TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.3, March 2014, pp. 1917 ~ 1923
DOI: http://dx.doi.org/10.11591/telkomnika.v12i3.4262

Improved K-means Clustering Algorithm Based on Genetic Algorithm

Tang Zhaoxia*, Zhang Hui
Faculty of Computer Engineering, Huaiyin Institute of Technology,
Huai'an 223003, P.R.China
*Corresponding author, e-mail: zx-tang@163.com

Abstract

Through comparison and analysis of clustering algorithms, this paper presents an improved K-means clustering algorithm. Using genetic algorithm to select the initial cluster centers, using Z-score to standardize data, and take a new method to evaluate cluster centers, all this reduce the affect of isolated points, and improve the accuracy of clustering. Experiments show that the algorithm to find the initial cluster centers is the same location, objective function value is smaller, the clustering effect is better and more stable when it has the outlier data, and it applies not only to simple data sets, but also to more complicated data sets.

Keywords: *K-means clustering algorithm, genetic algorithm, isolated points*

1. Introduction

Clustering is an unsupervised classification, in accordance with a specific standard to a data set divided into different classes or clusters, makes the same data within a cluster as large as possible similarity, at the same time not in the same cluster, the data differences are as large as possible. As an important data analysis techniques, Clustering can greatly improve the efficiency of the algorithm. Therefore, clustering has become very important field of data mining. In the market or customer segmentation, pattern recognition, spatial data analysis, web document classification and other fields, clustering has been widely studied and applied [1].

Clustering has been carried out forty years, it has gone through three stages of development in foreign，and classification in the mail, electronic conferences, information filtering and achieved a relatively wide range of applications. Clustering in our country was studied from 1980's，and it has been trying to develop a number of automatic classification and clustering systems. Clustering algorithm can be divided into [2]:

1.1. Grid-based Clustering Algorithm

Grid-based clustering algorithm divided the object space into limited number of units, to form a grid structure. All clustering operations are carried out in the grid structure. This has no direct advantage of the data processing, second the increase in data grid allows data from grid-based clustering of the order. Clustering quality depends on the lowest level of granularity of the grid structure, if the relatively small size, clustering operation costs will increase significantly, however, if the lowest level of granularity is too coarse will reduce the quality of cluster analysis. Therefore，when the space dimension big, they will naturally have a huge number of grid cells to the high computation complexity. Typical grid-based clustering algorithms are CLIQUE algorithm and STING algorithm [3].

1.2. K-means Clustering Algorithm

K-means clustering algorithm was a classical clustering algorithm, proposed in 1967 by the J.B.MacQueen [4]. K-means clustering algorithm is a method of clustering algorithm based on partition，K data were first randomly selected as the initial center of K clusters (the objects in the cluster average), then the rest of the data object according to its distance from the center of each cluster assigned to the nearest cluster, re-calculated for each cluster centers. This process is repeated, until the objective function is minimized. Thus seen K-means clustering algorithm

Received August 17, 2013; Revised September 29, 2013; Accepted October 17, 2013

for the K value must be determined in advance，secondly, the final clustering results depend on the selection of cluster centers, different centers were randomly selected will produce different clustering results, and the clustering results may be very large differences.

1.3. Fuzzy C Means (FCM) Clustering Algorithm

Fuzzy C Means (FCM) clustering algorithm proposed by Dunn, and developed by Bezdek [5], FCM is the application of the most widely used clustering algorithm. However, FCM clustering algorithm is essentially a hill climbing local search，on the initial cluster centers or membership matrix sensitive, not guarantee convergence to global optimal solution, in addition FCM clustering algorithm in the treatment of high dimensions, large data sets slower rate of convergence, makes the practical application of FCM clustering algorithm is subject to certain limitations.

1.4. Clustering Algorithm Based on Neural Network

Clustering algorithm based on neural networks will be described as a sample for each cluster, as a prototype of the cluster sample. According to some distance measure, the new object can be allocated to its most similar cluster sample. Neural network includes Rumelhart competitive learning neural networks and Kohonens self-organizing maps (SOM) neural network. In the SOM algorithm，clustering is also a competition through a number of units for the current object. Weight vector closest to the current object as the winning unit cell. In order to enter the object closer, of the winning unit and its nearest neighbors to adjust the weights which include data to be improvement in the type of expansion and efficient algorithms designed to handle large data sets to the clustering.

Among these clustering algorithms, each clustering algorithm has its own shortcomings, such as easy to fall into local optimum, the initial point through the high sensitive and complex. K-means algorithm is a classical algorithm to solve clustering problem, it is simple, quick, easy to improve, and the relatively high efficiency of the implementation. However, the algorithm is susceptible interference from noise and isolated points, lead to the next generation of the deviation of the cluster centers, and ultimately affect the clustering effect. There is no basis standard for the number of categories K value, sometimes K value is difficult to estimate, but the inaccuracy of K value will affect the quality of clustering. So the algorithm need to improve, this paper presents an improved K-means clustering algorithm to improve running efficiency, scope of application and accelerate the convergence rate of the algorithm.

2. Improved K-means Clustering Algorithm

2.1. Improved Initial Cluster Centers

The current method to select the initial cluster centers is: select K cluster centers from the pending clustering data arbitrarily, find out the K farthest point as the initial cluster centers by calculating the distance, then calculate the average value as the K initial cluster centers, all pending clustering the data are divided into K classes. Document [6] proposed to select K farthest points from the data set as the initial cluster centers by solving similarity. Document [6] proposed the method is first determine the furthest distance between two data points, then the data set is divided into K segments, calculate each mean as K initial cluster centers [7]. The experiments show that these methods not only improve the clustering efficiency but also become more accurate. In this paper, use the genetic algorithm to select the initial cluster centers.

1) Chromosome coding

Encoding methods are divided into three categories: binary coding, floating-point coding and symbol coding. There are two main genetic K-means clustering algorithm coding: one, randomly generate K cluster centers as chromosomes; Second, each data belongs the cluster class number as the chromosomal gene value, so that the chromosome length corresponding to the total number of the dataset. When the clustering data set is long, the cluster computing of this encoding will be very complex, it will reduce the efficiency of the algorithm. So this article choose the first coding method, the chromosome length is shorter, facilitate the calculation and understanding, it can improve the efficiency of clustering.

2) Initial population

Population size M is the number of individuals contained in the initial population, When M value is larger, it will reduce the search efficiency of the genetic algorithm, When M value is small, it can improve the searching speed, but it can reduce the diversity of the population, may lead to the premature phenomenon. So M-value general range is [20-100]. The operation of initial population is as follows:

(1) Set the number of individuals M

(2) do

Randomly select K cluster centers from the initial population as one chromosome

(3) Until population size reached M

As the initial K is not the best number of clusters, therefore, it can take individual of the largest fitness value as a sample of other individuals，the length of other individuals (K value) to learn from. If K value is increasing trend, for then increased gene, the farthest point from the largest cluster center of the current individuals are added. If K value is reducing trend, then merger this two cluster centers if two cluster centers is the minimum distance of all cluster centers, the new cluster centers after the merger is average of two cluster centers. It can significantly improve the real possibility of clustering results according to individual fitness learn the K value automatically.

3) Single-point crossover

Genetic cross is an important feature different from other algorithms. Cross is the main method to generate new individuals, and directly affect the ability of algorithm global search. Methods of cross commonly used are: single-point crossover, two-point crossover, uniform crossover, discrete crossover and arithmetic crossover [8]. Because chromosome length is not long, so in this paper take single-point crossover to protect the excellent individual. The steps are as follows:

(1) Randomly divided group into groups of two individuals, the population size is M, formed M / 2 individual groups

(2) For each group of individuals, set a gene Location for intersection by the random function

(3) For each group of individuals, swap the gene of the two chromosomes according to the crossover probability in determining the intersection point, it can format two new chromosomes

(4) Calculate two new individuals' fitness, compare the fitness with parent, choose two chromosomes from these four chromosomes

4) Choice

Choice is to improve the global convergence and computational efficiency. The choice of methods used: roulette wheel, strategy and ranking selection method etc. Simply using the basic genetic algorithm, lead to poor convergence of the algorithm. This paper introduces the concept of difference degree, by calculating the difference degree between cluster centers to limit individuals with poor fitness, when child difference degree must be greater than the parent individuals' difference degree, child can choose to genetic operation. To optimize the efficiency of genetic algorithm, let individual $C_j(i) = \{C_1 i , C_2 i ,\dots , C_k i\}$, the individuals' difference degree:

$$d(C_j(i)) = \frac{a}{k(k-1)} \sum_{x=1}^{k} \sum_{y=1}^{k} \left\| c_x i - c_y i \right\| \qquad 0 < a < 1 \tag{1}$$

If the individual is a quality individual，then the distance between cluster centers，difference degree large; if the individual is a poor individual，is very close to the cluster center，difference degree is small.

5) Variation

Variation can increase the diversity of the population，to prevent the occurrence of premature, increased the algorithm capacity of local random search[9]. However, genetic variation may exceed the range，this paper propose a new adaptive method. Let variation individual's gene: G_{ij} denotes value of the i cluster center of the j-dimensional, the variation of individual genes G_{ij}' as:

$$G_{ij}' = G_{ij} + m (GC_{ij,max} - G_{ij}) \qquad m >= 0 \tag{2}$$

$G_{ij}' = G_{ij} + m(G_{ij} — G_{ij,min})$ m<0

Where，$G_{ij,max}$ and $G_{ij,min}$ respect maximum and minimum of the i cluster center of j-dimensional, m generate uniformly distributed within [-(f-fmin)/(fmax-fmin), (f-fmin)/(fmax-fmin)], f is variation individual fitness, fmin and fmax respect the best and the worst fitness of population. In this way, the worst large individual participate in variation，the best individual does not participate in variation，to achieve adaptive variation of genes. Thus ensuring the clustering performance and the convergence rate faster.

6) Evaluation of individual fitness

Fitness function is the main basis for survival of the fittest individuals. Fitness function related to the quality of the next generation of population and number of，let the data set is divided into K classes, the fitness function:

$$f = \frac{\min \left\| c_i - c_j \right\|^2}{\frac{1}{k}\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j}\left\| x_j - c_i \right\|^2 / n_j\right)} \quad (3)$$

Among them, the molecule is the minimum distance between cluster centers，should be as large as possible; the denominator is the average distance within the cluster center, it cloud be as small as possible. Fitness function reflects the distance between cluster centers should be loose, the cluster centers should be as compact as possible.

7) Algorithm flow

(1) Initialization: the population size M, the maximum number of iterations T, the number of clusters K etc
(2) Randomly generated m chromosomes, one chromosome represents a set of initial cluster centers, it can form initial population
(3) Take K-means clustering of the data according to the initial cluster centers, then calculate each chromosome adapt according to the clustering results, retain the best
(4) Select population, take crossover and mutation operations to produce a new generation of groups
(5) If the average fitness of successive generations of individual differences in less than a minimum threshold (a smaller constant ε) or genetic algebra T= Gmax, algorithm stop to (6), otherwise to (3)
(6) Calculate the fitness of the new generation of group, identify the highest fitness individuals
(7) Take K-means clustering according to the initial cluster centers of the highest fitness chromosome for output clustering results.

2.2. Improved the Affect of Isolated Points

The property values may be quite different in pending clustering data set, such large value can affect the distance between the data attribute, it can lead to the failure of the cluster. Therefore, the data first take z-score normalization before performing clustering [10], use the following formula:

$$V' = \frac{V - \overline{A}}{\sigma_A} \quad (4)$$

Where A is the property, $\overline{A}$ is the mean, σ_A is the standard deviation, By the above formula, the value V can be normalized to V'. In order to reduce the impact of isolated points, this paper take a new evaluates methods of cluster centers to improve clustering accuracy. The concrete steps are as follows:

(1) Randomly select K clustering data as the initial cluster centers
(2) Do

Calculate the distance of each data with the K-th cluster centers, and this data is assigned to the cluster that it's nearest center point represent
Calculate the average distance d for all the data in the cluster with the cluster centers;
Select cluster center distance less than 2d data objects form a subset
Calculate the average value of each subset as the next round of cluster centers

(3) Until all cluster centers no longer change.

When the cluster has isolated points, the maximum distance between the cluster data and cluster centers is larger than the average distance between all data and cluster centers. Then twice the average both includes most of the data and rule out the isolated points. When the data is relatively dense, the maximum distance between the cluster data and cluster centers is smaller than the average distance between all data and cluster centers. Twice the average can include almost all of the data in the cluster. So using the above method can get better clustering effect, whether the data set have isolated points.

3. Experimental Results and Analysis

In order to verify the effectiveness of the algorithm in this paper, improved K-means clustering based on genetic algorithm will compare with the original K-means algorithm and two known algorithm. The algorithm parameters are set as follows: k=3, pc1=0.9, pc2=0.6, pm1=0.5, pm2=0.1, m=50, Maxgens=100.

Experiment 1: The case of non-isolated points

The experimental data included a set of manual data and the iris data from the UCI database, artificial data as shown in Table 1. Respectively from the initial cluster centers, the number of iterations, the minimum objective function value and the running time of the K-means algorithm, the two existing improved algorithm of K-means and this algorithm were validated, experimental results were shown in Table 2 and Table 3.

Table 1. Manual Data (k=3, n=15)

X	0	0	1	1	1	2	2	2	3	6	6	7	7	8	8
Y	0	1	0	2	5	1	2	5	6	6	7	6	8	7	8

Table 2. Manual Data of Non-isolated Points

Clustering algorithm	Initial cluster centers	The number of iterations	The minimum objective function value	The running time (s)
K-means 1	(3,2,1)	4	5.64	15
K-means 2	(4,8,5)	5	5.64	9
K-means 3	(7,0,9)	2	5.64	5
K-means 4	(13,8,11)	4	7.57	9
K-means 5	(11,13,14)	3	7.57	5
K-means 6	(8,1,12)	2	5.64	5
K-means 7	(11,8,9)	4	7.57	4
K-means 8	(13,8,10)	4	7.57	4
K-means 9	(5,2,13)	4	5.64	7
K-means 10	(7,5,14)	2	5.64	4

Table 3. Iris Data of Non-isolated Point

Clustering algorithm	Initial cluster centers	The number of iterations	The minimum objective function value	The running time (s)
K-means 1	(143,70,335)	7	129.4	14
K-means 2	(26,109,76)	9	129.4	10
K-means 3	(69,66,84)	12	129.4	5
K-means 4	(10,110,61)	5	129.4	7
K-means 5	(50,118,55)	12	129.5	5
K-means 6	(82,80,7)	12	129.4	5
K-means 7	(118,58,59)	9	129.5	5
K-means 8	(4,118,103)	11	129.4	6
K-means 9	(45,92,1)	4	145.4	5
K-means 10	(70,105,145)	11	135.6	6

Experiment 2: the case of isolated points

Several isolated-points were added to the above data, Manual data added a group (15, 15), the iris data added five groups (10,3.0,1.5,5), (5.8,3.6,20,0.2), (0, 0, 0, 0), (9.0,6.6,14,0), (6.9,9,1.4,9), experimental results were shown in Table 4 and Table 5.

Table 4. Manual Data of Isolated Points

Clustering algorithm	Initial cluster centers	The number of iterations	The minimum objective function value	The running time (s)
K-means 1	(8,10,5)	2	7.13	7
K-means 2	(2,9,8)	2	7.13	5
K-means 3	(6,2,9)	3	7.13	6
K-means 4	(7,14,4)	2	7.13	6
K-means 5	(8,10,6)	4	7.13	5
K-means 6	(12,15,1)	3	7.13	15
K-means 7	(11,14,11)	4	7.13	4
K-means 8	(10,12,0)	5	7.13	7
K-means 9	(13,11,3)	4	8.34	7
K-means 10	(11,15,14)	5	8.34	8

Table 5. Iris Data of Isolated Points

Clustering algorithm	Initial cluster centers	The number of iterations	The minimum objective function value	The running time (s)
K-means 1	(70,6,8)	9	123.8	10
K-means 2	(44,81,96)	11	123.8	7
K-means 3	(84,5,122)	9	123.8	7
K-means 4	(130,128,75)	13	123.8	8
K-means 5	(15,78,40)	5	140.5	8
K-means 6	(62,73,32)	9	123.8	8
K-means 7	(102,58,125)	13	123.8	5
K-means 8	(42,48,97)	8	139.4	5
K-means 9	(45,54,30)	7	140.4	4
K-means 10	(18,105,145)	12	125.6	8

It can be seen from the two experiments, when there was no isolated point data, the proposed algorithm can get more accurate clustering results. When there was isolated point, the proposed algorithm can be significantly reduced the impact of an isolated point.

Experiment 2: Convergence performance comparison

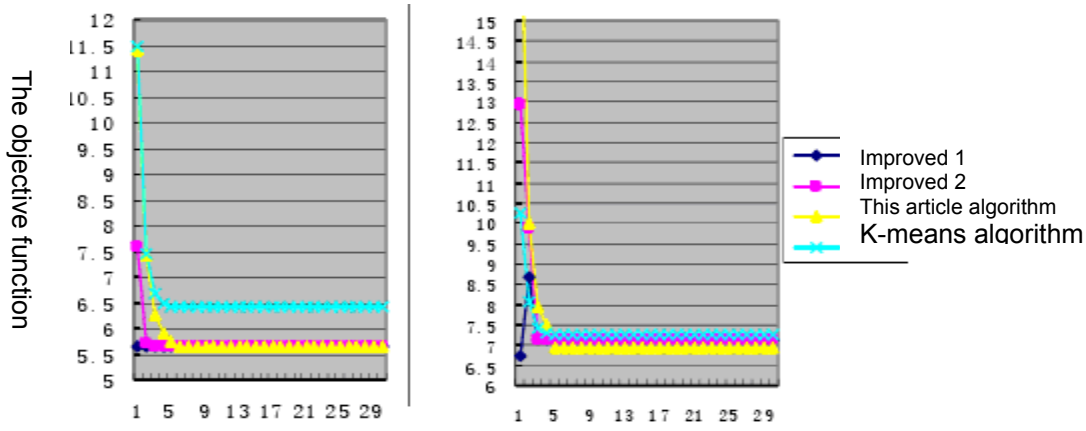

Figure 1. The Number of Iterations

Figure 1 gives a two-dimensional graph of objective function value changing with the number of iterations in each case, it is the convergence of the objective function curve. In both

cases, several algorithms convergence speed changed little. For more complex data sets, the proposed algorithm convergence rate is relatively not fast, however, the proposed algorithm is relatively converges to a minimum value of the objective function, the effect of clustering is best, the other algorithms precocious in a larger objective function value. Algorithm objective function value of iterations are smaller than other algorithms, it indicates that the proposed algorithm has better performance.

4. Conclusion

This article main innovation is: used the genetic algorithm to select the initial cluster centers, improved the affect of isolated points by data preprocessing. The experimental results show that regardless of whether there are isolated points, the algorithm is better than the other three algorithms in the number of iterations and objective function value. The proposed algorithm finds out more favorable initial cluster centers, the clustering effect is better and more stables, it is not only suitable for simple data set, but also for the slightly more complex data set.

References

[1] E Bagheri, H Edldari. *Dejong Function Optimization by Means of a Parallel Approach to Fuzzified Genetic Algorithm.* Computers and communications, ISCC 06 Proceedings. 2006; 675-680.
[2] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press.1981: 62-78
[3] Zhou Hong-wei, Yuan Jin-hui, Zhang Lai-sun. Improvement strategies of Genetic algorithm "premature" phenomenon. *Computer Engineering*. 2007: 33(19): 201-203
[4] Kennedy J, Eberhart RC. *Particle swarm optimization*. Institute of Electrical and Electronics Engineers. 1995; 1942-1948.
[5] Gao Yuelin, Duan Yuhong. An adaptive particle swarm optimization algorithm with new random inertia weight. *Communications in Computer and Information Science*. 2007; 342 -350.
[6] Tang zhaoxia, Zhang hui, Xu dongmei. Image Retrieval based on Improved PSO Algorithm and Relevance Feedback. *Computer Science*. 2011; 38(10): 279-285.
[7] A Goyal, B Verma. *A Neural Network based approach for the Vehicle classification*. IEEE Symposium on Computational Intelligence in Image and Signal Processing. 2007; 226 -231.
[8] Ze zhiChen, Pears N, Freeman M, Austin J Road. *Vehicle Classification using Support Vector Machines*. IEEE International conference on Intelligent computing and Intelligent Systems. 2009; 214 - 218.
[9] Tang zhaoxia, Zhang hui. Vehicle Recognition Based on Particle Swarm Optimization and Decision Tree, *JDCTA (Journal of Digital Content Technology and its Applications)*. 2012; 6(6): 860.
[10] Hong qi Li, Xu He, Xiaolong Xie, Li li, Jinyu Zhou, Xiongyan Li. A New Boundary Condition for Particle Swarm Optimization. *Journal of AICIT (Journal of Convergence Information Technology)*. 2010; 5(9): 215-221.