Research Article

An Improved Clustering Algorithm Based on Density Peak and Nearest Neighbors

Chao Zhao,1 Junchuang Yang,2 and Kexin Wen3

1School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China
2Beijing Wodong Tianjun Information Technology Co. Ltd., Beijing 102676, China
3School of Management Engineering and Business, Hebei University of Engineering, Handan 056038, China

Correspondence should be addressed to Chao Zhao; zhaochao@hebeu.edu.cn

Received 9 February 2022; Revised 27 June 2022; Accepted 8 July 2022; Published 10 August 2022

1. Introduction

Clustering is a set of unsupervised algorithms, which plays an important role in machine learning and data mining. Traditional clustering algorithms can be divided into five categories [1–3]: partitioning, hierarchical, density-based, grid-based, and model-based. K-means is one of the most popular and simplest partition-based clustering algorithms, which has been widely used in many fields [4, 5]. However, K-means algorithm has its own drawbacks, and then many scholars have proposed improvements [6]. It is mainly concentrated in two ways [7–12]: selection of the initial clustering center and determination of the number of clusters K.

In 2014, Rodriguez and Laio proposed the density peak clustering (DPC) [13] algorithm. It suggests that the cluster centers are always surrounded by those data points with low local density, whereas it is far away from those data points with high local density. The DPC algorithm is of great significance to the selection of initial cluster centers and the determination of K-values, but it also has limitations [14–16]. A new strategy [17] attempts to automatically determine the cutoff distance by information entropy, which is based on data potential energy in the data field instead of the local density between data points. There is an approach [18] determining the initial cluster centers according to an improved DPC algorithm, meanwhile using entropy to calculate the weighted Euclidean distance between data points to optimize the K-means algorithm. A DPCSA algorithm [19] is presented which improves the DPC algorithm based on K-nearest neighbor (KNN) algorithm in addition to weighted local density sequence. The fuzzy weighted KNN is used to calculate the local density of data points [20], and it employs two assignment strategies to classify data points which improves the robustness of this clustering algorithm.

The DPC-DLP algorithm [21] uses the KNN idea to calculate the global cutoff distance and the local density of the data points, and uses the graph-based label propagation algorithm to assign data points to clusters. The PageRank algorithm can be used to calculate the local density of data points [22], which can avoid the instability of clustering result caused by the cutoff distance. The CFSFDP-HD
algorithm [23] uses heat diffusion to calculate the local density of data points to reduce the influence of cutoff distance on clustering result. The ADPC-KNN algorithm [24] is proposed based on the KNN to calculate the local density and automatically selects the initial cluster centers. The DPADN algorithm [25] uses a continuous function to redesign the local density, which can automatically determine the cluster centers. A multi-feature fusion with adaptive graph learning model [26] is developed, which is an unsupervised algorithm to address the issue of person reidentification. An adaptive approach is constructed [27] which is effective in simultaneously learning the affinity graph and feature fusion, resulting in better clustering results. The novel RCFE method [28] is proposed where the number of clusters is guaranteed to converge to the ground truth via a rank constraint on the Laplacian matrix. Finally, there are some other recent work involving density peak and graph and feature fusion, resulting in better clustering results.

Inspired by the ideas of DPC and KNN, this paper proposes an improved clustering algorithm. First of all, using the sum of the squared errors (SSE) to determine the optimal number of clusters—the K-value, then we select K initial cluster centers based on an improved DPC algorithm. In addition, K-means algorithm is applied to carry out iteration, and data points are divided into core points and outliers according to the average distance within each cluster. Finally, combined with the nearest neighbor idea of KNN, outliers are assigned by voting. Experimental results show that our proposed algorithm can achieve better clustering effect in most cases, compared with several known clustering algorithms.

The remainder of this paper is organized as follows. Section 2 reviews the DPC algorithm and introduces several improved methods. Section 3 provides our K-value selection method. Section 4 describes the proposed algorithm in detail. Experimental results are presented and discussed in Section 5. The conclusion is stated in Section 6.

2. Related Work

2.1. The DPC Algorithm. The DPC algorithm indicates that cluster centers are characterized by a higher density than their neighbors and a relatively large distance from data points with higher densities. In this approach, for each data point $x_i$, two attributes need to be calculated: local density $\rho_i$ and distance $\sigma_i$.

There are two ways for calculating the local density $\rho_i$ of $x_i$, namely, cutoff kernel and Gaussian kernel. The cutoff kernel is often used when data size is large, and it is defined as follows:

$$\rho_i = \sum_j \chi \left( \frac{x_i - x_j}{\sigma} \right) \cdot e^{-\left( \frac{d_{ij}}{c} \right)^2}.$$  \hspace{1cm} (2)

Comparing the two local density calculation methods, it is found that local density obtained by the cutoff kernel is a discrete value, while the other is a continuous value. The use of the cutoff kernel may cause conflicts when calculating the local density of different data points; therefore, the Gaussian kernel is more generally used to calculate the local density.

The distance $\sigma_i$ of data point $x_i$ is measured by computing the minimum distance between the point $x_i$ and any other point with higher density:

$$\sigma_i = \min_{j \neq i} \left( d_{ij} \right).$$  \hspace{1cm} (3)

For the point with highest density, it conventionally takes

$$\sigma_i = \max_j \left( d_{ij} \right).$$  \hspace{1cm} (4)

Assume that the original data set is shown in Figure 1. The local density $\rho_i$ and distance $\sigma_i$ are calculated for each data point $x_i$, and a decision graph is illustrated in Figure 2. According to the DPC algorithm, data points that have both large $\rho$ and $\sigma$ are most likely to be the cluster centers, which are located in the upper right in Figure 2.

In order to determine the cluster centers and its number clearly, the quantity $y$ in [13] which is a comprehensive consideration of $\rho$ and $\sigma$ is defined as

$$y_i = \rho_i \sigma_i.$$  \hspace{1cm} (5)

According to $y_i$ and number of data points, a new decision graph is shown in Figure 3. Obviously, the larger the value $y$, the more likely the corresponding data point to be the cluster center. It can be seen from Figure 3 that $y$ of noncluster center is relatively smooth, and there is an obvious jump of the value $y$ between the cluster center and the noncluster center. It is demonstrated that the data point above the dotted line is the cluster center. Consequently, the remaining data points are assigned to the cluster which has the nearest point with larger local density.

2.2. Set the Cutoff Distance. It is not advisable to manually set the cutoff distance $d_c$ based on experiences for diverse data structures. It affects the local density of data points, which in turn leads to different clustering results.

Assuming $p$ is a percentage, we define

$$d_c = \text{dis} \left( N \ast p \right).$$  \hspace{1cm} (6)

$N$ is the number of distances between any two data points, and $\text{dis}$ is an ascending order of these distances. Here, it takes the $N \ast p$ position of $\text{dis}$ sequence as the value of the cutoff distance $d_c$. Figure 4 shows clustering results by running the DPC algorithm in terms of different percentage $p$.

We can see that the cutoff distance $d_c$ obtained at different percentage $p$ has a huge impact on clustering results. Hence, the setting of the cutoff distance should be flexible, and an appropriate $d_c$ should be selected to certain data structure.
In information theory, Shannon entropy is used to measure the uncertainty of a system. The greater the entropy, the greater the uncertainty [37]. Similarly, the uncertainty of data distribution can be expressed by entropy. Suppose that the local density is 

\[ \rho_i = \sum_j e^{-\left(\frac{d_{ij}}{\sigma}\right)^2}, \]  

(7)

\( \sigma > 0 \) is an impact factor which is used to optimize the cutoff distance \( d_c \). For a data set \( X = \{x_1, x_2, \ldots, x_n\} \), the local density of each point is \( \{\rho_1, \rho_2, \ldots, \rho_n\} \). Considering the information entropy to evaluate the rationality of local density estimation [38],

\[ H = -\sum_{i=1}^{n} \frac{\rho_i}{Z} \log \left( \frac{\rho_i}{Z} \right), \]  

(8)

\[ Z = \sum_{i=1}^{n} \rho_i. \]  

(9)

\( Z \) is the normalized factor. The relationship between information entropy \( H \) and impact factor \( \sigma \) is shown in Figure 5. We can see that when \( \sigma \) starts to increase from \( 0 \), the information entropy \( H \) first decreases rapidly, then increases slowly, and finally tends to be stable. It follows that \( \sigma \) is the optimal value when the entropy is the lowest.

According to the 3B rule of Gaussian distribution in [39], the data point has a radius of influence on other points which is \( 3\sigma/\sqrt{2} \). Similarly, we think that data points can only affect points within its radius in clustering algorithm [40]. Therefore, we set the cutoff distance as

\[ d_c = \frac{3}{\sqrt{2}} \sigma. \]  

(10)

### 2.3. Optimize the Local Density

The DPC algorithm only considers the global structure of data set, and the clustering result is not good enough when it comes to unevenly distributed data sets [41]. When the data distribution is relatively concentrated and the density of each cluster is quite different, the change of local density will lead to the difficulty of selecting the correct cluster centers, thus affecting the final clustering results. Based on the idea of KNN algorithm, the nearest neighbors are introduced into local density calculation. The closer the data point is to the target point, the more contribution it makes to the local density of the target point. The new local density is defined as follows:

\[ \rho_i = \sum_{j=1}^{m} e^{-\left(\frac{d_{ij}}{\sigma}\right)^2} + \sum_{j=m+1}^{n} \frac{e^{-d_{ij}^*}}{(j-m)^2}, \]  

(11)

where \( m = 2 \times K \), \( d_{ij}^* \) represents the ascending order of the distances between data point \( i \) and other data points, \( n \) is the number of data points in data set, \( m \) is the number of nearest neighbors of data point \( i \), and \( K \) denotes the number of clusters.

### 3. The K-Value Selection Method

In the DPC algorithm, the boundary between cluster center point and noncluster center point may be unclear, which cause different people to choose different numbers of clusters according to their own experiences. Similarly, in the K-means algorithm, the number of clusters also needs to be preset by the user based on experiences. However, it is often difficult to determine the value of \( K \).

In order to solve the problem that it is difficult to choose the optimal number of clusters, we proposed a \( K \)-value selection method based on the SSE and the ET-SSE algorithm in [42].
Figure 4: Clustering results of the DPC algorithm under different percentage \( p \): (a) \( p = 1\% \), (b) \( p = 2\% \), (c) \( p = 4\% \), and (d) \( p = 6\% \).

Figure 5: The relationship between information entropy \( H \) and impact factor \( \sigma \).
Ordinarily, the $K$-means algorithm employs $SSE$ to measure clustering quality as

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (d(x, c_i))^2,$$  \hspace{1cm} (12)

where $d(x, c)$ represents the Euclidean distance from data points $x$ and $c$, and $K$ is the number of clusters. Let $C_i$ denote the set of all data points in the $i$-th cluster, and $c_i$ represents the cluster center of $i$-th cluster. Usually, the relationship between $SSE$ and $K$-value is demonstrated in Figure 6(a). It is obvious that with the increase of $K$, the $SSE$ will decrease and eventually stabilize. When the $K$-value gradually increases and approaches the actual number of clusters, $SSE$ will decrease rapidly; when the $K$-value is greater than the actual number of clusters, $SSE$ decreases slowly. Consequently, the $K$-value corresponding to the obvious “elbow point” in Figure 6(a) can be used as the optimal number of clusters. Nevertheless, sometimes there will be situation as shown in Figure 6(b), where there is no obvious “elbow point” in the figure. In this case, it will be more difficult to select the optimal $K$-value, which will affect the final clustering results.

To solve the problem that the “elbow point” in Figure 6(b) is not clear, the exponential function $e^x$ is introduced into the $SSE$ formula. Using the property that the $e^x$ function is sensitive to positive number, the $SSE$ values of different clusters are scaled down to further improve the difference degree of $SSE$ values, when $K$-value is not equal to the actual number of clusters. Meanwhile, in order to prevent the exponential explosion, an adjustment factor $\delta$ is added to update the weight of the $SSE$ value. A new calculation formula of the $SSE$ is defined as follows:

$$SSE_K = \sum_{i=1}^{K} e^{\max_{x \in C_i} \sum_{x \in C_i} (d(x, c_i))/\delta} + \frac{\max_{x \in C_i} \sum_{x \in C_i} (d(x, c_i))/\delta}{e^{\max_{x \in C_i} \sum_{x \in C_i} (d(x, c_i))/\delta}},$$  \hspace{1cm} (13)

where $\max$ means to find the largest $SSE$ value in $K$ clusters. In order to reduce the influence of manual parameters on the clustering result, based on a large number of experiments on different data sets, it is found that when

$$\delta = \sqrt{\max_{i \in 1:K} \sum_{x \in C_i} (d(x, c_i))^2},$$  \hspace{1cm} (14)

the “elbow point” in Figure 6 is the most obvious. Then, the predicted number of clusters is closest to the actual number of clusters, and the clustering effect is the best.

4. The Improved Clustering Algorithm

Aiming at the problem of unstable clustering results caused by randomly selecting initial cluster centers in traditional $K$-means algorithm, this paper proposed an improved clustering algorithm based on the density peak and nearest neighbors. Firstly, using information entropy to improve the DPC algorithm, we find the optimal cutoff distance of data set and then calculate the local density of the data points; additionally, the optimal number of clusters is obtained by the $K$-value selection method proposed in Section 3. Finally, according to the descending order of values in (5), the top $K$ corresponding data points are selected as the initial cluster centers. The $K$-means algorithm is used for iterative clustering.

4.1. Weighted Euclidean Distance. Let set $X = \{x_1, x_2, \ldots, x_n\}$ be a data set containing $n$ data points, and each data point contains $m$-dimensional attribute characteristics, where $x_{ip}$ ($i = 1, \ldots, n; p = 1, \ldots, m$) represents the $p$-th attribute characteristic of the $i$-th data point.

In order to remove the unit restrictions of different attributes in the original data and avoid its impact on the clustering results, the original data need to be normalized and converted into pure numerical data. After the normalization, each attribute is in the same order of magnitude, which is suitable for comprehensive comparative evaluation. The normalization formula is as follows:

$$x_{ip} = \frac{x_{ip} - \min(x_{ip})}{\max(x_{ip}) - \min(x_{ip})},$$  \hspace{1cm} (15)

where $\max(x_{ip})$ and $\min(x_{ip})$, respectively, represent the maximum and minimum of the $p$-th attribute value in a data point. The traditional $K$-means algorithm employs Euclidean distance to measure the similarity between data points. It is applicable to the uniform measurement of each attribute in the data point, and it treats the difference between different attributes equally. But in practice, the contribution of different attributes to the clustering results is quite different. To solve this problem, a weighted Euclidean distance is used to measure the similarity between data points. Let the weight of the $l$-th dimension attribute of the data set be

$$w_l = \frac{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2}{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2}, \quad l = 1, 2, \ldots, m,$$  \hspace{1cm} (16)

where $\bar{x}_p$ denotes the average value of the $p$-th dimension attribute. Then, the weighted Euclidean distance between the data points $x_1$ and $x_2$ is

$$d(x_1, x_2) = \sqrt{\sum_{l=1}^{m} w_l (x_{1l} - x_{2l})^2}.$$  \hspace{1cm} (17)

4.2. The Framework of the Proposed Algorithm. The $K$-means algorithm is sensitive to outliers [43], and the improved DPC algorithm can exclude the influence of outliers on the selection of initial cluster centers. However, $K$-means is an iterative clustering algorithm, and each iteration will generate new cluster centers. The outliers will have an impact on the generation of new cluster centers. Hence, it is necessary to distinguish the outliers in each cluster.

Let $C = (C_1, C_2, \ldots, C_k)$ be the $K$ clusters after first iteration, and the cluster centers are $c = (c_1', c_2', \ldots, c_k')$. The average distance of the $i$-th cluster is

$$\text{MeanDist}_i = \frac{\sum_{j=1}^{m} \sum_{x, c_j} d(x, c_j)}{\text{sum}(C_i)},$$  \hspace{1cm} (18)
where \( \text{sum}(C_i) \) is the number of data points in the \( i \)-th cluster. According to the average distance MeanDist of each cluster, the data points in the cluster are divided into core points and outliers. If the distance between the data point \( x_j \) and its cluster center is less than MeanDist, \( x_j \) is regarded as a core point; otherwise it is an outlier. The average value of the data points marked as core points in the cluster is calculated as the new cluster center of the next iteration, and the outliers do not participate in the calculation of the new cluster center.

To ensure that the assignment of outliers is closer to real situation and improve the clustering accuracy, the idea of KNN is introduced into the assignment of outliers [44–46]. Suppose that \( x_i \) is an outlier, calculate the distance sequence \( \text{dist} = [\text{dist}_1, \text{dist}_2, \ldots, \text{dist}_i] \) from \( x_i \) to each core point by (17). Then, \( \text{dist'} = [\text{dist}_1', \text{dist}_2', \ldots, \text{dist}_i'] \) is an ascending order, and we select the first \( 2^*K \) corresponding core points as the core neighbor point \( \text{core} = \{\text{core}_1, \text{core}_2, \ldots, \text{core}_{2^K}\} \). Since the core points have completed the assignment, the cluster category corresponding to each data point is counted from the sequence \( \text{Core} \), and the cluster with the largest number of statistics is regarded as the cluster to which the outlier \( x_i \) belongs. If there is more than one cluster category with the largest number of statistics, the cluster belonging to the nearest center point is marked as the cluster category of \( x_i \), according to the distance from \( x_i \) to the center point.

In summary, the steps of our proposed algorithm are as follows:

1. Input the original data set \( X = \{x_1, x_2, \ldots, x_n\} \).
2. Normalize the data set \( X \) according to (15) to obtain the processed data set \( X^* \).
3. According to (17), the weighted Euclidean distance between each data point in data set \( X^* \) is obtained.
4. Calculate \( \rho_i \) and \( \sigma_i \) of each data point in data set \( X^* \) according to (3), (4), and (11) and determine the optimal cutoff distance \( d_i \) according to the information entropy in Section 2.2.
5. Calculate \( y_i = \rho_i \sigma_i \) and arrange the \( y \) value sequence in descending order to obtain the sequence \( y' \).
6. Determine the number of clusters \( K \) according to the \( K \)-value selection method presented in Section 3.
7. The first \( K \) corresponding data points are selected from the sequence \( y' \) as the initial cluster centers.
8. The distances from each data point to each cluster center are calculated, and a data point is classified into the cluster with the nearest cluster center.
9. Calculate the mean distance MeanDist of each cluster according to (18). The data points in each cluster are divided into core points and outliers.
10. Calculate the average value of the core points in each cluster as the new cluster centers of the next iteration. The idea of nearest neighbors in KNN algorithm is used to re-assign the outliers.
11. When the cluster centers no longer change, the algorithm terminates and outputs the clustering result. Otherwise, go to Step 8.

Furthermore, a flowchart for the overall process is provided in Figure 7.

4.3. The Time Complexity Analysis. The time complexity of the improved DPC algorithm mainly depends on the local density \( \rho \) and distance \( \sigma \), which is \( \mathcal{O}(n^2) \). The time complexity of \( K \)-value selection method mainly comes from the sum of the squared error (SSE), which is \( \mathcal{O}(Kn^2) \). The time complexity of the idea based on nearest neighbors in KNN algorithm is used to re-assign the outliers.

5. Experiments

5.1. Experimental Environment and the Data Sets. The hardware environment is based on Windows 10 Professional 64-bit, Inter Core i3-4000M CPU, 2.40 GHz, and 4 GB memory. The proposed algorithm is implemented in MATLAB R2011a. The data of Wine, Pima, WDBC, Iris, and
Parkinsons in the UCI real data sets [47] are used as the experimental data sets, as shown in Table 1.

5.2. The Evaluation Measures. This paper uses accuracy (ACC), adjusted Rand index (ARI), and adjusted mutual information (AMI) to evaluate the performance of the clustering algorithms. Assume that \( P_j \) is a known manually labeled cluster, and \( C_j \) is a cluster generated by the clustering algorithm. The ACC calculation formula is as follows:

\[
\text{ACC}(P_j, C_j) = \frac{|P_j \cap C_j|}{|P_j|} \tag{19}
\]

where the value range of ACC is [0, 1] and the value range of ARI and AMI is \([-1, 1]\). The specific meanings and calculations of ARI and AMI are referred to [48, 49]. The three evaluation measures are positively correlated with clustering performance. The larger the value, the better the clustering performance of the algorithm.
5.3. Experimental Results and Discussion. In this paper, we compare our proposed algorithm with several existing algorithms: $K$-means, DPC [13], CNACS-$K$-means [50], and DCC-$K$-means [51]. Due to the influence of the initial cluster centers, the $K$-means algorithm takes the mean value of 20 times. The experimental results on the UCI data sets are shown in Tables 2–4. The data in bold are the best experimental results on this data set.

As shown in Tables 2–4, the algorithm we proposed is generally better than the other four comparison algorithms in terms of the three evaluation measures. It achieves the best experimental results especially on the Wine and Iris datasets.

In the comparison of ACC in Table 2, our algorithm is slightly worse than the $K$-means and DCC-$K$-means algorithms on the Pima and WDBC data sets, but is better than the other two comparison algorithms. However, the situation is just the opposite on the Parkinsons data set.

As for ARI comparison in Table 3, the proposed algorithm is lower than the CNACS-$K$-means on the Pima data set, whereas it is higher than the other three comparison algorithms. On the WDBC data set, it is only lower than $K$-means and DCC-$K$-means algorithms by 1.7%. Furthermore, it is lower than CNACS-$K$-means and 0.02% difference from DPC on the Parkinsons data set, but it is higher than the other two comparison algorithms.

In Table 4, the comparison of AMI shows that our algorithm is superior to the other four comparison algorithms on the Pima data set. On the WDBC data set, it is close to the $K$-means and DCC-$K$-means algorithms, and superior to the other two comparison algorithms. Although our algorithm outperforms the DPC and CNACS-$K$-means algorithms, it is indeed inferior to the other two comparison algorithms on the Parkinsons data set.

6. Conclusion

Focusing on the problems of randomly selecting initial cluster centers, manually determining the number of clusters, and not considering the influence of outliers on clustering process, this paper proposes an improved clustering...
algorithm based on the density peak and the nearest neighbors. Our algorithm uses an improved DPC algorithm to determine the initial cluster centers and calculates the sum of the squared errors within the clusters to find the optimal cluster number $K$. Moreover, the average distance within the clusters and the nearest neighbor idea are combined to determine the outliers and its assignment. The experimental results show that the proposed algorithm can achieve better clustering results on the UCI real data sets. However, it is not sure that the algorithm can be applied to large-scale data sets. In future, how to improve the stability and running efficiency of our algorithm for large-scale data set will be the focus.

**Data Availability**

The data sets used to support the findings of this study are included within the article.

**Conflicts of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.

**Acknowledgments**

This work was supported by the National Natural Science Foundation of China (61902262) and Handan City Science and Technology Research and Development Program (19422031008-15).

**References**


