Research on Density Peak Clustering Algorithm Based on Artificial Bee Colony Optimization

Chunyan Qiu[a], Yihe Zhang, Yang Liu[b,]*, Jialing Han, Shuang Hu, Zhiyu Chen

School of Management Science and Information Engineering, Jilin, University of Finance and Economics, Changchun, China
[a]81692002@qq.com, b7923759@qq.com
**Corresponding author*

Keywords: Density Peak Clustering; Artificial Bee Colony Optimization; K-Means; Cluster

Abstract: This paper proposes a density peak algorithm based on artificial bee colony optimization. The improved DPC algorithm can better realize the automatic identification and reasonable clustering of data points between clusters, and reduce the difficulty of selecting the value of the original density peak clustering (DPC) algorithm and the limitation of the neighboring principle aggregation operation in the low density area. This paper verifies the clustering effectiveness of the proposed algorithm using some known data sets and custom data sets. Clustering comparison experiments with K-Means, AP, and DPC algorithms on multiple classical datasets. The experimental results show that compared with the DPC algorithm, the proposed algorithm automatically recognizes and reasonably clusters data points between clusters; automatically identifies the cluster center points and clusters and automatically handles the advantages of randomly distributed data sets.

1. Introduction

In the era of Internet big data, data has exploded. Various industries have gradually adopted computer technology to manage data, greatly improving the ability to generate, collect, store and process data. Cluster analysis can analyze data without any prior information, and identify the potential structure of data space [1], which is widely used in pattern recognition, data analysis, image processing, market research and many other fields. As a result, various clustering algorithms have emerged. In 1975 and 1979, Hartigan and Wong [2] proposed a simple, efficient and widely used K-means clustering algorithm. In 1998, Alsabti et al. proposed a K-means algorithm based on the cost function, which successfully avoided determining the number of pre-allocated clusters and partially reduced the limitations of K-means [3]. In 2007, Frey and Dueck proposed the affinity propagation (AP) algorithm, and successfully applied this algorithm to image recognition [4].

In 2014, Rodriguez and Laio proposed a clustering algorithm based on density peak (DPC algorithm). This method can quickly find the density peak points of any shape data set, and can

Published by CSP © 2018 the Authors
DOI: 10.23977/iceccs.2018.011

efficiently perform sample allocation and reject outliers. A large number of experiments have confirmed the excellent performance of the DPC algorithm [5], but the algorithm needs to be further verified for the identification and classification of data points between clusters.

2. Density peak clustering based on artificial bee colony optimization

The DPC algorithm can automatically find the cluster center point of the data set sample, and the selected cluster center point has a higher density and is relatively far away from other cluster centers. Therefore, there are certain limitations.

The density peak clustering algorithm based on artificial bee colony optimization proposed in this paper makes corresponding adjustments to the aggregation principle on the basis of fully inheriting the advantages of DPC algorithm. It mainly improves the sensitivity of DPC algorithm to data points between clusters, and proposes a more scientific and rational aggregation principle of cluster points. The algorithm proposed in this paper is mainly divided into the following six steps:

(1) Step 1: Calculate the density of the data points and generate a decision map;
(2) Step 2: Perform initial clustering;
(3) Step 3: Identify data points between clusters;
(4) Step 4: The cluster label of the data point between the primary clusters;
(5) Step 5: Determine the cluster label of the data point between the clusters;
(6) Step 6: Complete the clustering.

2.1 Calculate the density of data points and generate decision maps

Calculate the calculation method in the DPC algorithm on the data point density.

$$\rho_i = \sum_j x(d_{ij} - d_c) \tag{1}$$

In the formula, the density value of the first data point represents the distance from the first data point to the first data point, and when d_{ij}-d_c<0, $x(d_{ij}$-$d_c)=1$, otherwise, $x(d_{ij}$-$d_c)=1$,.

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

The δ_i represents the minimum of the distance from the i data point to all points higher than its density value. When the i data point density value is the maximum density value, the default $\delta_i = \max_{j:}(d_{ij})$.

2.2 Perform initial clustering

The algorithm proposed in this chapter not only considers the high-density point with the smallest distance from the point (relative to the point) before clustering each data point, but also considers the high-density point that is second from the point. When it is of the same class, it can be considered that the data point has a relatively clear clustering result, and is the same as the clustering of the recording point, and the data point is listed as a set of data points that can be clearly classified. When the two categories are different from the sub-genus, it is considered that the clustering result of the data point may have a controversy caused by its special position. It may be of the same class or the same class, so the data is Points are included in the set of data points between clusters.

2.3 Identify data points between clusters

After a preliminary classification of 2.2 pairs of data points, all data points can be divided into two categories: the classification of data points and the set of data points between clusters can be clarified. First, according to the proposed clustering principle, points that can be clearly classified in the data point cluster are clustered. Second, the data points in the data point set between the clusters are reclassified:

$$\gamma_i = | Dist_{i.nneith1} - Dist_{i.nneith2} | \quad (3)$$

$Dist_{i.nneith1}$ represents the distance from the *i* data point to the nearest high-density point *nneigh1*, $Dist_{i.nneith2}$ represents the distance from the *i* data point to the next high-density point nneigh2, and γ_i represents the minimum distance of the *i* data point The difference between the small distance and the second. For the different values of γ_i, the following two operations are performed on the data points in the classification point set:

(1) When $\gamma_i > d_c$, the distance between the *i* data point and nneigh1 and nneigh2 is considered to be significantly different, then the data point is at the edge position of the cluster with a smaller distance from *nneigh1* and medium n*neigh2*, so the class of the data point belongs to The cluster should be consistent with the clustering labels that are closer.

(2) When $\gamma_i \leq d_c$, the distance between the *i* data point and *nneigh1* and *nneigh2* is considered to be no significant difference, then the data point is in the middle of the two clusters to which *nneigh1*and *nneigh2* belong.

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

2.4 Class cluster labeling of data points between primary clusters

Before using the bee colony algorithm to find the optimal solution, it is first necessary to determine the most likely class cluster label of the data points between the clusters. For each inter-cluster data point, its most likely associated cluster label can be determined based on its known relationship with *nneigh1* and *nneigh2*.

2.5 Class cluster labeling for determining data points between clusters

Halkidi and Vazirgiannis proposed a classification evaluation index (CDbw) based on the tightness and resolution between clusters [6]. It can effectively measure the classification effect of arbitrarily distributed data points. The algorithm uses the bee colony algorithm with CDbw as the objective function. For each clustering result of data points between each cluster, the optimal solution is obtained according to different CDbw values, and finally the optimal solution is obtained, that is, the most data points between clusters. Good clustering results.

In 2006, the bee colony algorithm was proposed by Karaboga, which is a new optimization method to simulate honey bee collecting behavior[7]. In 2014, Karaboga et al. proved that the bee colony algorithm can be widely applied to feature selection [8], real parameter optimization, job scheduling [9], travel salesman problem [10] and so on.

The clustering result of the data points which can be clearly classified in 2.2 and the clustering result of the data points of the clusters in 2.4 are integrated into the final result, and the clustering result is evaluated by the contour coefficient (Sil) and the F value (F-Measure). .

3. Experimental result

In order to measure the validity and stability of the algorithm, some known data sets and custom 2D data sets are selected as input data on the data set; in the clustering algorithm, with K-Means algorithm, AP algorithm and DPC algorithm Comparison of clustering results. The known data sets used are shown in Table 1:

Table 1 Known data set to be tested

Set name data	Number of data points	Dimension	Number of classes
Flame	**240**	**2**	**2**
Aggregation	**788**	**2**	**7**
R15	**600**	**2**	**15**

In order to evaluate the classification effect of the clustering algorithm proposed in this chapter, the values obtained by the commonly used F-Measure and Sil evaluation indicators are based.

In this paper, three different clustering algorithms (K-means, AP, DPC) are used to map the clustering results of three different known data sets Table 2 and Table 3 (Flame, Aggregation, R15) and the proposed method. The clustering results are compared to evaluate the clustering effect of the proposed algorithm.

Table 2 Values of clustering results for three different data sets

Data set name	K-Means		AP		DPC		Algorithm	
	F-Measure value	*Sil value*	*F-Measure value*	*Sil value*	*F-Measure value*	*Sil value*	*F-Measure value*	*Sil value*
Flame	***0.752817***	***0.5298***	***0.084906***	***0.5201***	***1***	***0.425***	***0.992244***	***0.4291***
Aggregation	***0.715588***	***0.6303***	***0.041432***	***0.6115***	***1***	***0.6419***	***0.992008***	***0.647***
R15	***0.685978***	***0.6969***	***0.020289***	***0.2093***	***0.986671***	***0.8985***	***0.98654***	***0.891***

Table 3 Values of clustering results for three different data sets

Data set name	K-Means	AP	DPC	Algorithm
Flame	***0.5298***	***0.5201***	***0.425***	***0.4291***
Aggregation	***0.6303***	***0.6115***	***0.6419***	***0.647***
R15	***0.6969***	***0.2093***	***0.8985***	***0.891***

4. Result analysis

The algorithm proposed in this paper is based on the assumption of DPC algorithm, but the clustering principle is completely different. This algorithm implements reasonable clustering of data points between clusters. Through experimental comparison and analysis, and the results of the three clustering methods of K-Means, AP, DPC were compared. The algorithm experiments presented in this paper show that for arbitrary distribution of data sets, there are obvious advantages in the identification and classification of data points between clusters.

Acknowledgment

This paper is supported by the Changchun City Philosophical and Social Science Planning Project “Research on the Measurement of Public Happiness and Its Influencing Factors in Changchun City”(No.: CSKT2018ZX-010) , Jilin University of Finance and Economics Doctoral Fund Project

“Research on User Privacy Information Disclosure Behavior and Protection Mechanism in Online Medical Health Service”(No.: 2018B15) and Jilin Province Social Science Fund Project((No.: 2018B79,2017BS28).

References

[1] Amiri M, Eftekhari M, Keynia F. Using Naive Bayes Classifier to Accelerate Constructing Fuzzy Intrusion Detection Systems [J]. International Journal of Soft Computing & Engineering, 2013, 2(6):453-459.
[2] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm [J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.
[3] Alsabti K, Ranka S, Singh V. An Efficient K-Means Clustering Algorithm [C]. Proceedings of Ipps/spdp Workshop on High Performance Data Mining, 1998:9-15.
[4] Frey B J and Dueck D. Clustering by Passing Messages between Data Points [J]. Science, 2007, 315(5814):972-976.
[5] Rodriguez A and Laio A. Machine learning. Clustering by fast search and find of density peaks.[J]. Science, 2014, 344(6191):1492.
[6] Halkid I M,Vazirgiannis M,Batistakis Y. Quality Scheme Assessment in the Clustering Process[C]//Proc of theew 4th Eur Conf Principles and Practice of Knowledge Discovery in Databases,2000:165-276.
[7] Akay B and Karaboga D. A modified Artificial Bee Colony algorithm for real-parameter optimization [J]. Information Sciences, 2012, 192(1):120-142.
[8] Schiezaro M and Pedrini H. Data feature selection based on Artificial Bee Colony algorithm[J]. Eurasip Journal on Image & Video Processing, 2013, 2013(1):47.
[9] Banharnsakun A, Sirinaovakul B, Achalakul T. Job Shop Scheduling with the Best-so-far ABC[J]. Engineering Applications of Artificial Intelligence, 2012, 25(3):583-593.
[10]Yang W and Pei Z. Hybrid ABC/PSO to solve travelling salesman problem[J]. International Journal of Computing Science & Mathematics, 2013, 4(3):214-221.