PROCEEDINGS OF SCIENCE

A Fuzzy Density Peak Optimization Initial Centers Selection for K-medoids Clustering Algorithm

Cangsheng Liu[1a], **Qinglin Xu**[2b]
Guangdong University of Technology
Guangzhou, China
E-mail: [a]cang19910623@163.com;[b]422293273@qq.com

Xinran He
University of Rochester
Rochester, NY, United States
E-mail: hexinran0226@outlook.com

In order to solve the problem that the traditional K-medoids clustering algorithm needs to specify the number of clusters, which is sensitive to the initial cluster center and the slow convergence speed, the method of density peak optimization is used for solution. In this paper, we propose Fuzzy density peak K-medoids (FDP_K-mediods) algorithm. In the improved K-medoids algorithm, the local clustering center is obtained by calculating the local density and the high density distance, and then merged into the global clustering center, which can adaptively generate the initial clustering center and determine the number of clusters. The experimental results show that our scheme can adaptively generate the initial clustering center and determine the number of clusters with some practical and artificial data sets. Compared with the traditional K-medoids algorithm, the improved algorithm can accurately obtain the number of clusters and improve the algorithm's performance.

CENet2017
22-23 July 2017
Shanghai, China

[1]Speaker
[2]This study is supported by Guangdong Provincial Science and Technology Department of major projects (2016B030305002, 2016B030306003)

PoS(CENet2017)006

1.Introduction

K-medoids clustering algorithm [1-2] is based on the improvement of K-means clustering algorithm [1-2]. In the K-means implementation process, we firstly need to select initial center randomly. Only the first random selection of the initial centroid is the actual clustered point of concentration, and the subsequent non-centroid points assigned to the corresponding centroid point after re-calculation of the centroid point of concentration is not clustered. If some non-centroid is the outlier, which may result in the re-calculated centroid deviated from the cluster. K-medoids algorithm is a good solution to this problem, the algorithm does not use the average value of the objects in the cluster (center of mass) as the reference point, but the representative object is called the center point instead of center of mass . K-medoids algorithm has been applied to many fields such as the outlier analysis, the detection [3] and the distributed computing [4] etc.

The traditional K-medoids algorithm is a fast and efficient clustering algorithm. the algorithm is simple and effective, not sensitive to noise and outliers; but must be given the number of cluster K and the manual selection of the initial cluster center. It's easy to fall into the local optimal solution. In view of the shortcomings of the k-central point algorithm proposed above, many experts and scholars both at home and abroad have studied and put forward many solutions. The selection of initial cluster centers has been improved based on K-means++ [5-7], but this improvement can't significantly accelerate the convergence rate of the algorithm.

In this paper, in order to overcome the shortcomings of the traditional K-mediods algorithm, we propose a FDP_K-mediods (fuzzy density optimization K-medoids algorithm, FDP_K-mediods) with high density and distance local density structure decision diagram to determine the local density peak, then merge the local density peak to the global density peak. The global density peak is taken as the initial clustering center and the number of clusters is determined adaptively. In order to verify the clustering effect of the algorithm, the UCI data set and the artificial data set are selected for testing. The experimental results show that the proposed FDP_K-mediods algorithm can effectively select the initial cluster center, the number of clustering and the adaptive recognition data sets. It can effectively reduce the number of iterations, shorten the clustering time and improve the clustering accuracy.

2. Related Algorithm

2.1 K-medoids Clustering Algorithm

K-medoids algorithm, as associated with the K-means algorithm, can be regarded as a variant of K-means. K-medoids puts forward a new method of selecting particles by not simply using K-means algorithm to calculate the mean value. In the K-medoids algorithm, each iteration of the particle is selected from the sample points of the cluster as per the selection criteria that when the sample point becomes a new particle, it can improve the clustering quality of the cluster and make the cluster more compact.

The algorithm uses the absolute error criterion to define the compactness of a class of clusters for the objective function as follows:

$E=\sum_{i=1}^{k}\sum_{p\in c_i}|p-O_j|$, p is a sample point, and o_j is a cluster c_j of representatives

The K-medoids algorithm works as follows:

1. Initialize: randomly select k of the n data points as the medoids.
2. Assign each data point to the nearest medoids.
3. While the cost of objective function decreases less than the maximum number of iterations:

1) For each of the medoid o , for each non-medoid p , perform the following steps.

a) Exchange medoid o and non-medoid p , and recalculate the generated value of the division after the exchange.

b) If the exchange causes an increase in the cost, the exchange is canceled.

2.2 Fuzzy-CFSFDP Algorithm

CFSFDP algorithm is a new clustering algorithm proposed by the *Science Magazine* in 2014, which is a density-based clustering method. For CFSFDP Clustering Algorithm, two key parameters must be calculated for each data point: the local density and the distance from the point with higher local density, both depending on the distance between the data points.

The local density ρ_i of the data points i is defined as follows:

$$\rho_i=\sum_j \chi\left(d_{ij}-d_c\right) \tag{2.1}$$

Where $\chi(x)=\begin{matrix}1 & if\ x<0\\ 0 & ohterwise\end{matrix}$, d_c is a cut-off distance. For large amounts of data, the local density is essentially the relative density between the data points; so the choice of d_c is robust to the algorithm.

For the data point i , δ_i is the minimum value of a point to any point larger than its density, which is defined as follow :

$$\delta_i=\min_{j:\rho_j>\rho_i}\left(d_{ij}\right) \tag{2.2}$$

For the local density of the largest point, the general value of $\delta_i=\max_j\left(d_{ij}\right)$.

When two key parameters of local density ρ_i and the high density distance δ_i are acquired, ρ_i is taken as the horizontal axis and δ_i is taken as the horizontal axis to construct the decision diagram. The user selects the data with high horizontal axis and vertical axis on the decision diagram point as a cluster center, and then assign the other points to the cluster center. The decision chart is shown below.

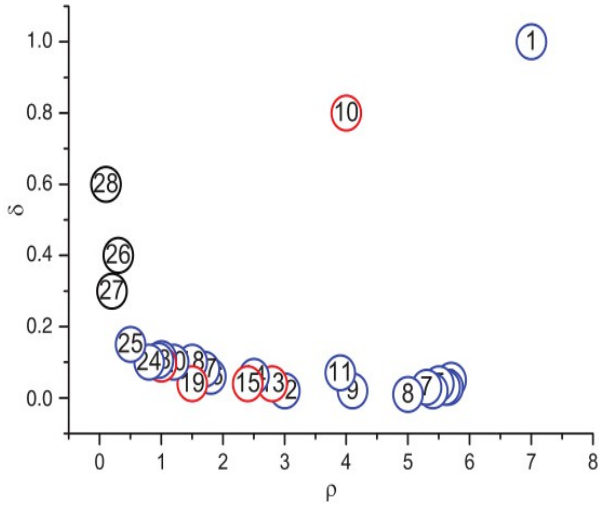

Figure 1: Decision Graph

Based on the CFSFDP algorithm, a Fuzzy-CFSFDP algorithm is proposed. After obtaining the two key parameters of local density ρ_i and high density distance δ_i, it is not to construct the decision graph but generate the cluster center adaptively without human intervention.

The Fuzzy-CFSFDP algorithm is based on the following two formulas:

$$EC_i = (\delta)_i \geq 2(\sigma_i) \tag{2.3}$$

Where EC_i refers to the expected cluster center, σ_i is the standard deviation of the high density distance δ_i. According to the CFSFDP algorithm, the distance between the cluster center and the rest of data-points in the cluster should be less than $2(\sigma_i)$; but there is still a low density with high value of δ_i. Therefore, these noises should be isolated via the following formula:

$$LC_i = EC_i \geq \mu(\rho_i) \tag{2.4}$$

LC_i are the local clustering center for the noise removal and $\mu(\rho_i)$ is the mean of ρ_i. By calculating the above two formulas out of the local cluster center is a large distance and neighbor data points compared with the higher density of data points. However, there may be different local clustering centers in the same cluster. In order to merge local clusters to obtain a global cluster, the fuzzy-CFSFDP finds the minimum distance between local clusters. If the cluster is located at a distance of d_c from other clusters having an average density, they are combined into a single cluster.

3. FDP_K-mediods Clustering Algorithm

To solve the above problems raised by k-mediods, FDP_K-mediods method is proposed. The FDP_K-medoids algorithm uses the local density ρ_i of the inverse of the sum of the nearest neighbors of the sample and measures the distance δ_i of the sample x_i using Equation (2). And Equation (4) and Equation (5) can get local clustering centers, then the local clustering centers are merged into the global clustering centers. The final clustering center is the center of the initial cluster and the number of density peak points is the number of clusters so that the initial cluster centers are in different clusters.

The FDP_K-medoids algorithm works as follows:

1. Select the initial medoids.
    - For sample x_i, calculate the distance d_{ij}, then calculate ρ_i and δ_i according to Equation (1) and Equation (2).
    - Use Equation (4) and Equation (5) to calculate local cluster centers.
    - Then local cluster centers are merged into global cluster centers. The global cluster centers are used as the medoids.
2. Assign each data point to the nearest medoids.

When the cost of objective function decreases:

- For each of the medoid o, for each non-medoid p:

a) Exchange medoid o and non-medoid p, and recalculate the generated value of the division after the exchange.

b) If the exchange causes an increase in the cost, the exchange is canceled.

4. Experiments

In order to verify the validity of the FDP_K-medoids algorithm proposed in this paper, we use R15 data set and D31 data set in one paper [4] and UCI data set: IRIS, GLASS, WINE and SONAR for verification. The details of the data set used are shown in Table 1.

Dataset	Objects	Dimensions	Classes
R15	1500	Numeric(2)	15
D31	3100	Numeric(2)	31
IRIS	150	Numeric(4)	3
GLASS	214	Numeric(10)	7
WINE	178	Numeric(13)	3
SONAR	208	Numeric(60)	2

Table 1: Data Set Description

4.1 Result and Discuss

The clustering analysis of the above data sets is carried out by using the traditional K-medoids algorithm and the improved algorithm proposed by this paper - FDP_K-mediods algorithm. The generated clustering results are shown in Fig. 2. In Fig. 2, the dots represent the data points, 'x' represents the initial cluster center and the same color represents the same cluster. As seen in Fig. 2 (a) and Fig. 2 (c), the initial clustering center deviates from the resulting clustering. As seen in Fig. 2 (b) and Fig.2 (d), FDP_K-mediod proposed algorithm has a good clustering effect, the initial clustering center and the final clustering results are close.

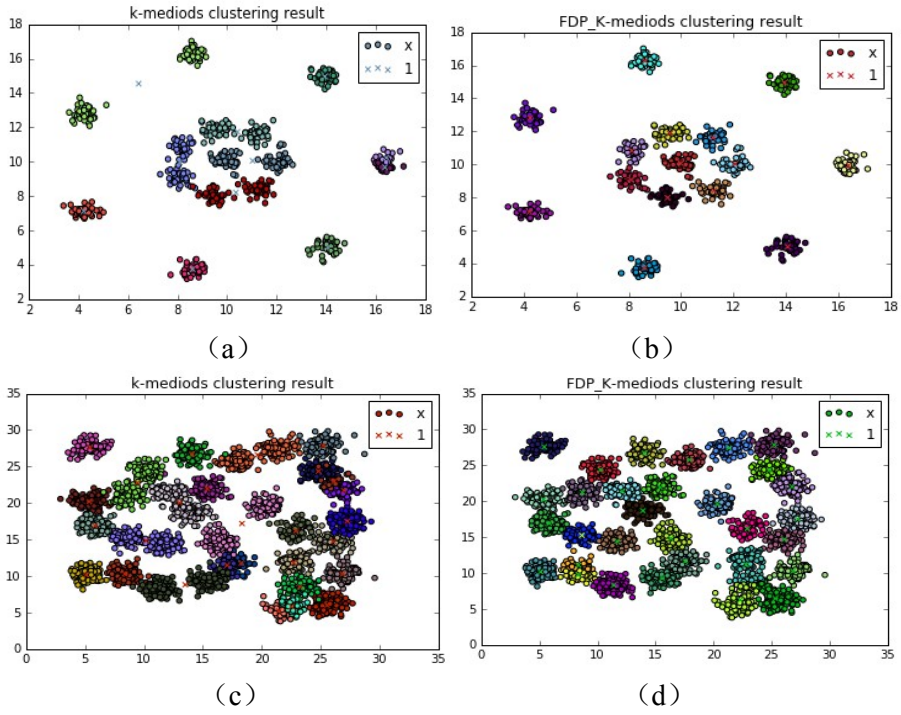

Figure 2: Clustering results

(a) R15 dataset clustering results for K-mediods. (b) R15 dataset clustering results for FDP_K-medoids. (c) D31 dataset clustering results for K-mediods. (d) D31 dataset clustering results for FDP_K-medoids.

5. Conclusion

In this paper, FDP_K-mediods algorithm is used to optimize the traditional k-mediods algorithm, which is inspired by the latest density clustering algorithm and fuzzy-CFSFDP. FDP_K-mediods algorithm uses local density and high density distance to construct the decision graph and chooses the peak as the local clustering, then uses the fuzzy rules to merge the local clusters to form the global clusters. FDP_K-mediods algorithm can solve the problem that the traditional k-mediods algorithm needs to manually specify the number of clusters and the sensitivity to the initial clustering center and it is easy to fall into the local optimization.

References

[1] MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). Berkeley: University of California Press

[2] Kaufmann L, Rousseeuw P J. *Clustering by Means of Medoids*[C]// Statistical Data Analysis Based on the L1-norm & Related Methods. North-Holland, 1987:405-416.

[3] Manikandan, R. P. S., A. M. Kalpana, and M. Naveenapriya. "*Outlier analysis and Detection using K-medoids with support vector machine.*" International Conference on Computer Communication and Informatics IEEE, 2016:1-7.

[4] Tao, Ye. "*Robust distributed k-mediods clustering algorithm*." Computer Engineering & Applications (2009).

[5] Arthur D, Vassilvitskii S. k-means++: *the advantages of careful seeding*[C]// Eighteenth Acm-Siam Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, Usa, January. 2007:1027-1035.

[6] Rodriguez A, Laio A. *Machine learning. Clustering by fast search and find of density peaks* [J]. Science, 2014, 344(6191):1492-6.

[7] R. Mehmood, R. Bie, H. Dawood, et al. *Fuzzy Clustering by Fast Search and Find of Density Peaks*[C]// International Conference on Identification, Information, and Knowledge in the Internet of Things. IEEE Computer Society, 2015:258-261.