ATLANTIS PRESS

Advances in Engineering Research, volume 141

5th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCE 2017)

Clustering by Fast Searching Density Peaks Based on Parameter Optimization

LV Zheng-hua[1,a], WANG Jun-hua[1,a], SHI Xia[1,a], ZHUANG Ya-de[1,a] and GE Shou-fu[1,a]

[1]Non-Commissioned Officer Academy of PLA Rocket Force, Qingzhou Shandong 262500, China

[a]skysword_wjh@126.com

Key words: clustering; density peak; cut-off distance parameter; local density

Abstract: An effective density clustering algorithm, called Clustering by Fast Search and Find of Density Peaks (CFSFDP), is appeared on *science* in 2014, which is simple and efficient and doesn't need many parameters. However, it needs make sure of cut-off distance parameter artificially. For the above problems, a new algorithm, called Clustering by Searching Density Peaks based on Parameter Optimization (CSDPPO), is proposed in this paper, which can estimate cut-off distance parameter adaptively. Firstly, local density information entropy function is constructed with cut-off distance parameter. And then cut-off distance parameter is estimated by solving minimization problem of local density information entropy. Because CSDPPO can obtain suitable cut-off distance parameter, its clustering performance is better than CFSFDP. Our experimental results validate the effectiveness of the proposed algorithm.

1. Introduction

Existing cluster methods can be divided into five types: Partitioning method, Hierarchical method, Density-based method, Grid-based method and Model-based method [1-5]. Although there are many kinds of cluster methods, they have their own shortcoming and can't satisfy the higher demand of clustering task at the same time.

This paper mainly explores clustering method based on density. The representative algorithm includes DBSCAN algorithm [6-7]、DENCLUE algorithm [8]、OPTICS algorithm [9] and AP (Affinity Propagation) algorithm [10-12] etc. DBSCAN algorithm is a kind of clustering algorithms based on high density connected region, which needs to artificially set two parameters: Eps and MinPts in advance. DENCLUE algorithm is a kind of clustering algorithms based on density distribution function. DENCLUE algorithm is more flexible and exact than DBSCAN algorithm in density but still needs to artificially set two parameters. OPTICS algorithm has a high computation burden of processing procedure and doesn't form clusters of data sets. AP algorithm is not designed to set cluster number in advance and require symmetry of similarity matrix among data, which acquire better effects in a practical application, while AP algorithm still exists two problems: cluster results are affected by bias parameters; it can't suit massive data clustering. In 2014, Alex Rodriguez and Alessandro Laio propose a new clustering algorithm in *science* named CFSFDP algorithm [13], which has a good computation efficiency and doesn't need iteration to discover cluster of arbitrary shapes. But the deficiency of CFSFDP algorithm is to make sure of parameter d_c. Bibliography [14] discusses the selection of parameter d_c and confirms parameter d_c by the percent of the average number of neighbors of every data points in overall data points. However, this algorithm chooses parameters artificially, and it can't ensure the effectiveness of parameter d_c.

For the deficiency of CFSFDP algorithm, in this paper we consider that parameter d_c can be determined by solving local density information entropy minimization problem. The proposed algorithm firstly forms local density information entropy function about parameter d_c, and then obtains the optimization solution of parameter d_c by solving minimum of the function. This algorithm can chooses parameter d_c adaptively and overcomes the deficiencies of CFSFDP algorithm, which needs artificial participation.

2. CFSFDP Algorithm

Firstly, we briefly introduce the CFSFDP algorithm. Order cluster data set $S = \{x_i\}_{i=1}^{N}$ and then CFSFDP algorithm mainly includes following four steps:

(1) Calculate local density

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c). \tag{1}$$

Where ρ_i is the local density of data point i. $d_c > 0$ is cut-off distance parameter. And d_{ij} is the normalized distance between two data points. Then function $\chi(x)$ can be defined as

$$\chi(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases}. \tag{2}$$

(2) Calculate the distance between data points

The distance of data point i refers to among all the data points whose local density is larger than ρ_i the distance between data point i and the data point that is nearest to data point x_i, which can be defined as

$$\delta_i = \begin{cases} \max\limits_{j}(d_{ij}) & \rho_i \textit{ is the} \max \\ \min\limits_{j:\rho_j > \rho_i}(d_{ij}) & \textit{else} \end{cases}. \tag{3}$$

(3) Ensure cluster center

Cluster center can be ensured, relying on local density ρ_i and data point distance δ_i. The data point, whose ρ_i and δ_i is larger at the same time, is treated as cluster center depending on decision diagram artificially.

(4) Assign data points to different class

Except cluster center data points is distributed to the same class of nearest data points. In the distributing process data points is in descending order.

The selection of parameter d_c affects the cluster results of algorithm. CFSFDP algorithm doesn't put forward appropriate method of select parameter d_c, which is designed to select parameter d_c by experience.

3. Clustering by Fast Searching Density Peaks Based on Parameter Optimization

3.1 cut-off distance parameter optimization

CFSFDP algorithm selects parameter d_c manually relying on experience and thus average number of neighbors of every data points is found in about $1\% - 2\%$ in overall data points. This paper presents that appropriate parameter d_c is determined by solving optimization solution of local density information entropy function and thus cut-off function (2) is replaced by Gaussian function, which is used to calculate local density [15] and can be written as

$$\rho_i = \sum_{j \neq i} e^{-\left(d_{ij}/d_c\right)^2} . \tag{4}$$

From equation (4), we can construct local density information entropy

$$H(d_c) = -\sum_{i=1}^{N} \frac{\rho_i}{Z} \log(\frac{\rho_i}{Z}) . \tag{5}$$

Where $Z = \sum_{i=1}^{N} \rho_i$ is local density sum of all data points.

When parameter d_c is selected the smaller the local density of all data points is, the larger the local density information entropy is. In other word, the larger the local density of all data points, the smaller the local density information entropy is. When local density information entropy is largest, all data points have the same local density. At present there is no way to ensure cluster center relying on local density. To make sure of cluster center by better using local density, the local density of data point is expected to reach to maximum differentiation, that is to say, local function information entropy is minimum. Therefore, parameter d_c can be determined by minimizing local density information entropy function $H(d_c)$, which is defined as

$$\hat{d}_c = \arg\min_{d_c} \left\{ H(d_c) = -\sum_{i=1}^{N} \frac{\rho_i}{Z} \log(\frac{\rho_i}{Z}) \right\} . \tag{6}$$

It's an unstrained optimization problem. The minimum of local density information entropy function $H(d_c)$ can be acquired using gradient-descent algorithm in order to determine cut-off distance parameter d_c.

3.2 calculate local density and distance

Local density is calculated by using Gaussian function thus corresponding local density can be written as

$$\rho_i = \sum_{j \neq i} e^{-\left(d_{ij}/d_c\right)^2} . \tag{7}$$

Let $I_S = \{1, 2, \cdots, N\}$, and thus data point distance can be calculated as

$$\delta_i = \begin{cases} \max\limits_{j}(d_{ij}) & I_S^i = \varnothing \\ \min\limits_{j:\rho_j > \rho_i}(d_{ij}) & I_S^i \neq \varnothing \end{cases} . \tag{8}$$

Where $I_S^i = \left\{ j \in I_S \middle| \rho_j > \rho_i \right\}$.

3.3 data point clustering

Similar to CFSFDP algorithm, cluster center is ensured according to the decision diagram of local density ρ_i and distance δ_i, and then the data points than doesn't belong to cluster center are classified. Suppose I_c contains K elements, which means the data has K cluster centers $\left\{ x_{I_c}^k \right\}_{k=1}^{K}$. Initialization cluster center classifier is $C\left(x_{I_c}^k\right) = k$. The data points not belonging to cluster center are in descending order according to local density and the data points are distributed to the same class whose distance is nearest to the data point, that is written as

$$C\left(x_i\right) = C\left(x_j : j = \min_{j \in I_S^i}(d_{ij}) \right) . \tag{9}$$

Where $C\left(x_i\right)$ is the class of data point x_i; and $I_S^i = \left\{ j \in I_S \middle| \rho_j > \rho_i \right\}$.

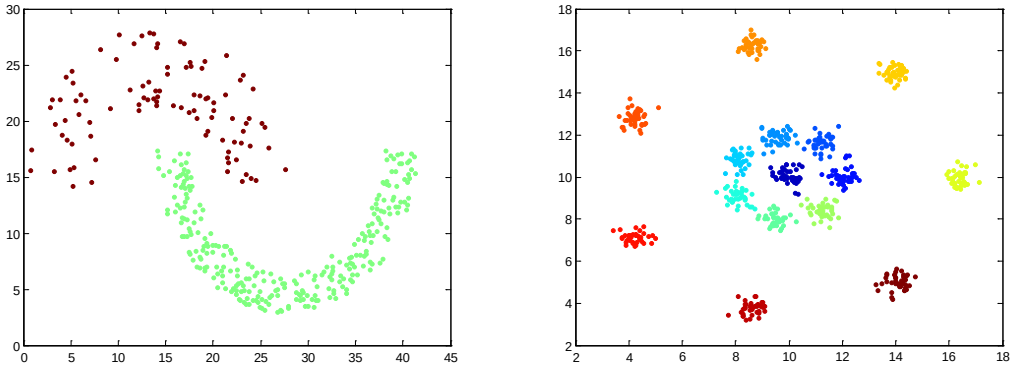
4. Experimental result and analysis

In this section, we will make experiments upon data set in [13] and UCI standard data set to testify the performance of the proposed algorithm, and compare it to CFSFDP algorithm. In experiment 1 the clustering effect of two algorithms is compared on data set in [13]; in experiment 2 the clustering performance of two algorithms on UCI standard data set is compared.

Data set in experiment: Jain data set and R15 data set in [13]; Iris and wine data set in UCI standard data set.

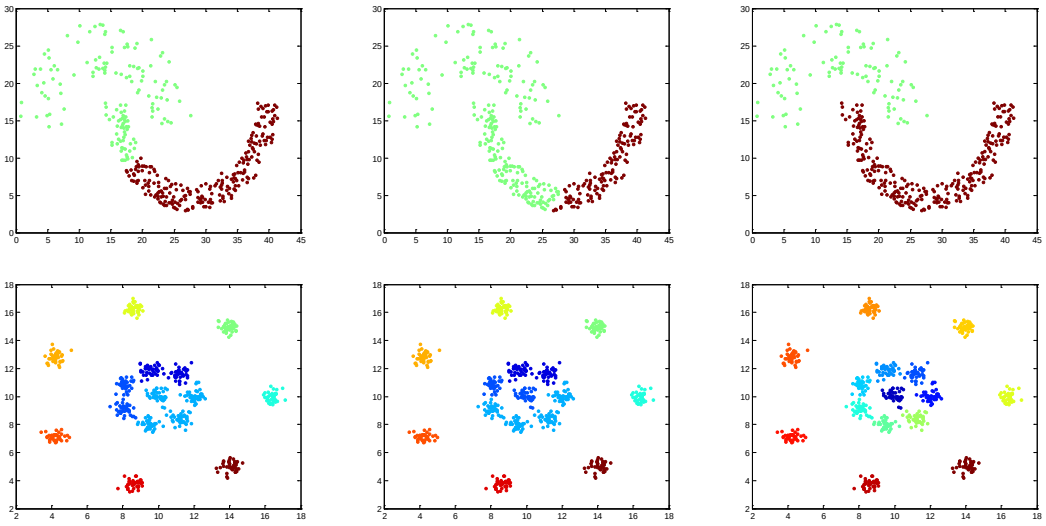
Parameter setting in experiment: CFSFDP algorithm respectively selects parameter d_c to make the average number of neighbors of every data point 1% and 2% of overall data set, which is denoted as $per = 1\%$ and $per = 2\%$.

Experiment 1: comparison on clustering effect of data set in [13]



(a) Jain data set (b) R15 data set

Fig.1. Data set in [13]



(a)CFSFDP with $per = 1\%$ (b)CFSFDP with $per = 2\%$ (c) The proposed algorithm

Fig.2. Clustering result comparison with two algorithms

Jain and R 15 data set is showed in fig.1, and clustering effect of two algorithms is compared in fig.2. It is shown from fig.2 that the proposed algorithm can obtain more suitable clustering result than CFSFDP algorithm on Jain and R15 data set.

Experiment 2: UCI standard data set experiment

In this section, the clustering accuracy of two algorithms is compared according to clustering experiment of UCI standard data set. Clustering accuracy of two algorithms is showed in table 1. It is shown form table 1 that the clustering accuracy of the proposed algorithm is higher than CFSFDP algorithm with $per = 1\%$ and $per = 2\%$.

Table 1 comparison on clustering accuracy of different algorithms

Data set	CFSFDP algorithm($per = 1\%$)	CFSFDP algorithm ($per = 2\%$)	The proposed algorithm
Iris	0.8312	0.8612	0.9235
Wine	0.7356	0.7156	0.9026

5. Conclusion

For the shortage of the CFSFDP algorithm that it needs to artificially ensure cut-off distance parameter d_c, we propose a new algorithm, called clustering by fast searching density peaks based on Parameter Optimization in this paper. The main innovation points of this algorithm are that it forms local density information entropy function $H(d_c)$ and make sure of parameter d_c by minimizing function $H(d_c)$. The experimental results show that cut-off distance parameter d_c that is determined according to the proposed algorithm in this paper can acquire better clustering effect than CFSFDP algorithm.

References

[1] L. Coletta, L. Vendramin, and E. Hruschka. Collaborative fuzzy clustering algorithms: some refinements and design guidelines [J]. IEEE Trans. Fuzzy Syst., 2012, 20(3):444–462

[2] Xie Juanying, Wang yane. K-means Algorithm of Minimum Variance Optimization Initial Cluster Centers [J]. Computer Engineering, 2015, 40(8): 206-223

[3] Wang Xin, Wang Hongguo. Cluster Analysis Method and Instrument Research [J]. Computer Science，2006, 33(2): 17-20

[4] Wang Chong, Lei Xiujuan. The New Niche Fireflies Divide Cluster Algorithm[J]. Computer Engineering, 2016, 40(5): 175-179

[5] Xiao Yu, Yu Jian．Semi-supervised Clustering Based on Affinity Propagation Algorithm Journal of Software, 2008, 19(11): 2803-2813

[6] M. Ester, H. Kriegel, and J. S ander． Adensity-based algorithm for discovering clustersin large spatial databases with noise[C]．Proceedings of 2nd international conference on knowledge discovery and data mining (KDD), 1996: 226-231

[7] Li Minghua, Liu Quan, Liu Zhong． New Development of Clustering in Data Mining[J]．Computer Application Research, 2008(01): 32-39

[8] Chen Yin, Cheng Yan. Data Mining: Concept、Model、Method and Algorithm[M]. Tsinghua University Press, 2003

[9] M. Ankerst, M. Breunig, and H. Krie. OPTICS: ordering points to identify the clustering structure[C]．Proceedings of ACM conference on Management of Data (SIGMOD 99) , 1999: 49-60

[10] B. Frey, D. Dueck. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972- 97

[11] Xie Wenbin, Tong Nan, Wang Zhongqiu. Affinity Propagation Algorithm Based on Particle Swarm[J].Computer System Applications，2016, 23(3): 103-108

[12] Liu Xiaonan, Yin Meijuan, Li Mingtao. Hierarchical Affinity Propagation Clustering Algorithm Facing Data Base [J]. Computer Science, 2015, 41(3): 185-192

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

[14] R.Seghers,R.Laoi.Clustering[Online].Available:http://rseghers.com/machine-learning/rodriguez-laoi-clustering, 2014

[15] R. Albarakati. Density based data clustering [D]. California State University, San Bernardino Scholar Works, 2016