascending distance sequence, and the definition processes are shown as follows (assuming Sample = $\{s_1, s_2, s_3, \ldots, s_n\}$):

$$\text{position} = \text{round}(N \times \text{percent}/100), \tag{1}$$

$$\text{disorder} = \text{sort}\big(\text{dis}(s_i, s_j)\big), \tag{2}$$

$$d_c = \text{disorder}(\text{position}), \tag{3}$$

where N indicates the manual inputting value and dis (s_i, s_j) represents the distance between the point s_i and the point s_j. Rodriguez and Laio (2014) define that ρ_i denotes the number of points in a circle with the point s_i as the center and the d_c value as the radius, and the process is shown as follows:

$$\rho_i = \sum_{s_i, s_j \in \text{Sample}}^{n} \chi\big(\text{dis}(s_i, s_j) - d_c\big), \tag{4}$$

where the function $\chi(o)$ is equal to 1 or 0. If the variable o is greater than 0, $\chi(o)$ is equal to 0. Otherwise, $\chi(o)$ is equal to 1. In addition, the calculation process of the δ value is shown as follows:

$$\delta_{s_i} = \min_{\substack{s_i, s_j \in \text{Sample} \\ \rho_{s_i} < \rho_{s_j}}} \text{dis}(s_i, s_j). \tag{5}$$

2.2 Related work

Based on the aforementioned contents, it is clear that the δ value and the ρ value are limited by the threshold parameter, i.e. d_c value, and utilizing different d_c values could even provide completely different clustering results when analyzing the same data set (Hou *et al.*, 2020; Lu *et al.*, 2020; Jangra and Toshniwal, 2020; Flores and Garza, 2020; Zhu *et al.*, 2020). For addressing the threshold parameter selection issue, Xu *et al.* (2020) proposed a robust DP with density-sensitive similarity to find accurate cluster centers automatically and reduce the effect of the d_c value selection on clustering results. D'Errico *et al.* (2021) provided a feasible approach for solving the classification problem of data with different shapes and distributions in order to avoid the drawback of the d_c value. Ding *et al.* (2018) developed an automatic DP based on a generalized extreme value distribution. At the same time,

the assignment strategy of non-cluster center points often affects the final clustering results. To address the assignment issues, Jiang *et al.* (2019) introduced logistic distribution theory and K-nearest neighbor (kNN) theory into DP. Xu *et al.* (2021) designed a novel sparse search strategy to measure the similarity between the nearest neighbors of each point. Yu *et al.* (2021) proposed a three-way density peak clustering method based on evidence theory. Seyedi *et al.* (2019) utilized a graph-based label propagation to assign labels to remaining points and proposed the dynamic graph-based label propagation for density peak clustering. Apart from the d_c value selection issue and the non-center point assignment issue, it is challenging to identify the potential centers in low-density regions and to analyze data with varying density distributions using the DP. For solving these issues, Yan *et al.* (2021) proposed a rotation-DPeak algorithm to solve the imbalanced data and data with sparse regions. Liu *et al.* (2018) presented three novel definitions, i.e. shared nearest neighbor (SNN) similarity, local density ρ and the distance from the nearest larger density point δ, and proposed an SNN-based clustering by fast search and find of density peaks algorithm. Du *et al.* (2019) provided a new option based on the sensitivity of the local density, redefined the δ value and redesigned the assignation strategy based on a new density-adaptive metric, while Chen and Yu (2021) proposed a domain-adaptive density clustering algorithm, which consisted of three steps: domain-adaptive density measurement, cluster center self-identification and cluster self-ensemble. In addition, the DP could not effectively identify the noise data and outliers and it has high computational complexity when solving large-scale data. For avoiding the drawbacks and accelerating the DP, Parmar *et al.* (2019)

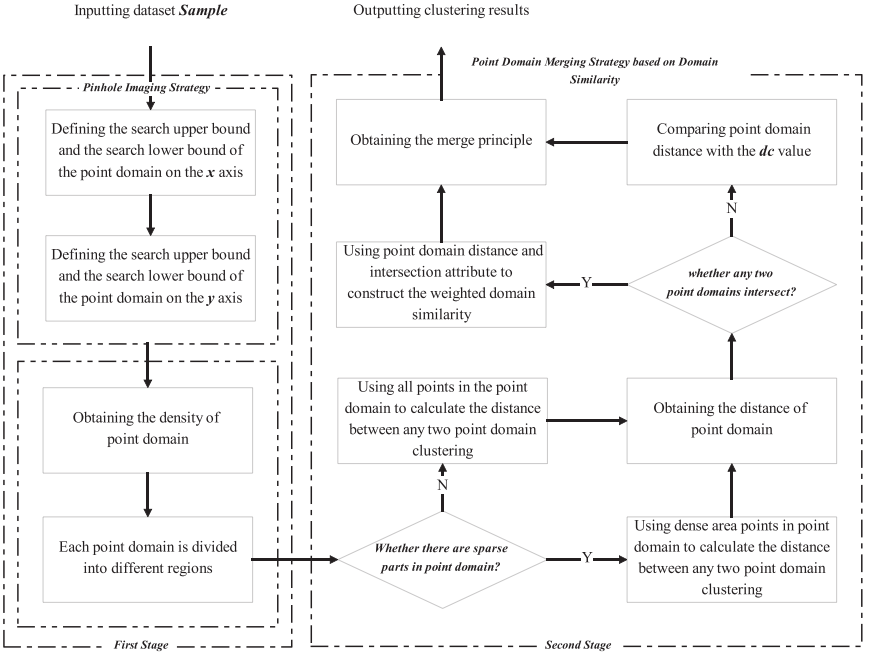


Figure 3.
The framework of the TMsDP algorithm

proposed a residual error-based DP to better identify overlapping clusters. Wang *et al.* (2020) combined the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and proposed a systematic density-based clustering method using anchor points. Chen *et al.* (2020) replaced density with kNN density and proposed a fast DP, i.e. FastDPeak. Fan *et al.* (2019) proposed a fast algorithm that accelerates the density computation about 50 times over the original one.

To optimize the performance of original DP, the authors delineate a novel DP-based clustering method in this paper. In the novel method, they propose four main significant strategies, i.e. point-domain, point-domain density, domain distance and domain similarity. The framework of TMsDP is shown in Figure 3.

3. The proposed clustering method

3.1 Point-domain strategy based on pinhole imaging theory

In order to explore the potential cluster centers in low-density regions, the proposed TMsDP constructs the point-domain by introducing the pinhole imaging theory. Pinhole imaging is a physics phenomenon where a light source passes through a pinhole and its inverted image will be formed on a screen (Long *et al.*, 2021). Inspired by the related literature (Long *et al.*, 2021; Lu *et al.*, 2018), this paper introduces the pinhole imaging theory into the search strategy of potential cluster centers, which can help the TMsDP to expand the range of center exploration. Assume that the point $S_i(x_{s_i}, y_{s_i})$ is a potential cluster center in Sample and

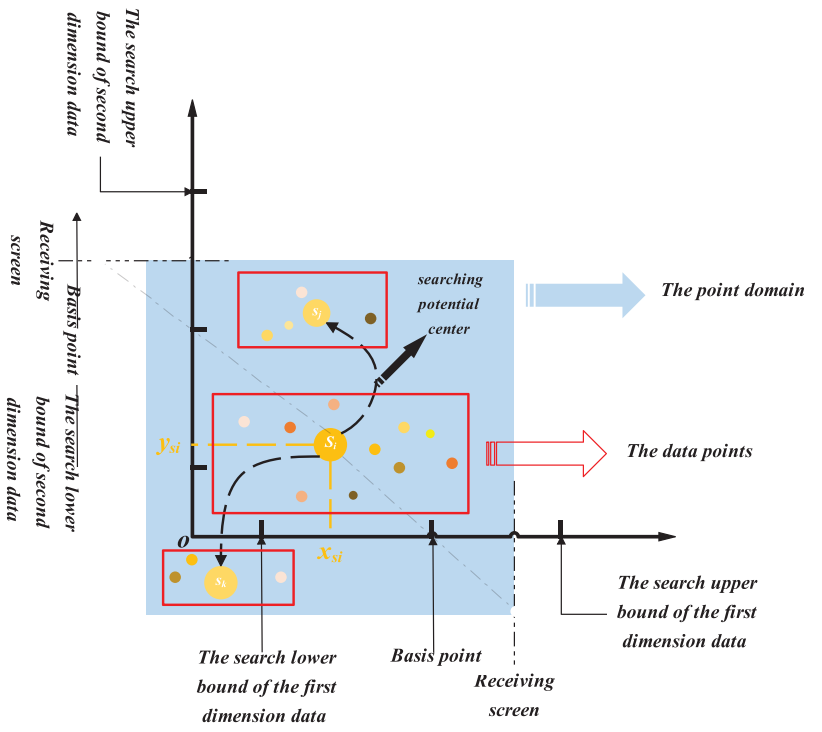


Figure 4.
The whole process of constructing point-domains by utilizing the pinhole imaging strategy

other potential cluster centers, like the point $S_k(x_{s_k}, y_{s_k})$ and the point $S_j(x_{s_j}, y_{s_j})$, may also exist in the same point-domain. If we want to construct a point-domain for data S_i, we should comprehend the preliminary definitions which are shown in Figure 4 (this paper mainly utilizes two-dimensional (2D) data as the examples to explain the following preliminary definitions).

Definition 1. (upper bound for searching of the first dimension data). As a rule of thumb, if the point $S_i(x_{s_i}, y_{s_i})$ is a cluster center, the ρ value of other potential cluster centers should be close to ρ_{s_i}. Therefore, this study should determine a searching range to explore these potential cluster centers. The first exploration concept is the search upper set (SUS); the SUS is a point-set where the points have a higher first dimension data value and a higher ρ value than the point S_i and they are nearly closest to the point S_i. Based on the SUS, the calculation processes of upper bound for searching of the first dimension data are shown as follows:

$$\mathrm{SUS}\, x = \left\{ S_n \in \mathrm{Sample} | \rho_{S_n} > \rho_{S_i},\ x_{S_n} > x_{S_i},\ \text{nearest neighbor}(S_n, S_i) \right\}, \tag{6}$$

$$\text{upper bound for searching} = \left\{ S_n \in \mathrm{SUS}\, x | x_{S_n} - x_{S_i} = \tau d_c \right\}. \tag{7}$$

Definition 2. (lower bound for searching of the first dimension data). To maximize the odds of finding more potential cluster centers, this study should consider a situation where some potential centers may exist in a region with a slightly lower ρ value than the point S_i. Therefore, the second exploration concept is the search lower set (SLS); the SLS is also a point-set where the points have a lower first dimension data value than the point S_i, and the ρ values of these points are much closer to ρ_{s_i}. Based on the abovementioned contents, the calculation processes of lower bound for searching of the first dimension data are shown as follows:

$$\mathrm{SLS}\, x = \left\{ S_m \in \mathrm{Sample} | \rho_{S_m} < \rho_{S_i},\ x_{S_m} < x_{S_i},\ \text{nearest neighbor}(S_m, S_i) \right\}, \tag{8}$$

$$\text{lower bound for searching} = \left\{ S_m \in \mathrm{SLS}\, x | \max_{S_m \neq S_i} (x_{S_m}) \right\}. \tag{9}$$

Definition 3. (basis point in the first dimension). In this paper, the basis point in the first dimension denotes a middle value between upper bound for searching of the first dimension data and lower bound for searching of the first dimension data. The definition is shown as follows:

$$\text{basis point}\, x = \frac{x_{S_m} + x_{S_n}}{2}. \tag{10}$$

Definition 4. (upper bound for searching of the second dimension data). The process of upper bound for searching of the second dimension data is similar to Definition 1, and the definitions are shown as follows:

$$\mathrm{SUS}\,y = \left\{ S_a \in \mathrm{Sample} | \rho_{S_a} > \rho_{S_i}, y_{S_a} > y_{S_i}, \text{ nearest neighbor}(S_a, S_i) \right\}, \tag{11}$$

$$\text{upper bound for searching} = \left\{ S_a \in \mathrm{SUS}\,y | y_{S_a} - y_{S_i} = \omega d_{\mathrm{c}} \right\}. \tag{12}$$

Definition 5. (lower bound for searching of the second dimension data). The process of lower bound for searching of the second dimension data is similar to Definition 2, and the definitions are shown as follows:

$$\mathrm{SLS}\,y = \left\{ S_b \in \mathrm{Sample} | \rho_{S_b} < \rho_{S_i}, y_{S_b} < y_{S_i}, \text{ nearest neighbor}(S_b, S_i) \right\}, \tag{13}$$

$$\text{lower bound for searching} = \left\{ S_b \in \mathrm{SLS}\,y | \max_{S_b \neq S_i} \left(y_{S_b} \right) \right\}. \tag{14}$$

Definition 6. (basis point in the second dimension). The definition of the basis point in the second dimension is similar to Definition 3, and it is shown as follows:

$$\text{basis point}\, y = \frac{y_{S_a} + y_{S_b}}{2}. \tag{15}$$

For the example shown in Figure 4, it is a point-domain of the point S_i. In the point-domain S_i, the authors set the x-axis value of receiving screen (first dimension) to x_{r}, the y-axis value of receiving screen (first dimension) to y_{r}, the x-axis value of receiving screen (second dimension) to x'_{r} and the y-axis value of receiving screen (second dimension) to y'_{r}. Based on the triangular similarity theory, the relationships between four searching bounds and two basis points are shown as follows:

$$\frac{\frac{x_{S_n} - x_{S_m}}{2} + x_{S_m} - x_{S_i}}{x_{\mathrm{r}} - \left(\frac{x_{S_n} - x_{S_m}}{2} + x_{S_m} \right)} = \frac{y_{S_i}}{y_{\mathrm{r}}} = \psi, \tag{16}$$

$$\frac{\frac{y_{S_a} - y_{S_b}}{2} + y_{S_b} - y_{S_i}}{y'_{\mathrm{r}} - \left(\frac{y_{S_a} - y_{S_b}}{2} + y_{S_b} \right)} = \frac{x_{S_i}}{x'_{\mathrm{r}}} = \xi, \tag{17}$$

where the control thresholds ψ and ξ could be set manually for different clustering demands. According to formula (16) and formula (17), the side values of the point-domain can be obtained as follows:

$$\text{side} = \left\{ \text{side} \in (\text{side}\, x \cap \text{side}\, y) \left| \begin{array}{l} \text{side}\, x = \dfrac{2+2\psi}{\psi}\left(\dfrac{x_{S_m}+x_{S_n}}{2} - x_{S_i}\right) \\ \text{side}\, y = \dfrac{2+2\xi}{\xi}\left(\dfrac{y_{S_a}+y_{S_b}}{2} - y_{S_i}\right) \end{array} \right. \right\}. \tag{18}$$

3.2 Domain merging strategy based on point-domain similarity

Although the TMsDP transforms the relationships between points into that between point-domains, it is still a density-based clustering method. Therefore, how to perform the density analysis on point-domains is a highlight in this section. This paper defines the point-domain density as follows:

Definition 7. (point-domain density). In this paper, the point-domain density denotes the amount of points per unit area of a point-domain (the definition emphasizes the distribution of data points, which has statistical significance). According to the aforementioned contents, the authors could assume a set $D = \{D_1, D_2, D_3, \ldots, D_n\}$, where n indicates the amount of point-domains and D indicates a domain-set which includes all point-domains, and the calculation process of point-domain density is shown as follows (applying the function amount(θ) to calculate the amount of data points in a point-domain):

$$\text{PD}\,\rho_i = \frac{\text{amount}(D_i)}{\left(\frac{2+2\psi}{\psi}\left(\frac{x_{S_m}+x_{S_n}}{2} - x_{S_i}\right)\right)\left(\frac{2+2\xi}{\xi}\left(\frac{y_{S_a}+y_{S_b}}{2} - y_{S_i}\right)\right)}. \tag{19}$$

The point-domain density could show the inner characteristic of a point-domain; moreover, the authors consider the outer characteristics between point-domains. Therefore, this paper constructs a novel distance definition, i.e. domain distance.

Definition 8. (domain distance). Inspired by the literature (Vavpetic and Zagar, 2021; Ryu and Kamata, 2021; Nie *et al.*, 2021), the authors adopt the Hausdorff distance to calculate the domain distance between point-domains. Assume that a point-domain $D1 = \{d1_1, d1_2, d1_3, \ldots, d1_i\}$ and the other point-domain $D2 = \{d2_1, d2_2, d2_3, \ldots, d2_j\}$, where $d1_i$ and $d2_j$ denote the two different points and i and j denote the serial number of data points in $D1$ and $D2$, respectively. The calculation process of the domain distance is shown as follows:

$$\text{domain dis}(D1, D2) = \max_{d1_i \in D1, d2_j \in D2} \min\left(\text{dis}\left(d1_i, d2_j\right)\right). \tag{20}$$

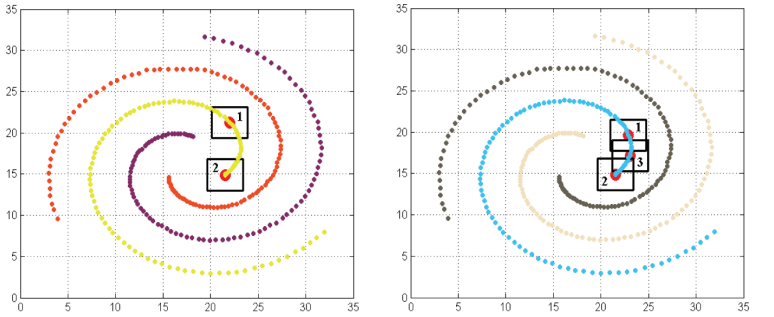
For calculating the domain distance, the authors still need to consider two additional situations: (1) is there an intersection part between the two point-domains? (2) whether the points in these two point-domains are uniformly distributed? The authors take the data set *spiral* as an example to describe these two situations, and the results are shown in Figures 5 and 6.

As shown in Figure 5(a), point-domain 1 and point-domain 2 have no intersection part, which means that the point-domain similarity could take the domain distance as the only calculation criterion. But in Figure 5(b), point-domain 1 and point-domain 3 have an intersection part and there is also an intersection part between point-domain 2 and point-domain 3. When there exist

intersection parts between point-domains, the calculation of point-domain similarity needs to take into account the intersection part, and it is shown as follows:

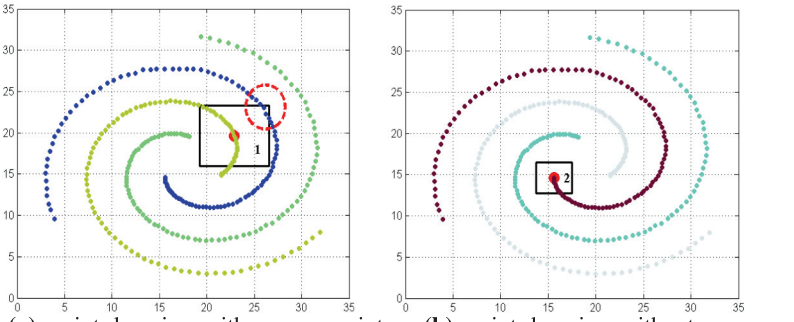
$$\text{Intersection} = \left\{\text{amount}\left(D_i \cap D_j\right) \middle| \left(D_i, D_j \in D\right)\right\}. \tag{21}$$

As shown in Figure 6(a), point-domain 1 has some independent sparse points in the red circle region, and it could be clearly seen that these sparse points deviate from the overall distribution trend of the points in the point-domain. Therefore, the calculation process of point-domain similarity should be performed on the points in the overall distribution trend other than the sparse points. In Figure 6(b), point-domain 2 does not have sparse points and the overall distribution trend of points is relatively stable. Inspired by the literature (Yarinezhad and Hashemi, 2019), the authors propose a strategy to identify the sparse points in this paper. Obviously, the points in the manifold data sets could identify easily whether they are the sparse points. However, for other data sets with different types, the sparse points could not be judged visibly. Assume that a point-domain could be divided into two regions with equal areas and the density values of these two regions are set to ρ_1 and ρ_2, respectively. In addition, the authors assume that ρ_1 is greater than ρ_2, the density value of the whole point-domain is ρ and the difference between ρ_1 and ρ_2 will be compared with the value of



(a) point domains have no intersection **(b)** point domains have intersection

Figure 5.
The case of intersection and non-intersection between point-domains



(a) point-domains with sparse point **(b)** point-domains without sparse

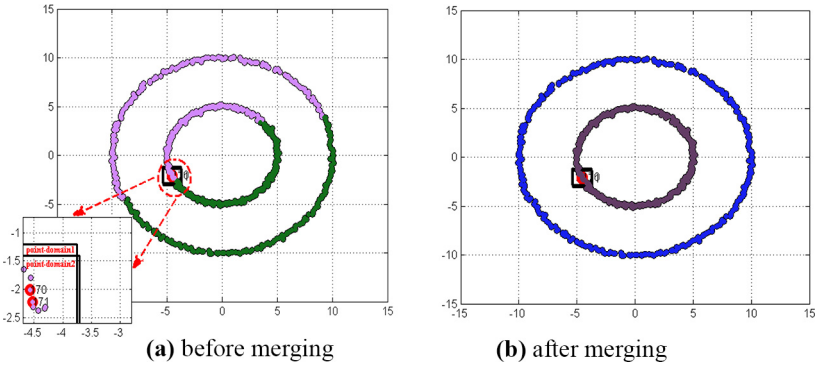
Figure 6.
The case of a point-domain with sparse point distribution and without sparse points

0.8ρ. If the difference between ρ_1 and ρ_2 is greater than 0.8ρ, the points in the region with small density could be identified as the sparse points.

According to the rule of thumb, if there are more intersection parts between two subjects and these two subjects are much nearer, the two subjects are more likely to merge into one. Therefore, the TMsDP adopts the domain distance and the intersection part between two point-domains to calculate the domain similarity. The calculation formula is shown as follows:

$$\underset{D_i,D_j\in\text{Sample}}{\text{sim}} = \begin{cases} \dfrac{\text{amount}(D_i\cap D_j)}{\sqrt{\text{amount}(D_i)}\times\sqrt{\text{amount}(D_j)}} \\ \quad \times exp\left(-\dfrac{\max\limits_{d_{i_m}\in D_i, d_{j_n}\in D_j}\min\left(\text{dis}\left(d_{i_m},d_{j_n}\right)\right)}{\max\limits_{d_{i_m}\in D_i, d_{j_n}\in D_j}\min\left(\text{dis}\left(d_{i_m},d_{j_n}\right)\right)+o(\theta)}\right) & D_i\cap D_j\neq\varnothing \\ \text{intersection sim}i_{D_i,D_j\in\text{Sample}} = \dfrac{\text{amount}(D_i\cap D_j)}{\sqrt{\text{amount}(D_i)}\times\sqrt{\text{amount}(D_j)}} & D_i\cap D_j\neq\varnothing \\ min(\gamma\times\text{intersection sim}i) \\ \quad \times exp\left(-\dfrac{\max\limits_{d_{i_m}\in D_i, d_{j_n}\in D_j}\min\left(\text{dis}\left(d_{i_m},d_{j_n}\right)\right)}{\max\limits_{d_{i_m}\in D_i, d_{j_n}\in D_j}\min\left(\text{dis}\left(d_{i_m},d_{j_n}\right)\right)+o(\theta)}\right) & D_i\cap D_j\neq\varnothing \end{cases} \tag{22}$$

where sim denotes the domain similarity, γ denotes a random parameter with a range of values in (0, 1) and s denotes the adjustment operator which aims to make the value of domain similarity in (0, 1). Considering that the distance value between point-domains with intersection must be smaller than that between point-domains without intersection, the larger the distance value between point-domains is, the smaller the similarity is. Therefore, this paper adds the adjustment operator $o(\theta)$ and the adjustment parameter γ to ensure that the domain similarity between point-domains without intersection is less than that between



(a) before merging **(b)** after merging

Figure 7.
The case of utilizing domain similarity to merge any two different point-domains

point-domains with intersection. Figure 7 shows the merging situation of two point-domains, which takes the data set *2circles* as an example.

In fact, these strategies and methods proposed in this paper increase the impact of the parameters on the clustering result. Apart from the original parameters d_c, the TMsDP adds the parameters τ and ω to determine the exploration range of the potential cluster centers, adds the parameters ψ and ξ to determine the size of the point-domain and the value of the domain density and adds the parameters $o(\theta)$ and γ to determine the domain similarity. Actually, the most significant parameter in the TMsDP is the side value of different point-domains, and the parameters mentioned above are finally utilized to calculate the side value. The side value of point-domains will be shown in the following specific experimental results (in the following experiments, the authors set the side length and side width to equal values in a point-domain). The overall procedures of the TMsDP are shown in Algorithm 1 (Table I).

Table I.
The part core process of the proposed TMsDP

Algorithm 1: TMsDP method

Input: The initial parameter *average percentage of neighbour*, the data set *Sample*, the parameters τ and ω, the parameters ψ and ξ, the parameters $o(\theta)$ and γ
Output: The final clustering result

calculating the distance between any two different data points and obtaining the local density (ρ value) of each data point.
pinhole imaging strategy
  for ***i***=1 to TRow **do** // inputting the τ and ω
    $S1$ = find((ρ>ρ_i) && (x>x_i) && nearest_neighbor(i))
    search upper bound_x = $\{p_1 \in S1 |\ x_{p1}\text{-}x_i\text{=}\tau\text{*}d_c\}$ (see the *Eq*6-*Eq*7)
    $S2$ = find((ρ<ρ_i) && (x<x_i) && nearest_neighbor(i))
    search lower bound_x = $\{p_2 \in S2 |\ max\ (x_{p2})\}$ (see the *Eq*8-*Eq*9)
    $S3$ = find((ρ>ρ_i) && (y>y_i) && nearest_neighbor(i))
    search upper bound_y = $\{p_3 \in S3 |\ y_{p3}\text{-}y_i\text{=}\omega\text{*}d_c\}$ (see the *Eq*11-*Eq*12)
    $S4$ = find((ρ<ρ_i) && (y<y_i) && nearest_neighbor(i))
    search lower bound_y = $\{p_4 \in S4 |\ max\ (y_{p4})\}$ (see the *Eq*13-*Eq*14)
  end
  constructing the point-domain and obtaining the *side* value (see the *Eq*16-*Eq*18) // inputting the ψ and ξ
  calculating strategy
  for ***i***=1 to domain_number-1 **do**
    for ***j***=***i***+1 to domain_number **do**
      domain_density(i) (see the *Eq*19) // obtaining the value of point-domain density
      domain_density(j)
    end
  end
  for ***i***=1 to domain_number-1 **do**
    for ***j***=***i***+1 to domain_number **do**
      if the domain has sparse points **then** // Refer to the judgment principle in the text
        use the other data point to calculate domain distance(i, j)
      else
        use tall data points in point-domain to calculate domain distance(i, j)
      end
      if the point-domain(i) and point-domain(j) have intersection **then**
        combining the intersection and domain distance to calculate similarity
      else
        just utilized the domain distance to calculate domain similarity // inputting the $o(\theta)$ and γ
      end
    end
  end

3.3 Time complexity analysis

For the TMsDP, the time complexity analysis is considered from the following aspects: (1) the time complexity of the point-domain is close to $O(n)$; (2) the time complexity of the calculation about the domain distance is close to $O(n^2)$ and (3) the time complexity of the domain similarity is close to $O(n^2)$. Thus, the time complexity of the TMsDP is close to $O(n^2 + n^2 + n)$, which is close to the original DP (the time complexity of the DP clustering is $O(n^2)$).

4. Experimental results and analysis

To illustrate the performance of the proposed method, this section selects 12 synthetic data sets and 12 real-world data sets as the experiment samples [1]. The 12 synthetic data sets include *2circles, compound, twocirclesnoise2, spiral, pathbase, jain, flame, D1, D2, DS5, skewed* and *unbalance*. The 12 real-world data sets include *thyroid, breast, glass, liver, heart, seeds, zoo, wine, vote, iris, dna* and *msplice*. The specific characteristics of these experiment data sets are shown in Table II. In addition, to further demonstrate the clustering performance of the proposed method, the TMsDP is compared with DP (Rodriguez and Laio, 2014), density peaks clustering based on logistic distribution and gravitation (DPC-LG) (Jiang *et al.*, 2019), DBSCAN (Ester *et al.*, 1996), Affinity Propagation Algorithm (AP) (Frey and Dueck, 2007) and K-means (Jain, 2010). This paper takes the Rand index (RI, the range of values is from −1.0 to 1.0), F-measure (FM, the range of values is from −1.0 to 1.0), Jaccard index (JI, the range of values is from 0 to 1.0) and normalized mutual information (NMI, the range of values is from −1.0 to 1.0) as the evaluation criteria to measure the clustering performance.

Table II. The basic attributions of experiment data sets

Data set type	Order	Data set name	Dimension	Data volume	Real cluster number
Synthetic	1	*2circles*	2	600	2
Synthetic	2	*compound*	2	399	6
Synthetic	3	*twocirclesnoise2*	2	610	3
Synthetic	4	*spiral*	2	312	3
Synthetic	5	*pathbase*	2	300	3
Synthetic	6	*jain*	2	373	2
Synthetic	7	*flame*	2	240	2
Synthetic	8	*D1*	2	87	3
Synthetic	9	*D2*	2	85	4
Synthetic	10	*DS5*	2	500	5
Synthetic	11	*skewed*	2	1,000	6
Synthetic	12	*unbalance*	2	6,500	8
Real world	13	*thyroid*	6	215	3
Real world	14	*breast*	9	277	2
Real world	15	*glass*	9	214	6
Real world	16	*liver*	6	345	2
Real world	17	*heart*	13	303	2
Real world	18	*seeds*	7	210	3
Real world	19	*zoo*	16	101	7
Real world	20	*wine*	13	178	3
Real world	21	*vote*	16	435	2
Real world	22	*iris*	4	150	3
Real world	23	*dna*	180	2,000	3
Real world	24	*msplice*	240	3,175	3

Data set	Method	Parameter value	FM	JI	RI	NMI
2circles	AP	31/0.9	0.3559	0.2027	0.4992	–
	K-means	2	0.4983	0.3318	0.4992	0
	DBSCAN	3/3	1	1	1	1
	DP	3.1702	0.5028	0.3358	0.5026	0.0050
	DPC-LG	5.6133	0.4991	0.3325	0.4998	0.0010
	TMsDP	0.0859/side = 0.0017	1	1	1	1
spiral	AP	30/0.9	0.3279	0.1961	0.5538	0.000369
	K-means	3	0.3274	0.1957	0.5541	0.000351
	DBSCAN	2.5/2	1	1	1	1
	DP	2.5812	1	1	1	1
	DPC-LG	2.5812	1	1	1	1
	TMsDP	1.7443/side = 0.0349	1	1	1	1
twocirclesnoise2	AP	30/0.9	0.3649	0.2029	0.5033	–
	K-means	3	0.4026	0.2461	0.5019	0.0004
	DBSCAN	1.8/5	0.9967	0.9934	0.9967	0.9850
	DP	1.3646	0.4833	0.3185	0.5041	0.0211
	DPC-LG	0.5302	0.4947	0.3286	0.5066	0.0254
	TMsDP	0.0872/side = 0.0017	0.9918	0.9837	0.9918	0.9901
jain	AP	43/0.7	0.5847	0.3900	0.5793	–
	K-means	2	0.6977	0.5315	0.6591	0.3672
	DBSCAN	2.9/20	1	1	1	1
	DP	11.812	1	1	1	1
	DPC-LG	6.027	1	1	1	1
	TMsDP	1.3537/side = 0.0271	1	1	1	1
pathbase	AP	71/0.7	0.6321	0.4483	0.6822	0.2804
	K-means	3	0.6617	0.4908	0.7476	0.5470
	DBSCAN	2/5	0.7518	0.5727	0.7594	0.6965
	DP	1.1011	0.6654	0.4950	0.7509	0.5530
	DPC-LG	1.8974	0.6473	0.4693	0.7134	0.5039
	TMsDP	2.5500/side = 0.4878	0.9739	0.9491	0.9826	0.9363
compound	AP	16/0.4	0.6838	0.5193	0.8471	0.7469
	K-means	6	0.6422	0.4650	0.8432	0.7202
	DBSCAN	1/5	0.9103	0.8335	0.9528	0.8708
	DP	0.8732	0.7223	0.5648	0.8670	0.8363
	DPC-LG	0.8732	0.6437	0.4731	0.8340	0.7665
	TMsDP	0.4472/side = 0.0447	0.8703	0.7584	0.9216	0.8653

Table III.
The performance benchmark of synthetic data sets

4.1 Experimental results of synthetic data sets

These 12 synthetic data sets can actually be divided into several different types, including manifold data sets, multiple center data sets, data sets with unbalanced and skewed size and data sets with varying sizes. The authors present the experimental results of the 12 synthetic data sets in Tables III and IV and the clustering results of these data sets in Figures 8–19. In the clustering result figures, the *original distribution* denotes the real distribution of a data set, panel (a) shows the clustering result of the AP algorithm, panel (b) shows the clustering result of the K-means algorithm, panel (c) shows the clustering result of the DBSCAN algorithm, panel (d) shows the clustering result of the DP algorithm, panel (e) shows the clustering result of the DPC-LG algorithm and panel (f) shows the clustering result of the TMsDP algorithm.

Data set	Method	Parameter value	FM	JI	RI	NMI
D1	AP	2.3/0.6	0.8766	0.7684	0.9201	–
	K-means	3	0.9745	0.9503	0.9824	0.9515
	DBSCAN	0.65/2.3	0.9193	0.8451	0.9465	–
	DP	0.6374	1	1	1	1
	DPC-LG	1.3231	1	1	1	1
	TMsDP	0.1934/side = 0.3868	1	1	1	1
D2	AP	2.3/0.6	0.9756	0.9524	0.9882	0.9655
	K-means	4	0.9756	0.9524	0.9882	0.9655
	DBSCAN	2.4/18	0.9524	0.9092	0.9770	0.9427
	DP	0.263	0.9756	0.9524	0.9882	0.9655
	DPC-LG	0.4094	0.9756	0.9524	0.9882	0.9655
	TMsDP	0.3542/side = 0.7083	0.9756	0.9524	0.9882	0.9655
DS5	AP	52/0.8	0.7834	0.6275	0.8896	0.7913
	K-means	5	0.8093	0.6792	0.9218	0.8206
	DBSCAN	0.03/6	0.8408	0.7086	0.9190	0.9018
	DP	0.041	0.7884	0.6413	0.9001	0.8663
	DPC-LG	0.0595	0.8205	0.6818	0.9114	0.8857
	TMsDP	0.0595/side = 0.1547	0.9099	0.8346	0.9637	0.9242
flame	AP	42/0.9	0.7473	0.5959	0.7381	0.4345
	K-means	2	0.7364	0.5822	0.7267	0.3989
	DBSCAN	1/6	0.9659	0.9336	0.9641	0.5312
	DP	1.1336	1	1	1	1
	DPC-LG	0.9301	1	1	1	1
	TMsDP	0.9301/side = 3.7202	0.9922	0.9845	0.9917	0.9635
skewed	AP	26.3/0.6	0.7082	0.5482	0.9024	0.7245
	K-means	6	0.7203	0.5629	0.9065	0.7422
	DBSCAN	49/6	0.9772	0.9550	0.9925	0.8755
	DP	71.5542	0.9901	0.9803	0.9967	0.9845
	DPC-LG	35.5106	0.9942	0.9884	0.9981	0.9906
	TMsDP	71.5542/side = 143.1084	0.9942	0.9884	0.9981	0.9906
unbalance	AP	33/0.8	0.9989	0.9978	0.9994	0.9943
	K-means	8	0.9989	0.9978	0.9994	0.9943
	DBSCAN	6000/6	0.9991	0.9983	0.9995	0.9603
	DP	1.1808e+3	0.9994	0.9988	0.9997	0.9956
	DPC-LG	3.8471e+3	0.9958	0.9916	0.9976	0.9844
	TMsDP	1.6846e+3/side = 3.3692e+3	0.9996	0.9992	0.9998	0.9964

Table IV. The performance benchmark of synthetic data sets

According to the visualization of the clustering results, this study could find that the proposed method, DBSCAN, DP and DPC-LG can obtain more accurate clustering results when analyzing some manifold data sets (such as *jain* and *spiral*). However, when analyzing some data sets with multiple centers (such as *2circles, compound* and *twocirclesnoise2*) and the data sets with unbalanced and skewed size (such as *unbalance* and *skewed*), only the proposed TMsDP can obtain more accurate clustering results among the six algorithms in the comparison experiments. Meanwhile, when analyzing some data sets with varying sizes (such as *D1* and *DS5*) and some data sets with irregular shapes (such as *flame* and *DS5*), the TMsDP still obtains more accurate clustering results than the other five comparison algorithms. In order to compare the clustering performance of these six methods more sharply, Tables III and IV present the evaluation index values of different algorithms with different parameter value settings, which demonstrate that the TMsDP outperforms other compared algorithms.

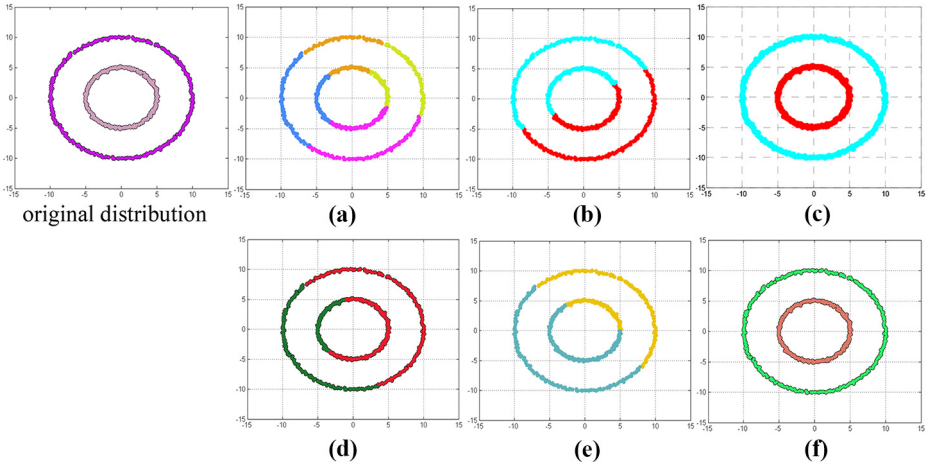


Figure 8.
The clustering result for data set *2circles*

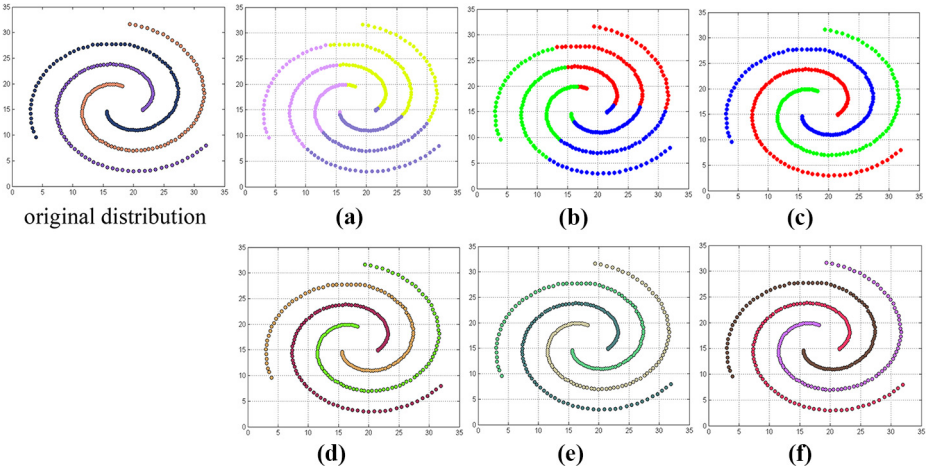


Figure 9.
The clustering result for data set *spiral*

4.2 Experimental results of real-world data sets

As shown in Tables V and VI, the TMsDP could obtain larger values in almost all the four evaluation metrics than the other five comparison algorithms when analyzing 12 real-world data sets. Of course, considering the diversity of data structural characteristics, the TMsDP could not obtain the best values in all evaluation metrics when analyzing all test data sets. Nevertheless, according to the available comparison results, the better clustering performance of TMsDP could still be shown.

4.3 Robustness analysis

In this experiment, the authors select the *seeds* and *liver* with different degrees of noise to evaluate the robustness of the compared algorithms. The authors generate different amounts of random data points as noise in the value space of the original data set. The

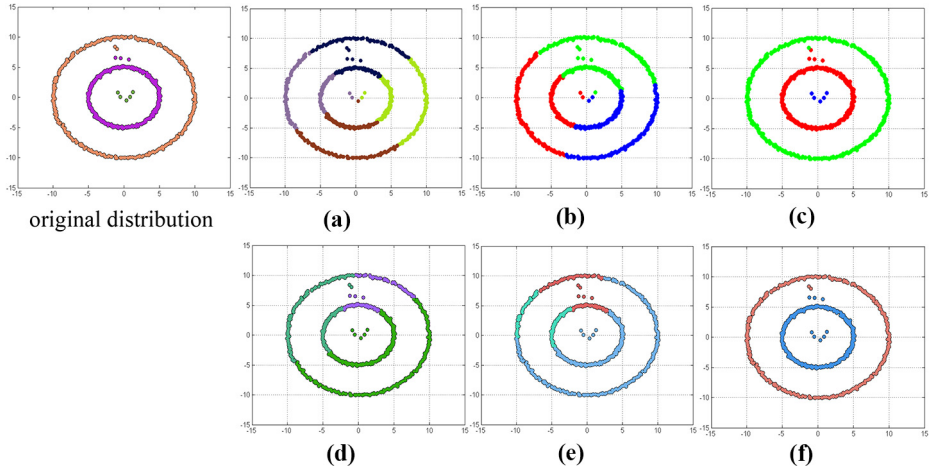


Figure 10.
The clustering result for data set *twocirclesnoise2*

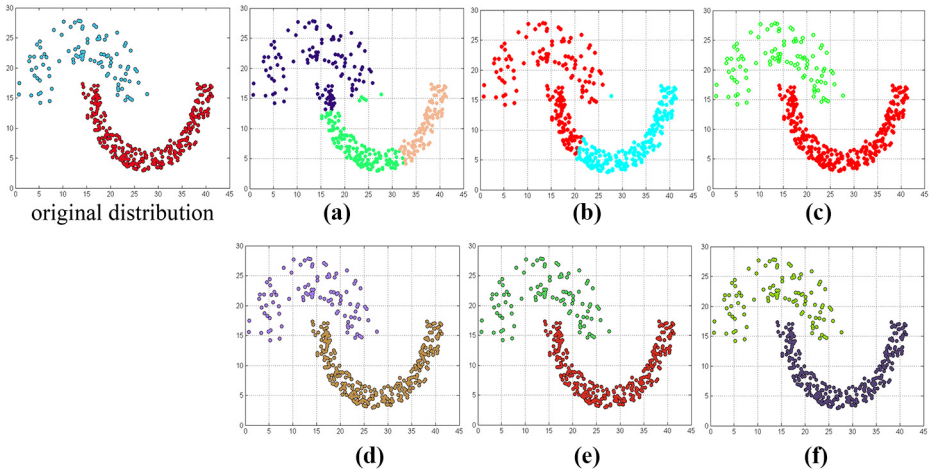


Figure 11.
The clustering result for data set *jain*

noise level of each data set gradually increases from 1.0 per cent to 10.0 per cent. The experimental results are presented in Figure 20.

As shown in Figure 20, with the increasing proportion of noise, the average FM value of each algorithm decreases. However, the average FM value of the TMsDP drops at a minimum rate, while that of AP drops at a maximum rate. Due to the small sample size of the data sets, the average FM values of TMsDP, DP, DPC-LG, DBSCAN and K-means are almost identical when the noise level rises from 1.0 per cent to 10.0 per cent. Therefore, the TMsDP retains higher accuracy in each case and illustrates higher robustness than the compared algorithms.

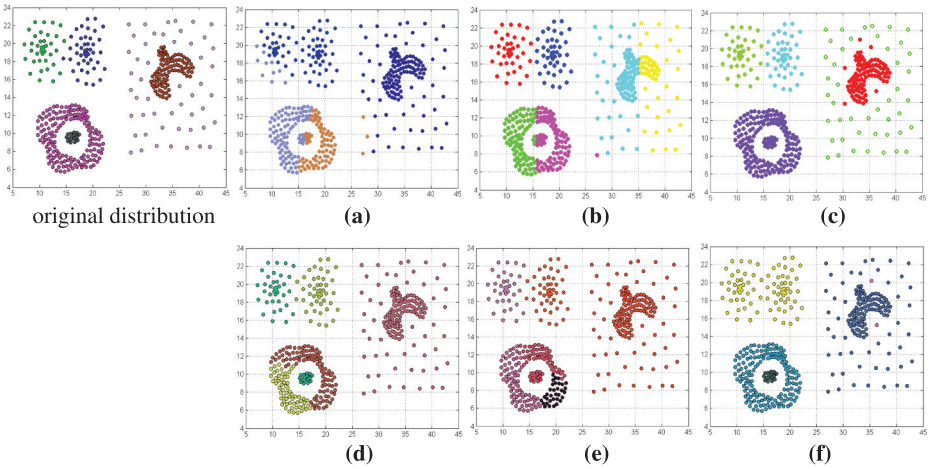


Figure 12.
The clustering result for data set *compound*

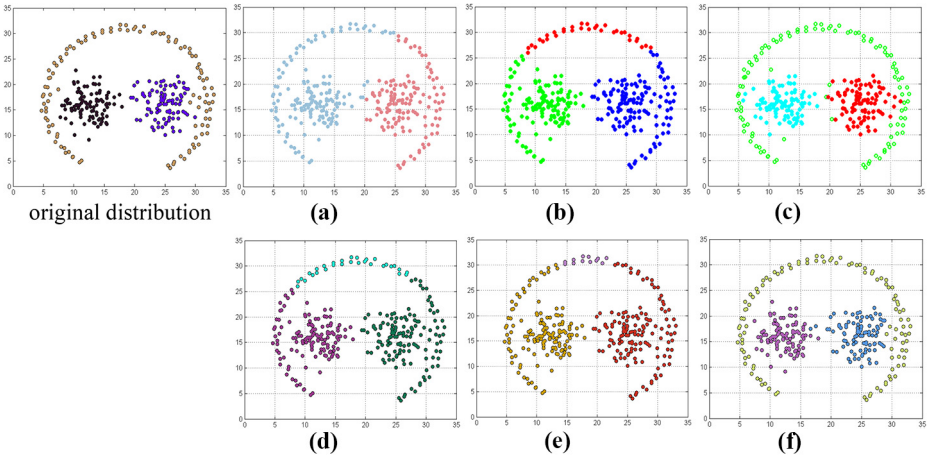


Figure 13.
The clustering result for data set *pathbase*

4.4 Running time analysis

In this section, the authors compare the running time of TMsDP with DPC-LG and DP on the 24 data sets, which include five different categories, i.e. (1) the synthetic 2D data sets with the data volume being less than 1,000, (2) the synthetic 2D data sets with the data volume being greater than or equal to 1,000, (3) the real-world data sets with the range of dimensions being from 2 to 10 and the data volume being less than 1,000, (4) the real-world data sets with the dimensions being greater than 10 and the data volume being less than 1,000 and (5) the real-world data sets with the dimensions being greater than 150 and the data volume being greater than or equal to 2,000 (selecting the average running time in 30 times of these three algorithms). The overall running speed is slow when dealing with higher dimensional data sets due to the limited running environment (Intel Core i5,

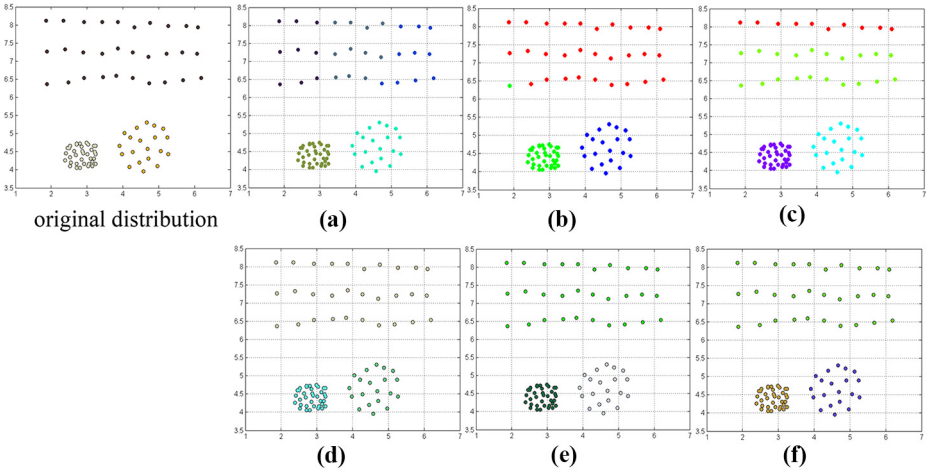


Figure 14.
The clustering result for data set *D1*

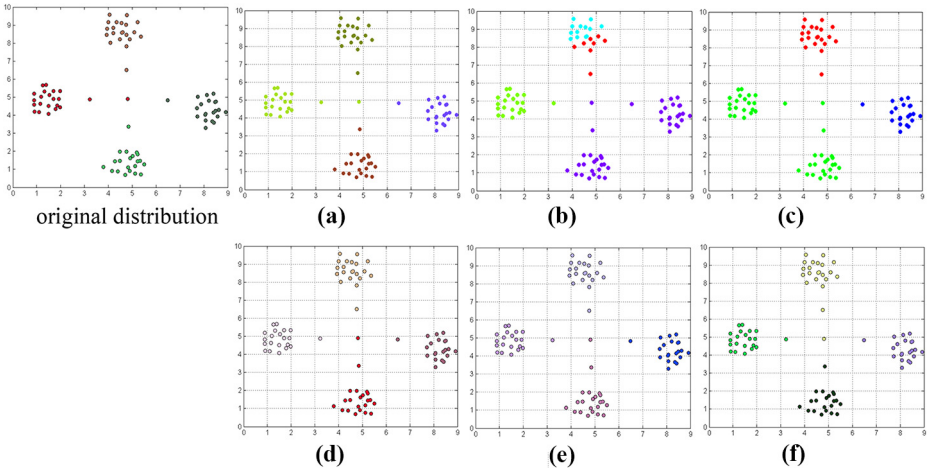


Figure 15.
The clustering result for data set *D2*

2.40 GHz, 8 GB RAM and MATLAB 2014a); therefore, when running some data sets with a large sample size and high dimensions, the overall running time of these three comparison algorithms is relatively long. In addition, because the TMsDP is an improved algorithm based on the DP, three DP-based algorithms (TMsDP, DPC-LG and DP) are selected for comparison. The running time result is shown in Table VII.

As shown in Table VII, the running time of the TMsDP is about twice as long as that of the DP. According to Section 3.3, the time complexity of the TMsDP is close to $O(n^2 + n^2 + n)$, which is close to DP (the time complexity of the DP is $O(n^2)$). Therefore, the actual running time of the TMsDP is not more than twice as long as that of the traditional DP.

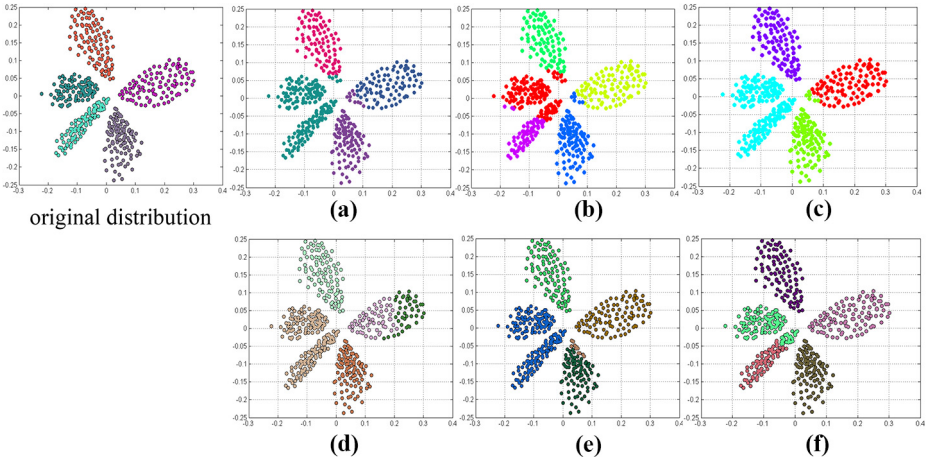


Figure 16.
The clustering result for data set *DS5*

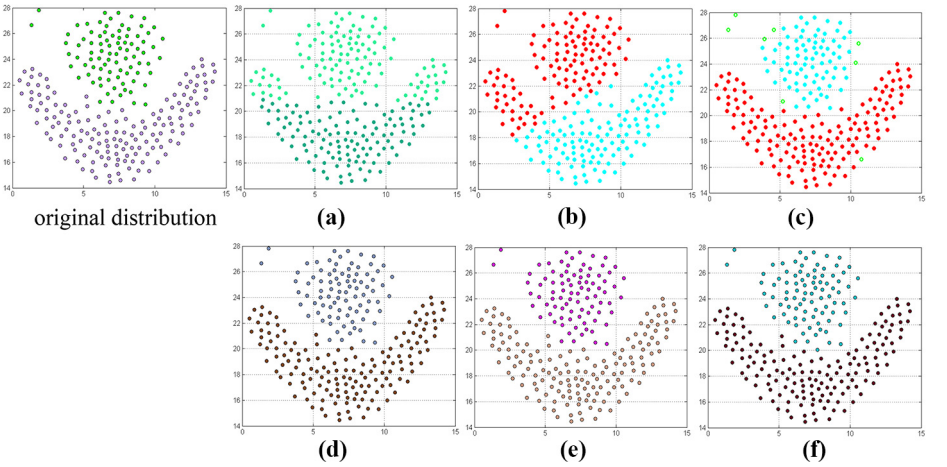


Figure 17.
The clustering result for data set *flame*

4.5 Overall performance review

In this paper, 12 synthetic data sets and 12 real-world data sets are utilized as experimental samples to demonstrate the clustering performance of the proposed method.

According to the clustering results of the 12 synthetic data sets, it could be seen that the proposed TMsDP shows better clustering performance than others when facing the manifold data sets, such as the *spiral* and *jain*. Moreover, when facing the multiple center data sets, such as the *2circles, compound* and *twocirclesnoise2*, and the data sets with an unbalanced and skewed size, such as the *unbalance* and *skewed*, the original DP is challenging to find potential centers in low-density regions, while the TMsDP method could adopt point-domains to explore more potential cluster centers

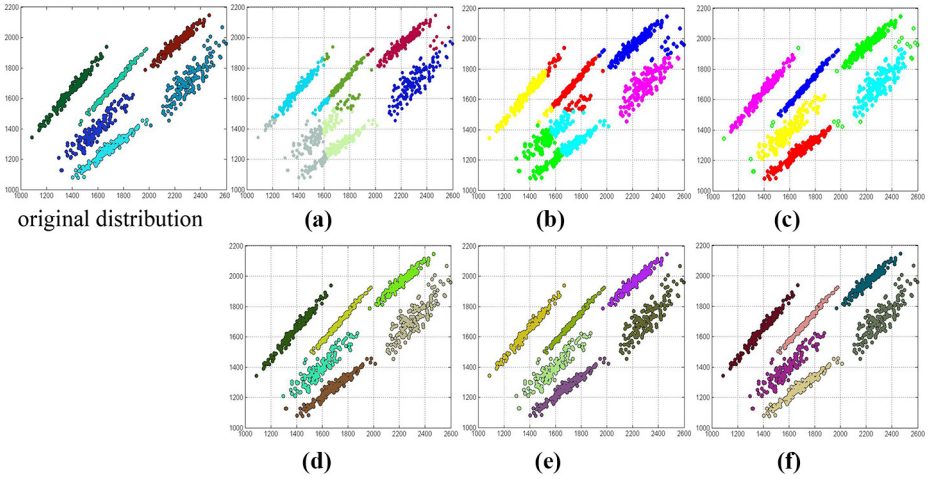


Figure 18.
The clustering result for data set *skewed*

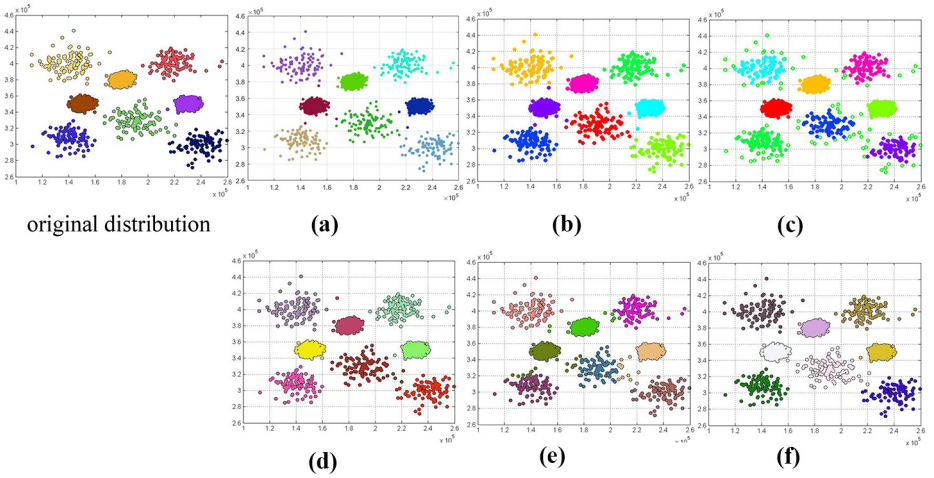


Figure 19.
The clustering result for data set *unbalance*

for achieving better clustering results. In addition, when facing the irregularly shaped data sets, such as *DS5, flame* and *spiral*, and the data sets with varying sizes, such as *D1* and *DS5*, the TMsDP could still obtain more accurate clustering results than other compared algorithms.

According to the clustering results of 12 real-world data sets, it is clearly shown that the TMsDP method could obtain better values in almost all evaluation metrics than other mentioned algorithms. In summary, the TMsDP improves the clustering performance compared with the original DP and expands the theoretical prospects of the density-based algorithms.

Data set	Method	Parameter value	FM	JI	RI	NMI
thyroid	AP	42/0.9	0.5336	0.3636	0.5199	0.0567
	K-means	3	0.8211	0.6960	0.8041	0.1497
	DBSCAN	4/6	0.7989	0.6651	0.7837	0.4446
	DP	9.2661	0.7380	0.5471	0.5653	0.0824
	DPC-LG	7.0661	0.7536	0.5721	0.6099	0.1360
	TMsDP	2.1307/side = 4.2615	0.7972	0.6428	0.7151	0.4673
breast	AP	14.3/0.9	0.5449	0.3732	0.5016	0.0007
	K-means	2	0.6510	0.4826	0.5939	0.0829
	DBSCAN	2.3/7.5	0.6141	0.4416	0.5759	0.0657
	DP	2.8002	0.7647	0.5886	0.5877	0.0371
	DPC-LG	1.8622	0.7647	0.5856	0.5877	0.0371
	TMsDP	0.4639/side = 0.1856	0.7651	0.5887	0.5971	0.0843
glass	AP	6.1/0.7	0.4200	0.2647	0.7211	0.3593
	K-means	6	0.5052	0.3298	0.6764	–
	DBSCAN	1.4/2	0.5638	0.3512	0.5927	0.2905
	DP	0.3132	0.5542	0.3333	0.5432	0.3922
	DPC-LG	0.3350	0.5428	0.3091	0.4591	0.3589
	TMsDP	0.3053/side = 0.6106	0.5520	0.3363	0.5619	0.4123
liver	AP	31/0.9	0.6102	0.4285	0.4998	0.0070
	K-means	2	0.6407	0.4538	0.5043	0.0009
	DBSCAN	10/3.6	0.5016	0.3347	0.4981	0.0037
	DP	28.1603	0.7124	0.5092	0.5104	0.0136
	DPC-LG	9.4868	0.7124	0.5092	0.5104	0.0136
	TMsDP	9.4868/side = 7.4403	0.7073	0.5064	0.5142	0.0196
heart	AP	31/0.9	0.5995	0.4280	0.5950	0.1611
	K-means	2	0.6162	0.4444	0.5921	0.1461
	DBSCAN	0.8/26.9	0.5514	0.3795	0.5187	0.0385
	DP	0.4834	0.6251	0.4510	0.5757	0.1375
	DPC-LG	0.4158	0.5897	0.4181	0.5893	0.1396
	TMsDP	0.3936/side = 0.3696	0.5964	0.4247	0.5837	0.1251
dna	AP	1.5/0.6	0.2077	0.0704	0.6195	0.0002
	K-means	3	0.6119	0.4384	0.7149	0.3646
	DBSCAN	4.2/5.1	0.6221	0.3903	0.4015	0.0478
	DP	0.3978	0.4974	0.3185	0.4731	0.0371
	DPC-LG	0.3979	0.4974	0.3274	0.5490	0.0632
	TMsDP	7.1414/side = 8.5697	0.4974	0.3185	0.4731	0.0371

Table V.
The performance benchmark of real-world data sets

5. Conclusion

To address the deficiencies of DP (i.e. failing to identify the cluster centers in low-density regions and being challenging to analyze a category with multi-centers), this paper proposes the TMsDP. The TMsDP shows three significant contributions: (1) constructing point-domain by introducing the pinhole imaging strategy to expand the search range for finding potential cluster centers; (2) proposing the novel methods to calculate point-domain density, domain distance and domain similarity and (3) finishing the clustering process based on domain similarity. The experimental results on 12 synthetic data sets and 12 real-world data sets illustrate that the TMsDP shows significantly improved clustering performance compared with original DP and the other algorithms experimentally compared in the paper.

Data set	Method	Parameter value	FM	JI	RI	NMI
seeds	AP	21/0.9	0.8068	0.6761	0.8714	0.7101
	K-means	3	0.8106	0.6815	0.8744	0.7061
	DBSCAN	1.1/12	0.5701	0.3950	0.6766	0.2685
	DP	0.6674	0.8026	0.6702	0.8673	0.6983
	DPC-LG	0.6674	0.7803	0.6396	0.8530	0.6833
	TMsDP	8.1971/side = 2.1964	0.8458	0.7328	0.8977	0.7436
zoo	AP	2.1/0.8	0.5893	0.4064	0.8354	0.6914
	K-means	7	0.6588	0.4826	0.8590	–
	DBSCAN	1/3.5	0.7716	0.6192	0.9032	0.8149
	DP	3.3166	0.5816	0.4053	0.7736	0.5841
	DPC-LG	2.8284	0.6121	0.4368	0.7935	0.6415
	TMsDP	3.3166/side = 33.1662	0.8813	0.7877	0.9453	0.8421
wine	AP	25/0.6	0.5828	0.4113	0.7161	0.4376
	K-means	3	0.5835	0.4120	0.7187	0.1505
	DBSCAN	100/0.8	0.5783	0.3368	0.3418	0.0296
	DP	101.3462	0.5892	0.3985	0.6102	0.3982
	DPC-LG	367.0219	0.6461	0.4496	0.6435	0.4624
	TMsDP	4.7849/side = 9.5699	0.6192	0.4247	0.6262	0.4158
vote	AP	22.9/0.7	0.7681	0.6232	0.7616	0.4380
	K-means	2	0.7742	0.6312	0.7684	0.4694
	DBSCAN	1/30	0.7086	0.5323	0.6011	0.1813
	DP	2.6458	0.7807	0.6400	0.7752	0.4900
	DPC-LG	2.4495	0.7237	0.5245	0.5259	0.0210
	TMsDP	2.6458/side = 0.6667	0.7483	0.5976	0.7420	0.4180
iris	AP	11.9/0.7	0.8208	0.6959	0.8797	0.7582
	K-means	3	0.8208	0.6959	0.8797	0.8688
	DBSCAN	1/30	0.7490	0.5779	0.7777	0.6952
	DP	0.8832	0.7635	0.5891	0.7766	0.7355
	DPC-LG	0.8832	0.7673	0.5920	0.7764	0.7452
	TMsDP	0.1732/side = 0.3750	0.8668	0.7649	0.9124	0.7900
msplice	AP	0.8/0.6	0.0274	0.0008	0.6154	–
	K-means	3	0.5470	0.3753	0.6715	0.3045
	DBSCAN	2.9/7.9	0.6192	0.3858	0.3931	0.0381
	DP	8.3666	0.4206	0.2649	0.5072	0.0068
	DPC-LG	8.3666	0.4355	0.2759	0.5050	0.0054
	TMsDP	8.3666/side = 10.0392	0.4729	0.3044	0.5107	0.0176

Table VI. The performance benchmark of real-world data sets

Although the proposed method shows better clustering performance, it adds six additional parameters, which could have more impacts on the clustering results. Therefore, the authors divide the future research plane into two aspects. In the theoretical aspect, the first part is to explore an improved calculation method of side values for reducing the number of parameters while preserving the clustering performance of the TMsDP; the second part is to update the calculation strategies of point-domain similarity and domain distance to accelerate the algorithm and the third part is to redesign a novel search mechanism and structure to automatically explore the potential cluster centers. In the application aspect, the authors extend the application fields of the TMsDP. When facing some data from the real-world problems, it could be found that the structures of these data are different from those of the experimental data sets mentioned above. Most of these data have diverse characteristics, including having multiple clustering centers, the clustering centers in the low-density region, unbalanced density distribution and unbalanced sample size distribution. For example, the text data, the consumption data of consumers, the stock

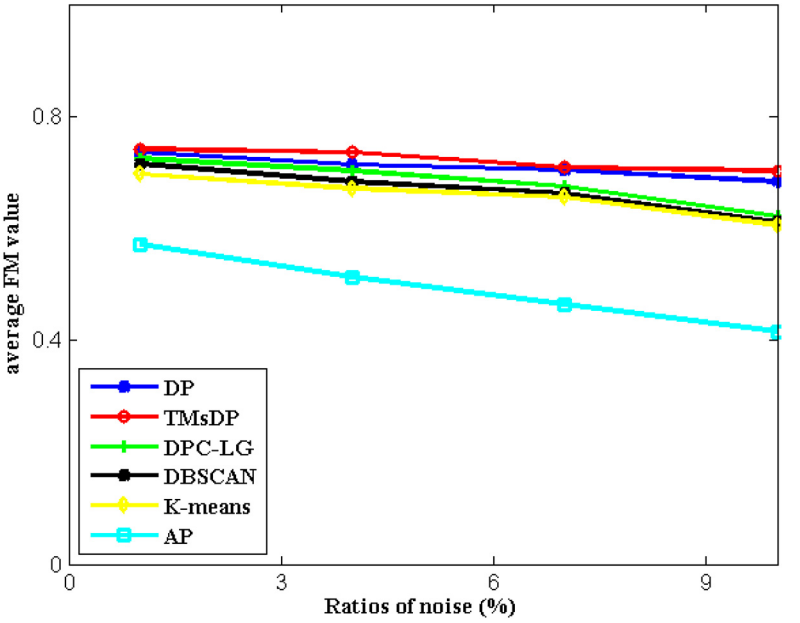


Figure 20.
Comparison of algorithm robustness

Table VII.
The running time result (unit: second)

Data sets characteristic	Data sets	TMsDP	DPC-LG	DP
Synthetic 2D data sets with data volume < 1,000	*2circles*	1.05514	0.72348	0.69241
	compound	0.74211	0.64561	0.63548
	twocirclesnoise2	0.94496	0.76864	0.67681
	spiral	0.92489	0.61591	0.59479
	pathbase	0.82898	0.58107	0.55721
	jain	1.16562	0.69894	0.65485
	flame	0.65291	0.55439	0.54409
	D1	0.58841	0.49088	0.45262
	D2	0.64021	0.48474	0.47937
	DS5	0.99563	0.70887	0.68828
Synthetic 2D data sets with data volume ≥ 1,000	*skewed*	1.80377	1.21797	1.01533
	unbalance	11.70188	9.75382	9.57304
Real-world data sets with 10 > dimension > 2 and data volume < 1,000	*iris*	0.73448	0.65967	0.60074
	thyroid	5.17884	3.82771	3.38868
	liver	31.46669	26.77568	25.87881
	seeds	6.54661	5.36481	5.01853
	breast	8.25847	5.21467	5.14373
	glass	2.01151	1.70734	1.31521
Real-world data sets with dimension >10 and data volume < 1,000	*heart*	7.07824	4.44836	4.25349
	wine	1.06705	1.00945	0.71224
	zoo	2.07121	1.42512	1.37737
	vote	47.03792	38.68827	38.03321
Real-world data sets with dimension >150 and data volume ≥ 2,000	*dna*	428.44746	336.21582	334.39406
	msplice	1205.55255	850.19014	837.18927

data, the financial data and the image data all have complex data features. Therefore, this study could apply the TMsDP to solve some related real-world problems, such as the topic identification of the online public opinion (mainly performing the text clustering), the customer segmentation for some enterprises (mainly performing the clustering analysis on the consumption data of consumers) and the problems of the facial image segmentation and detecting the CT scan images (mainly performing the image recognition). In addition, this study could also combine the TMsDP with some swarm intelligence optimization algorithms to solve the optimization problems in the real world.

ORCID iD
Jie Ma http://orcid.org/0000-0002-9005-0751

Note

1. Available at https://archive.ics.uci.edu/ml/index.php.

References

Abbas, M., El-Zoghabi, A. and Shoukry, A. (2021), "DenMune: density peak based clustering using mutual nearest neighbors", *Pattern Recognition*, Vol. 109, p. 107589.

Ansari, M.Y., Mainuddin, A.A. and Bhushan, G. (2021), "Spatiotemporal trajectory clustering: a clustering algorithm for spatiotemporal data", *Expert Systems and Applications*, Vol. 178, p. 115048.

Chen, J.G. and Yu, P.S. (2021), "A domain adaptive density clustering algorithm for data with varying density distribution", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, pp. 2310-2321.

Chen, Y.W., Hu, X.L., Fan, W.T., Shen, L.L., Zhang, Z., Liu, X., Du, J.X., Li, H.B., Chen, Y. and Li, H.L. (2020), "Fast density peak clustering for large scale data based on kNN", *Knowledge-Based Systems*, Vol. 187, p. 104824.

D'Errico, M., Facco, E., Laio, A. and Rodriguez, A. (2021), "Automatic topography of high-dimensional data sets by non-parametric density peak clustering", *Information Sciences*, Vol. 560, pp. 476-492.

Ding, J.J., He, X.X., Yuan, J.Q. and Jiang, B. (2018), "Automatic clustering based on density peak detection using generalized extreme value distribution", *Soft Computing*, Vol. 22 No. 9, pp. 2777-2796.

Du, M.J., Ding, S.F., Xue, Y. and Shi, Z.Z. (2019), "A novel density peaks clustering with sensitivity of local density and density-adaptive metric", *Knowledge and Information Systems*, Vol. 59, pp. 285-309.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996), *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In KDD, Vol. 96 No. 34, pp. 226-231.

Fan, X., Duan, Y.Z., Cheng, S.C., Zhang, Y.X. and Cheng, H. (2019), "Fast density-peaks clustering for registration-free pediatric white matter tract analysis", *Artificial Intelligence in Medicine*, Vol. 96, pp. 1-11.

Flores, K.G. and Garza, S.E. (2020), "Density peaks clustering with gap-based automatic center detection", *Knowledge-Based Systems*, Vol. 206, p. 106350.

Frey, B.J. and Dueck, D. (2007), "Clustering by passing messages between data points", *Science*, Vol. 315 No. 5814, pp. 972-976.

Guo, J.W., Zhang, J., Di, S., Zhang, Y.X., Xu, P.J., Li, L.T., Xie, Z.Q. and Li, Q.L. (2021), "An improved density-based approach to risk assessment on railway investment", *Data Technologies and Applications*, Vol. 56 No. 3, pp. 382-408. doi: 10.1108/DTA-11-2020-0291.

He, Y.L., Wu, Y.Y., Qin, H.L., Huang, J.Z.X. and Jin, Y. (2021), "Improved I-nice clustering algorithm based on density peaks mechanism", *Information Sciences*, Vol. 548, pp. 177-190.

Hou, J., Zhang, A.H. and Qi, N.M. (2020), "Density peak clustering based on relative density relationship", *Pattern Recognition*, Vol. 108, p. 107554.

Jain, A.K. (2010), "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, Vol. 31 No. 8, pp. 651-666.

Jangra, S. and Toshniwal, D. (2020), "VIDPSO: victim item deletion based PSO inspired sensitive pattern hiding algorithm for dense datasets", *Information Processing and Management*, Vol. 57 No. 5, p. 102255.

Jiang, J.H., Chen, Y.J., Hao, D.H. and Li, K.Q. (2019), "DPC-LG: density peaks clustering based on logistic distribution and gravitation", *Physica A*, Vol. 514, pp. 25-35.

Jo, T. (2020), "Semantic string operation for specializing AHC algorithm for text clustering", *Annals of Mathematics and Artificial Intelligence*, Vol. 88 No. 10, pp. 1083-1100.

Li, Y., Chu, X.Q., Tian, D., Feng, J.Y. and Mu, W.S. (2021), "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm", *Applied Soft Computing*, Vol. 113, p. 107924.

Liu, R., Wang, H. and Yu, X.M. (2018), "Shared-nearest-neighbor-based clustering by fast search and find of density peaks", *Information Sciences*, Vol. 450, pp. 200-226.

Long, W., Jiao, J.J., Liang, X.M., Wu, T.B., Xu, M. and Cai, S.H. (2021), "Pinhole-imaging-based learning butterfly optimization algorithm for global optimization and feature selection", *Applied Soft Computing*, Vol. 103, p. 107146.

Lu, G.L., Zhu, Y.B., Su, G.Z., Zhang, Z.M. and Yan, P. (2018), "Efficient block matching using improved particle swarm optimization with application to displacement measurement for nano motion systems", *Optics and Lasers in Engineering*, Vol. 111, pp. 246-254.

Lu, H., Shen, Z., Sang, X.S., Zhao, Q.H. and Lu, J.F. (2020), "Community detection method using improved density peak clustering and nonnegative matrix factorization", *Neurocomputing*, Vol. 415, pp. 247-257.

Luo, S., Liu, H. and Qi, E. (2021), "Recognition and labeling of faults in wind turbines with a density-based clustering algorithm", *Data Technologies and Applications*, Vol. 55 No. 5, pp. 841-868.

Medeghri, H. and Sabeur, S.A. (2021), "Anatomic compartments extraction from diffusion medical images using factorial analysis and K-means clustering methods: a combined analysis tool", *Multimedia Tools and Applications*, Vol. 80 No. 16, pp. 23949-23962.

Nie, B., Liu, D.Q., Liu, X.H. and Ye, W.J. (2021), "Phase I non-linear profiles monitoring using a modified Hausdorff distance algorithm and clustering analysis", *International Journal of Quality & Reliability Management*, Vol. 38 No. 2, pp. 536-550.

Parmar, M., Wang, D., Zhang, X.F., Tan, A.H., Miao, C.Y., Jiang, J.H. and Zhou, Y. (2019), "REDPC: a residual error-based density peak clustering algorithm", *Neurocomputing*, Vol. 348, pp. 82-96.

Rodriguez, A. and Laio, A. (2014), "Clustering by fast search and find of density peaks", *Science*, Vol. 344 No. 6191, pp. 1492-1496.

Ryu, J. and Kamata, S. (2021), "An efficient computational algorithm for Hausdorff distance based on points-ruling-out and systematic random sampling", *Pattern Recognition*, Vol. 114, p. 107857.

Seyedi, S.A., Lotfi, A., Moradi, P. and Qader, N.N. (2019), "Dynamic graph-based label propagation for density peaks clustering", *Expert Systems with Applications*, Vol. 115, pp. 314-328.

Singh, P. and Bose, S.S. (2021), "Ambiguous D-means fusion clustering algorithm based on ambiguous set theory: special application in clustering of CT scan images of COVID-19", *Knowledge-Based Systems*, Vol. 231, p. 107432.

Vavpetic, A. and Zagar, E. (2021), "On optimal polynomial geometric interpolation of circular arcs according to the Hausdorff distance", *Journal of Computational and Applied Mathematics*, Vol. 392, p. 113491.

Wang, S., Hua, W.Q., Liu, H.Y. and Jiao, L.C. (2019), "Unsupervised classification for polarimetric SAR images based on the improved CFSFDP algorithm", *International Journal of Remote Sensing*, Vol. 40 No. 8, pp. 3154-3178.

Wang, S.L., Li, Q., Zhao, C.F., Zhu, X.Q., Yuan, H.N. and Dai, T.R. (2021), "Extreme clustering-A clustering method via density extreme points", *Information Sciences*, Vol. 542, pp. 24-39.

Wang, Y.Z., Wang, D., Zhang, X.F., Peng, W., Miao, C.Y., Tan, A.E. and Zhou, Y. (2020), "McDPC: multi-center density peak clustering", *Neural Computing and Applications*, Vol. 32 No. 17, pp. 13465-13478.

Xu, X., Ding, S.F., Wang, L.J. and Wang, Y.R. (2020), "A robust density peaks clustering algorithm with density-sensitive similarity", *Knowledge-Based Systems*, Vol. 200, p. 106028.

Xu, X., Ding, S.F., Wang, Y.R., Wang, L.J. and Jia, W.K. (2021), "A fast density peaks clustering algorithm with sparse search", *Information Sciences*, Vol. 554, pp. 61-83.

Yan, M., Chen, Y.W., Hu, X.L., Cheng, D.D., Chen, Y. and Du, J.X. (2021), "Intrusion detection based on improved density peak clustering for imbalanced data on sensor-cloud systems", *Journal of Systems Architecture*, Vol. 118, p. 102212.

Yarinezhad, R. and Hashemi, S.N. (2019), "Solving the load balanced clustering and routing problems in WSNs with an fpt-approximation algorithm and a grid structure", *Pervasive and Mobile Computing*, Vol. 58, p. 101033.

Yu, H., Chen, L.Y. and Yao, J.T. (2021), "A three-way density peak clustering method based on evidence theory", *Knowledge-Based Systems*, Vol. 211, p. 106532.

Zhu, Y.L., Zhang, B., Dou, Z.H., Zou, H., Li, S.T., Sun, K. and Liao, Q.L. (2020), "Short-Term Load forecasting based on Gaussian process regression with density peak clustering and information sharing antlion optimizer", *IEEE Transactions on Electrical and Electronic Engineering*, Vol. 15 No. 9, pp. 1312-1320.

Further Reading

Wang, Y.Z., Wang, D., Peng, W., Miao, C.Y., Tan, A.E. and Zhou, Y. (2020), "A systematic density-based clustering method using anchor points", *Neurocomputing*, Vol. 400, pp. 352-370.

About the authors

Jie Ma, PhD, is Professor at the School of Business and Management, Jilin University in the People's Republic of China. She has published more than 100 academic articles in *Chinese Social Sciences Citation Index*, and her research interests include information resources management, information behavior, machine learning, deep learning and text analysis.

Zhiyuan Hao is a PhD candidate at the School of Business and Management, Jilin University in the People's Republic of China. He has published a number of papers in *Chinese Social Sciences Citation Index* and international SSCI/SCI journals, such as *Information Processing and Management, IEEE Access* and *Tehnicki Vjesnik-Technical Gazette*. His research interests include information behavior, machine learning, deep learning and text analysis. Jie Ma and Zhiyuan Hao contributed equally to this work. Zhiyuan Hao is the first corresponding author and can be contacted at: 15391910163@163.com

Mo Hu is working in the Department of Network and New Media, School of Journalism and Communication, Nanjing Normal University in the People's Republic of China. She received her PhD from the School of Business and Management, Jilin University. She majored in information resources management, machine learning and text analysis. Mo Hu is the second corresponding author and can be contacted at: 959539150@qq.com