ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.038

A Density Peak Clustering Method based on Improved Siphon Effect

Jiyuan Liu[1] and Xiaona Xia[2+]

[1] Affiliation (School of Co[1]mputer Science, Qufu Normal University, Rizhao Shandong 276826, China)

[2+] Affiliation (Faculty of Education, Qufu Normal University, Qufu, Shandong, 273165; Chinese Academy of Education Big Data, Qufu Normal University, Qufu, Shandong, 273165; School of Computer Science, Qufu Normal University, Rizhao, Shandong, 276826, China)

Abstract. To solve the problem that the density peak clustering algorithm needs to manually select the clustering center, a density peak clustering algorithm based on improved siphon effect (IDPC) is designed. IDPC completes the clustering of research objects with various distribution shapes by removing the maximum cluster center weight in the descending sequence, calculating and iterating the potential difference, and selecting the relative equilibrium point. A comprehensive comparative experiment is carried out on UCI data set. The experimental results show that IDPC can accurately and automatically select the cluster center and cluster number, improve the performance indexes such as Fowles mallows score, Rand index and adjusted mutual information, and can adaptively process data sets with low dimension and various distribution shapes.

Keywords: peak clustering, siphon effect, dichotomy, density clustering, unsupervised learning

1. Introduction

Clustering is an unsupervised in data mining field (according to the category of the unknown training sample solve various problems in pattern recognition [1-3]) clustering method [4].The reference [5] proposed a peak density clustering (Density Peaks Clustering, DPC) algorithm. This method could recognize arbitrary shapes of data, intuitively find the number of clusters, and easily find noise points. This method has good robustness. Reference [6] proposed four directions for optimizing DPC: The first is to increase speed; The second is the adaptive parameter is determined, but there is still a lack of automatic selection of cluster number selection. The third is to improve the clustering accuracy and robustness. The fourth is the processing of high dimensional data. To optimize the selection problem of DPC clustering center, this paper proposes a density peak clustering algorithm based on improved siphon effect. The algorithm of DPC is optimized to determine parameters adaptive PCA algorithm is used to research object dimension, using the dichotomy to adjust the size of the truncated distance, using the improved siphon effect to seek relative balance automatically divided into center of cluster and the cluster center, achieving the purpose of automatically determine the cluster centers and the number of cluster [7].

2. Introduction to DPC Algorithm

The DPC is based on the assumption that: for a data set, its cluster center is usually surrounded by points with lower local density than the cluster center, and the distance between these points with lower local density and other high density points is also larger. In this clustering model, the quantity to be calculated has two aspects, one is the local density of each data point ρ_i , and the other is the central offset distance of each data point δ_i . There are usually two ways to calculate local density, namely truncated kernel and Gaussian kernel. Truncated kernel is selected in this paper.

Truncated kernel is a way to calculate the density of discrete points, which calculated density is equal to the data points within the d_c neighborhood I the number of data points, could be easy to understand for calculating, centered on the i , d_c for the radius of circle all the number of data points, but does not include the data point itself, d_{ij} represent data point of the Euclidean distance between i and j , n is the number of data

+ Corresponding author.
E-mail address: xiaxn@sina.com.

points, x is the indicator function. It is not difficult to see that the truncated kernel results in discrete values. The Truncated kernel model is defined as follows:

$$\rho_i = \sum_j x(d_{ij} - d_c), \quad x(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \tag{1}$$

The central offset distance model is defined as follows:

$$\delta_i = \min_{j:\ \rho_j > \rho_i} \left(d_{ij}\right) \tag{2}$$

DPC helps us distinguish cluster centers, normal points and outliers through decision graphs based on truncation distance and local density. However, when the decision graph is very complex, manual selection will make mistakes. Therefore, the selection of truncation distance and cluster center needs quantitative analysis.

3. Peak Density Clustering Method based on Improved Siphon Equilibrium Law

3.1. Dichotomy

According to the selection range of truncation distance given in reference [5] (the average value of the number of points whose distance around each point is less than d_c accounts for 1% ~ 2% of the total points), the truncation distance is automatically determined by dichotomy. Set the initial truncation distance as half of the sum of the maximum and minimum Euclidean distances. If the ratio of the average number of points whose distance around each point is less than d_c to the total number of points is less than 1%, select 1%. If it is in the middle of 1% to 2%, select the actual ratio. If it is greater than 2%, it is 2%.

3.2. Improved Siphon Effect

Reference [8] provides quantitative analysis and theoretical basis for the division of cluster centers, and proposes an improved siphon balance method based on siphon phenomenon. The principle of siphon balance phenomenon is that when there is a pressure difference (potential energy difference) between the two ends of the tubular structure, water will flow from the high-pressure side to the low-pressure side until the whole device reaches equilibrium. The extreme value is removed to prevent too much influence on the average value. In the density peak clustering algorithm, for the data set whose weight of the first cluster center is much greater than that of other cluster centers after descending sorting, the selection of Q value is unstable. If the sum of the weights of the first Q cluster center is divided into two parts, the influence of the extreme value will be greater, so the extreme value will be removed.

3.2.1Cluster center weight

Reference [9] introduces the concept of cluster center weight from the perspective of quantification, and makes a quantitative analysis of the theory based on the improved siphon balance method. The weight of cluster center is the product of ρ_i and δ_i . Since all data class clusters are far smaller than the number of data points, the weight of cluster center is sorted in descending order to reduce the complexity of its calculation, and the firs γ_q points are selected as the candidate points of cluster center. Construct a candidate set of cluster centers $Q_q = \{\gamma_1, \gamma_2, \ldots\ldots \gamma_q\}$. Before introducing the Potential Difference, the point γ_c is introduced. The cluster center weights from γ_2 to γ_c be added to form the upper region, and the cluster center weights from γ_q to γ_c be added to form the lower region. The model is defined as:

$$\gamma_i = \rho_i * \delta_i \tag{3}$$

3.2.2 Potential Difference

Cluster center weights are defined as follows：

$$h(c) = \sum_{i=c}^{q} \gamma_i - \sum_{i=2}^{c-1} \gamma_i \tag{4}$$

Where $\sum_{i=2}^{c-1} \gamma_i$ represents the sum of the weights of the upper region excluding the weights of the first cluster center, and $\sum_{i=c}^{q} \gamma_i$ represents the sum of the weights of the cluster center of the lower region. Set the value of the initial difference segmentation point as $\gamma_c = \gamma_{q/2}$, and when there is a difference between the lower region and

the upper region, $h(c)$ will be less than zero, and the c value will gradually move upward until it reaches the relative equilibrium point. Next, the specific definition of relative equilibrium point is introduced.

3.2.3Relative equilibrium Point

The model of relative equilibrium point is defined as follows：

$$arg\ min\ h(c) = \{c \mid c \in [2, q/2], h(c) \geq 0\} \tag{5}$$

When γ_c subscript range in $[2, q/2]$, there are numerical c makes $h(c) \geq 0$, is considered before $q-1$ cluster center weights reach relative balance, stop action. According to the theory, when the sum of the weights of a few cluster centers reaches a relative balance with the sum of the weights of the majority of cluster centers, it is considered that a few high weight points could interpret the majority of low weight points, so as to select the cluster centers. The following are the specific steps of the density peak clustering algorithm based on the improved siphon effect.

Use the “Header 2” style, shown above, for subheads.

3.3. Algorithm Process

Input: data set

Output: clustering result Process

Step1: The dimensions of the data set is reduced by PCA, the Euclidean distance between data points d_{ij} is calculated, the truncation distance d_c is calculated, and the local density of each data point ρ_i , the center offset distance δ_i , and the cluster center weight I are calculated.

Step2: sort γ_i in descending order, and take the first q points as the candidate points of cluster center.

Step3: calculate the difference between the weights of the last $q-1$ cluster centers and the first $c-1$ cluster centers (excluding the first one), and calculate the bit difference to find the relative equilibrium point.

Step4: keep iterating forward the bit difference value, and gradually reduce c until the bit difference is greater than or equal to zero, when c is considered to be in relative equilibrium, the data point before c(including c) is taken as the center of the cluster.

Step5: cluster the data points according to the offset distance of the center, and allocate the remaining data points to the cluster center with the closest offset distance of the cluster center.

Step6: End of clustering of all data points.

4. Experimental Results and Process Analysis

In order to test and evaluate the effectiveness of IDPC, five UCI data sets, namely Wine, Abalone(AB), Iris, Connectionist Bench (CB) and Seeds, are used for relevant experiments. To determine the number of clusters, the time complexity, Fowlkes-Mallows scores (FM), RI (Rand Index), AMI (Adjusted Mutual Information) as IDPC evaluation the effectiveness of the algorithm. IDPC and decisions on cluster determine the number of figure and inflection point method are compared, in Fowlkes-Mallows scores, RI, AMI compared with three algorithms are compared.

The three indicators are as follows:

- FM: FM is defined as the geometric mean of accuracy and recall rate, and is often used to evaluate the quality of clustering and classification models. A higher value indicates that the predicted result is more similar to the real data. The ISBN assigned: 978-1-84626-xxx-x, etc.
- RI: The value could be $[0,1]$.The more consistent the clustering result is with the real situation.
- AMI: The larger the value is $[-1,1]$.The more consistent the clustering result is with the real situation.

In order to better test the relevant indicators of IDPC, this paper selects three similar and better comparison algorithms to carry out the experiment: the algorithm proposed in reference [5]、[10]、[11].

4.1. Experimental Data Set

The data sets in table 1 were all from UCI (http://archive.ics.uci.edu/ml/index.php).

Table 1: Data set description

Data	Instances	Features	Classes
Wine	178	13	3
Abalone	4177	8	3

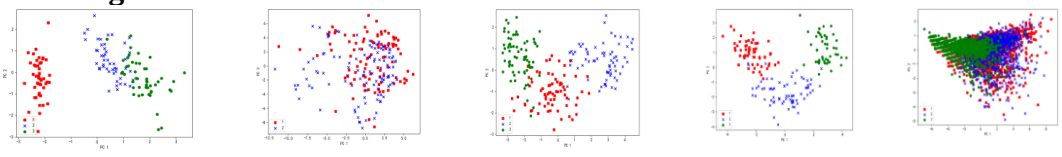
Iris	150	4	3
CB	208	60	2
Seeds	210	7	3

4.2. Experiment Parameter Setting

The selection criteria of q value is: within a certain range, the IDPC could obtain the correct number of clusters, and the intersection could be obtained in multiple data sets. In order to objectively select the q value of IDPC and the initial value of truncation distance, two small data sets wine and Iris and a large data set Abalone that is selected according to the number of data points in the data set for the experiment. Training set and test set account for 70% and 30%.

In terms of the selection of candidate number q value of cluster center: for data sets Abalone, Iris and Wine, the correct number of clustering could be obtained when $q \in [35,47]$, $q \in [27,52]$, $q \in [27,41]$ and Drink are met. Therefore, based on the selection requirements of the q values of the three data sets, the q value is set as 40. On the setting of initial truncation distance: When the initial values of dataset Abalone, Iris and Wine are set to half of the maximum and minimum Euclidean distances of data points x_i and x_j in the dataset respectively, the percentages of truncation distances are 1.8023%, 1.8882% and 1.3369% respectively, all between 1% and 2%. Therefore, the initial value of truncation distance is set as half of the maximum and minimum Euclidean distances of data points x_i and x_j in the dataset. The initial values of candidate cluster centers q and truncation distance are: $q = 40$, $d_c = \left(max\left(d_{ij}\right) + min\left(d_{ij}\right)\right)/2$.

4.3. Clustering Results



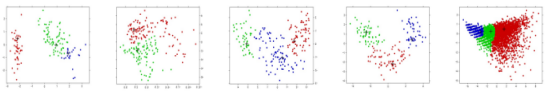
(a)Iris (b) CB (c)Sees (d) Wine (e)Abalone

Fig.1: Accurate depiction of five data sets after dimension reduction

The accurate description of the data set, the clustering effects of DPC, IDPC, spectral clustering (SP) and DBSCAN clustering algorithms are shown in Figures 1, 2, 3, 4 and 5. In iris, sees and wine datasets, IDPC and DPC clustering are reasonable. SP has obvious error in Iris data set, which is reasonably divided in sees and wine data sets. DBSCAN algorithm has more obvious errors in iris and sees classification. From the performance of the three data sets in the four algorithms, IDPC and DPC algorithms can reasonably divide the data sets of the research object set, while the spectral clustering algorithm will mistakenly divide the relatively independent research objects, divide a class of clusters into multi class clusters, and divide multiple class clusters into one class cluster. DBSCAN algorithm always clusters the research objects with low edge density. In CB and abalone data sets, the data points of different clusters in the two groups of data are staggered. It is difficult to distinguish which algorithm has better clustering effect by naked eyes, which will be judged by the evaluation index of clustering algorithm.

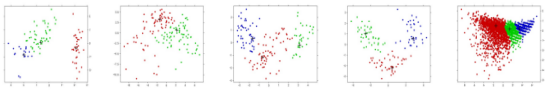
4.4. Comparison of Cluster Center Selection Methods

The comparison of different cluster center number selection methods is shown in Table 2. Taking wine, iris and abalone data sets as examples.The accuracy of selecting the number of cluster centers based on decision graph, inflection point method and improved siphon balance method is demonstrated respectively. Table 2 shows that the improved siphon balance method (ISE) is better than the decision graph and inflection point method in selecting the number of cluster centers. The reason for this problem is that decision graph (DG) and inflection point method (IPM) are highly subjective in selecting cluster centers. When the spatial structure of data is dense, human eyes can't distinguish small differences, and it is easy to make mistakes. ISE can accurately find the number of clusters in low dimensional data sets and effectively reduce the risk of artificial selection.



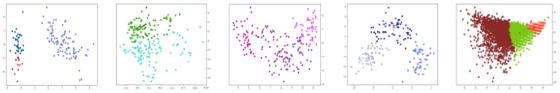
(a) Iris (b) CB (c)Sees (d) Wine (e)Abalone

Fig. 2: Clustering effect diagram of five data sets on DPC



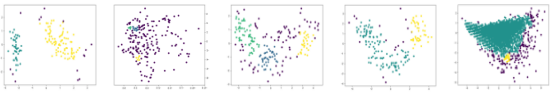
(a) Iris (b) CB (c)Sees (d)Wine (e)Abalone

Fig. 3: Clustering effect diagram of five data sets on IDPC



(a)Iris (b) CB (c) Sees (d) Wine (e)Abalone

Fig. 4: Clustering effect diagram of five data sets on SP



(a)Iris (b) CB (c) Sees (d) Wine (e)Abalone

Fig. 5: Clustering effect diagram of five data sets on DBSCAN

4.5. Time Complexity Analysis

Table 2: Comparison of different cluster center selection Methods

Data	DG	IPM	ISE	Number
Abalone	5	4	3	3
Iris	3	2	3	3
Wine	2	3	3	3

Table 3: Time complexity of four algorithms

algorithm	Time complexity
DPC	**o(n^2)**
DBSCAN	**o(n^2)**
Spectral clustering	**o(Dn^2)**
IDPC	**o(n^2)**

Analyze the time cost of IDPC algorithm and the other three comparison algorithms: as could be seen from Table 3, the time cost is well controlled after additional optimization steps are added, and the time complexity of the IDPC did not increase, which is the same as that of the DPC and DBSCAN. The time complexity of the SP is o(Dn^2), and D(dimension reduction) is usually much smaller than the number of data sets. The time complexity of the SP is the same as that of the IDPC, with no additional low cost. In conclusion, the time complexity of IDPC is equivalent to that of SP, DPC and DBSCAN.

Table 4: Fowlkes-Mallows scores

Data	DPC	DBSCAN	SP	IDPC
Wine	0.88701	0.90874	**0.91696**	0.90871
Seeds	0.73549	0.62199	**0.79493**	0.75166
AB	0.45237	0.42403	0.48629	**0.48683**
Iris	0.76398	0.74811	0.65800	**0.76725**
CB	0.47925	0.56376	0.52237	**0.59227**

Table 5: Rand index

Data	DPC	DBSCAN	SP	IDPC
Wine	0.92446	0.71427	0.72692	**0.93829**
Seeds	0.81986	0.78049	**0 .86441**	0.83098
AB	0.59147	0.58624	0.59608	**0.59622**
Iris	0.84890	0.75812	0.76575	**0.85221**
CB	0.49770	0.49828	**0.51398**	0.49777

Table 6: Adjusted mutual information

Data	DPC	DBSCAN	SP	IDPC
Wine	0.81638	0.51980	0.53552	**0.82037**
Seeds	0.62643	0.52375	0.63443	**0.64705**
AB	0.18520	0.18374	0.18193	**0.59623**
Iris	**0.75245**	0.58122	0.57099	0.73626
CB	-0.00333	0.01504	**0.02101**	-0.00301

Table 7: Intercept Distance and Average Time

Data	d_c /%	IDPCAT	DPCAT
AB	1.8023	104.7803s	108.4410s
Wine	1.8883	1.8200s	1.8610s
Iris	1.3369	1.5780s	1.5580s
CB	1.8870	2.1250s	2.2960s
Seeds	1.6867	2.1715	2.1360s

4.6. The Algorithm Classifies the Quality Evaluation Index

As could be seen from Table 4, 5 and 6: the clustering effect of these four algorithms could be ranked as follows: IDPC>Spectral clustering>DPC>DBSCAN. DBSCAN has the worst clustering effect, and the DPC without improvement could perform best in a group of data. It is obvious that DBSCAN is inefficient and of low quality. Although Spectral clustering uses the same PCA dimension reduction method as IDPC and

could easily be clustered by comparing the similarity matrix between data, it is still not as effective as IDPC. IDPC has both the accuracy of cluster center selection and higher clustering accuracy. Among the five groups of data, IDPC has the highest Fowlkes-Mallows score, indicating that IDPC clustering model is superior to the clustering model of other comparison algorithms and has the highest RAND coefficient, indicating that IDPC clustering model has the highest clustering accuracy. It has the most maximum adjustment mutual information, indicating that the clustering data set of IDPC clustering model has the highest similarity with the original data set, and its clustering result is closer to the real data set, which is the most advantageous among the four methods.

4.7. Average Running Time and d_c

The d_c and average running time (AT) of the algorithm are shown in Table 7. The final output results of d_c are all within the range of 1% and 2%, so as to solve the selection problem of d_c. In the case that the accurate clustering center could be obtained for q value, each q value is applied to the data set, run for several times and take the average value. The density peak clustering algorithm based on the siphon effect sacrifices the running time for clustering accuracy, so the AT is only compared with the density peak clustering algorithm.

IDPCAT and DPCAT are taken as the average values of 50 runs. The average running time of IDPC is 0.7634s less than that of DPC, which indicates that automatic selection of cluster center could reduce the time cost.

5. Conclusion

Aiming at the problem that DPC can not automatically select cluster center and cluster number, a density peak clustering algorithm (IDPC) based on improved siphon balance principle is proposed to reduce the selection risk caused by manual experience. Therefore, the designed IDPC algorithm and the other three comparison algorithms are applied to five groups of UCI data sets respectively, and the effectiveness and reliability of the algorithm are verified by sufficient simulation experiments. Firstly, the clustering effects of the four algorithms are compared. Secondly, the siphon balance method, decision method and inflection point method are improved respectively. The experimental results show that this method has obvious advantages. The peak density clustering algorithm based on the improved siphon effect provides a quantitative analysis basis for the selection of cluster centers and can accurately predict the number of cluster centers of the data set. In addition, The Fowlkes mallows score, RI and AMI are improved, and the model is optimized. The model has good evaluation ability.

6. References

[1] Xia X. Decision application mechanism of regression analysis of multi-category learning behaviors in interactive learning environment [J]. Interactive Learning Environments, 2021: 1-13.

[2] Xia X. Random field design and collaborative inference strategies for learning interaction activities [J]. Interactive Learning Environments, 2020: 1-25.A. Gray. Modern Differential Geometry. CRE Press, 1998.

[3] Xia X. Interaction recognition and intervention based on context feature fusion of learning behaviors in interactive learning environments [J]. Interactive Learning Environments, 2021: 1-18.

[4] Rosenberger C, Chehdi K. Unsupervised clustering method with optimal estimation of the number of clusters: Application to image segmentation[C]//Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. IEEE, 2000, 1: 656-659.

[5] Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

[6] Chen Yewang, Shen Lianlian, Zhong Caiming, Wang Tian, Chen Yi, Du Jixiang. A review of density peak clustering algorithms [J]. Journal of Computer Research and Development, 2020, 57(02):378-394.

[7] Yang Zhen, Wang Hongjun, Zhou Yu. An adaptive clustering algorithm based on truncation distance and clustering center [J]. Data analysis and knowledge discovery, 2018, 2 (03): 39-48.

[8] Ji D Y, Kim S H, Lee K Y, et al. Experimental study of small scale siphon breaker to verify Siphon Breaker

Simulation Program (SBSP) [J]. Annals of Nuclear Energy, 2018, 121：406-413.

[9] Ma Chunlai, Shan Hong, Ma tao. A density peak clustering algorithm based on automatic cluster center selection strategy [J]. Computer science, 2016, 43 (07): 255-258,280.

[10] Wang G, Yang J, Xu J. Granular computing: from granularity optimization to multi-granularity joint problem solving [J]. Granular Computing, 2017, 2(3): 105-120.

[11] Bai Lu, Zhao Xin, Kong Yu-ting, Zhang Zheng-hang, Shao Jin-xin, Qian Yu-rong. A review of Spectral Clustering Algorithms [J]. Computer Engineering and Applications, 2021, 57 (14):15-26.