ATLANTIS PRESS

Advances in Computer Science Research, volume 78

3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018)

K-Means Clustering Efficient Algorithm with Initial Class Center Selection

Huang Suyu, Hu Pingfang

School of Computer Science, Wuhan Donghu University, Wuhan, Hubei

Keywords: Clustering algorithm, particle swarm optimization algorithm; dissimilarity matrix, k-means

Abstract: The algorithm herein adopts density-based method and max-min distance method to define initial clustering center to eliminate the need for defining clustering center in advance in k-means algorithm, and normalize the data set to reduce the influence of fluctuation of attribute value for each dimension of sample set on accuracy of clustering result. Besides, it obtains dissimilarity matrix and takes advantage of good global convergence ability of particle swarm optimization algorithm to improve proneness of K-means algorithm to be trapped in local optimum. The effectiveness of the algorithm was verified via experiment. However, although the algorithm herein performs well in part of small low dimensional data set, while how to effectively make cluster analysis on large high dimensional data still needs to be further researched.

1. Introduction

The partitioning and hierarchical methods in clustering algorithm are the most popular clustering techniques. The partitioning-based clustering algorithms are mainly k-means and its optimization algorithm. The clustering algorithms based on hierarchy include UPGMA and improved algorithm. The k-means algorithm based on objective function suggested by MacQueen in 1967 is a representative partitioning method [2], which is simple, rapid and can effectively process large data set. However, it is sensitive to initial clustering centre, and different clustering centers largely influences the clustering result, it is also prone to be trapped in local optimum. Literature [3] introduces rough set to effectively process fuzzy boundary data, and introduces the concept of kernel to improve clustering accuracy. The particle swarm optimization (PSO) is an optimization algorithm based on swarm intelligence, which has quick convergence speed, simple procedure, fewer parameters needed to be set, can more quickly converge to globally optimal solution, and can avoid degradation of random optimization. [4]. Literature [5] adopts linear decreasing strategy for inertia weight in particle swarm algorithm, which is favorable for balancing global search and local search for particle swarm algorithm, yet the balancing effect needs to be strengthened. The improved particle swarm algorithm is used by many scholars to optimize k-means algorithm. Literature [6] uses strong global optimization ability of particle swarm to improve the proneness of k-means algorithm to be trapped in local optimum, but does not well take advantage of strong local search ability of k-means algorithm.

2. Improved k-means clustering algorithm

K-means algorithm is a widely applied clustering algorithm, but is extremely sensitive to the position of initial clustering center due to its attribute of gradient-descending. Besides, its random selection of clustering centre leads to significant fluctuation of clustering result. To obtain better clustering result, the density parameter is firstly set (MP and ε) to define the data points in high density region, i.e. the data points that are in ε region and are at least MP in number, to obtain a high density region D.

The object in the highest-density region is selected from D as the first clustering center C_1, the high-density point which is the furthest from C_1 is selected as the second clustering center C_2, the distance $d(X_i, C_1)$ of arbitrary data point Xi in region D to C_1 and C_2 is calculated, then the

third clustering center is data point Xi of $\max(\min(d(X_i, C_1), \min(d(X_i, C_2)))$

Thus, the Kth centering center is data point Xi meeting following

$$\begin{aligned} &\max(\min(d(X_i, C_1), \min(Xi, C_2), \\ &\cdots, \min(d(X_i, C_k)))) \end{aligned} \tag{1}$$

Then k initial clustering centers can be automatically defined by analogy.

2.1 Dissimilarity matrix

The dissimilarity matrix among n data points of data set is defined as s, which is a $n \times n$ matrix.

$$D = \begin{bmatrix} 0 & d(1,2) & d(1,3) & \cdots & d(1,4) \\ d(2,1) & 0 & d(2,3) & \cdots & d(2,n) \\ d(3,1) & d(3,2) & 0 & \cdots & d(3,n) \\ \vdots & \vdots & \vdots & 0 & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \cdots & 0 \end{bmatrix} \tag{2}$$

$$d(i,j) = \sqrt{(\left|x_{i1} - x_{j1}\right|^2 + \cdots + \left|x_{id} - x_{jd}\right|^2)} \tag{3}$$

The dissimilarity between arbitrary data points x_i and x_j is $s(i,j)$, the smaller the value is, the larger the similarity between the two points is; the larger the value is, the smaller the similarity is. $d(i,j) = d(j,i), d(i,i) = 0$. So the dissimilarity matrix is symmetric and self-reciprocal.

To organically combine PSO and k-means needs to define operation time of k means.

K-means calculation is not needed in global searching in PSO. When PSO begins to enter the convergence state, K-means calculation is started, thereby starting local search, which effectively improves operation efficiency of mixed algorithm and shortens the time needed for operation of algorithm. The convergence origin of PSO is judged and measured by the global change of all particles' fitness.

Definition: Inclusive fitness variance of particle swarm

$$\sigma^2 = -\sum_{i=1}^{n} \left(\frac{f_i - f_{avg}}{f} \right)^2 \tag{4}$$

Where, n is quantity of particles, f_i is fitness value of the ith particle, f_{avg} is the current average fitness of particle swarm, when $\sigma^2 < m$, m is a certain threshold, indicating PSO has entered the convergence phase and K-mean algorithm can be executed.

$c_1^1 c_1^2 \cdots c_1^d c_2^1 c_2^2 \cdots c_2^d \cdots c_k^1 c_k^2 \cdots c_k^d v_1^1 \cdots v_1^d \cdots v_k^d f$

The particle swarm algorithm is used to optimize k-means algorithm. The encoding mode is based on clustering center, i.e. each particle is potential optimal solution, and the position composed of k clustering centers also contains speed and fitness value of particle. Set the data set as having N data points, and set attribute dimension number of data point as d, then the position and speed of data point can be expressed by $N \times d$ matrix. The global coding way of each particle is: $c_1^1 c_1^2 \cdots c_1^d c_2^1 c_2^2 \cdots c_2^d \cdots c_k^1 c_k^2 \cdots c_k^d v_1^1 \cdots v_1^d \cdots v_k^d f$

The concrete algorithm flow is as follows:

Step 1: Initiation of population: First randomly classify the data point into a certain classification as the initial clustering classification, calculate the clustering center of different classifications as position code of initial particle, calculate fitness of particle as code of particle's individual optimal solution, randomly initialize the speed of particle, repeat N times to generate initialization population with N particles.

Step 2: Individual optimization: Compare the fitness of each particle with the best fitness it has

experienced, if the result is better, update the position of this particle.

Step 3: Swarm optimization: Compare the fitness of each particle with the optimal fitness value of swarm, if the result is better, update the global best position.

Step 4: Update position and speed: Update position and speed of particle according to equation (2) and (3)

Step 5: Judge whether particle swarm converges according to swarm fitness variance, i.e. equation (8). If yes, output the clustering partitioning contained in particle with the optimal fitness value.

Step 6: Code of clustering center is obtained according to optimal particle, then define clustering partitioning with data concentration according to the nearest-neighbor principle.

Step 7: Re-calculate new clustering centre according to k-means algorithm, and re-partition the data set

Step 8: Repeat step 7, until convergence law achieves convergence or iteration achieves the maximal number of iterations, and now cycle can be ended.

2.2 Experiment and analysis

To further verify effectiveness of algorithm herein, the k-means algorithm, PSO+k-means algorithm and the paper's algorithm were tested on Iris and Wine data sets of UCI standard data set to calculate its accuracy. Table 1 shows the characteristic indices of Iris and Wine data sets. Table 2 compares the clustering accuracy after the 3 algorithms run on the two data sets. Fig.1 and 2 are respectively convergence curves of fitness value for the 3 algorithms on Iris and Wine data sets.

Table 1 Statistics of information of data sets used in the experiment

Data set	Sample number	Classification number	Attribute number	Classification distribution
Iris	150	3	4	50,50,50
Wine	178	3	13	59,71,48

Table 2 Comparison of clustering correctness of the three algorithms

Algorithm / Data set	k-menas	PSO+k-means	Algorithm herein
Iris	0.7812	0.8953	0.9285
Wine	0.6821	0.7334	0.9317

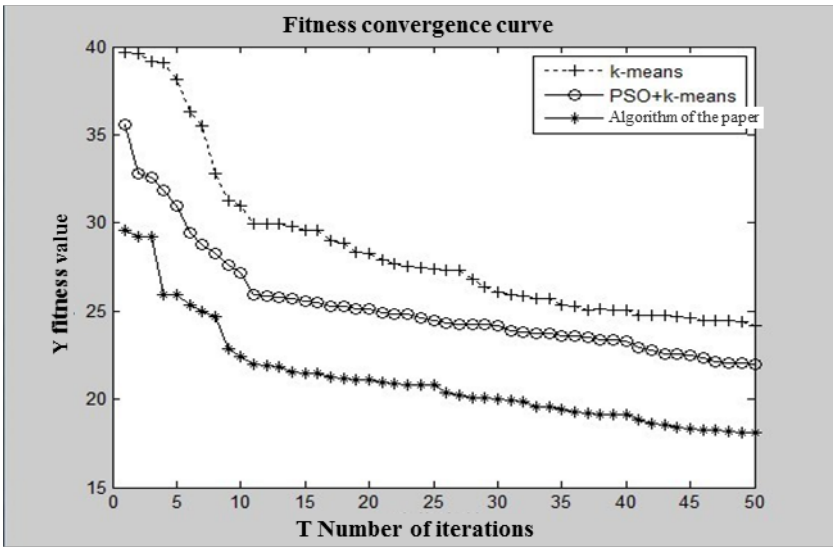

Fig.1. Fitness convergence curve of the three algorithms on iris data set

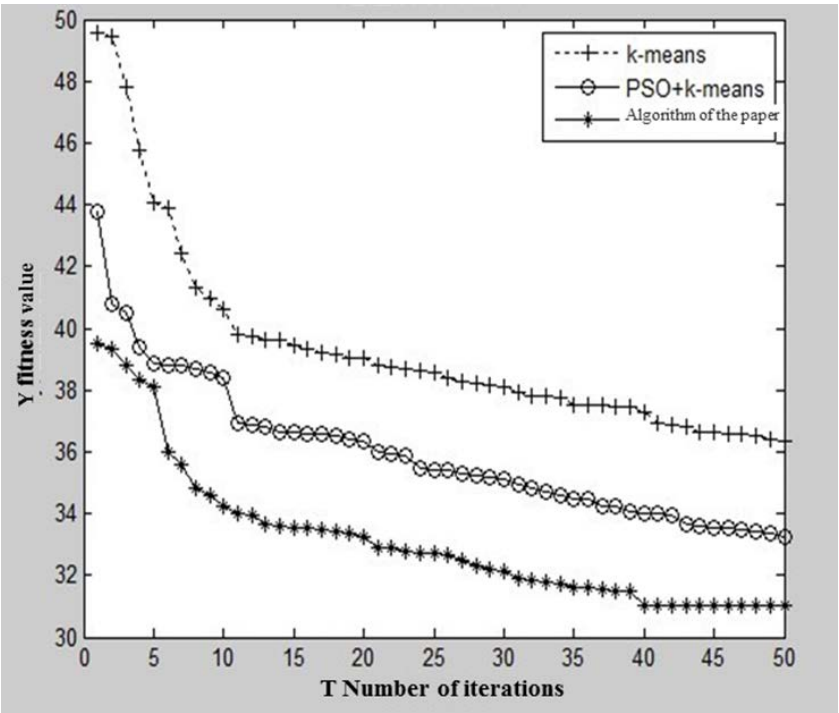

Fig.2. Fitness convergence curve of the three algorithms on wine data set

From the experiment: the experimental result in table 2 shows that on Iris data set, the paper's algorithm has an accuracy of 92.85%, 89.53% for PSO+k-means algorithm and 78.12% for k-means algorithm; on Wine data set, the paper's algorithm also outperforms PSO+k-means algorithm and k-means algorithm in accuracy. This shows that the clustering accuracy of PSO+k-means algorithm is higher than that of k-means algorithm, which indicates that particle swarm algorithm improves the disadvantage of proneness to be trapped in local optimum in k-means algorithm, while the accuracy of the paper's algorithm is higher than that of PSO+k-means algorithm, indicating the optimization of inertia weight of basic particle swarm algorithm improves clustering result by calculating dissimilarity matrix. It is observed from Fig.4 and 5 that on Iris and Wine data set, the clustering convergence rate of the paper's algorithm is quicker than that of PSO+k-means, and PSO+k-means is quicker than k-means. Besides, in terms of optimal fitness value of criteria function, the paper's algorithm outperforms PSO+k-means, and PSO+k-means outperforms k-menas, indicating the algorithm herein has a stronger ability of global search and convergence.

3. Summary

To cope with the k-means algorithm's drawbacks of sensitivity to initial clustering centre and proneness to be trapped in local optimum, the paper proposes an improved clustering algorithm based on particle swarm, which defines initial clustering centre by combining the density-based method and max-min distance method to solve the sensitiveness of k-means to initial value. Then the strong global optimization ability of particle swarm algorithm is used to avoid k-means being trapped in local optimum. The mixed algorithm is further improved by normalizing attributes of each dimension of sample set, decreasing inertia weight by degrees with concave function, calculating dissimilarity matrix, introducing swarm fitness variance. The experimental result shows the algorithm herein has higher accuracy and stronger convergence ability.

Acknowledgements

The Youth Natural Science Fund Project of 2017 Wuhan Donghu University.

References

[1] Weisen Pan, Shizhan Chen, Zhiyong Feng. Investigating the Collaborative Intention and Semantic Structure among Co-occurring Tags using Graph Theory. 2012 International Enterprise Distributed Object Computing Conference, IEEE, Beijing, pp. 190-195.

[2] Du X, Zhen S, Peng Z, Zhao C, Zhang Y, Zhe W, Li X, Liu G, Li X. 2017c. Acetoacetate induces hepatocytes apoptosis by the ROS-mediated MAPKs pathway in ketotic cows. Journal of Cellular Physiology 232: 3296-3308, 2017.

[3] Yingyue Zhang, Qi Li, William J. Welsh, Prabhas V. Moghe, and Kathryn E. Uhrich, Micellar and Structural Stability of Nanoscale Amphiphilic Polymers: Implications for Anti-atherosclerotic Bioactivity, Biomaterials, 2016, 84, 230-240.

[4] Stephygraph, L.R., Arunkumar, N., Venkatraman, V. Wireless mobile robot control through human machine interface using brain signals (2015) 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings, art. no. 7225484, pp. 596-603.

[5] Sun X, Xue Y, Liang C, Wang T, Zhe W, Sun G, Li X, Li X, Liu G. 2017. Histamine Induces Bovine Rumen Epithelial Cell Inflammatory Response via NF-κB Pathway. Cellular Physiology & Biochemistry 42(3):1109-1119.