Improved K-means Clustering Based on Genetic Algorithm

Wang Min
North University of China
School of Electronics and Computer Science and Technology
Taiyuan City, Shanxi Province, China
lansebingxi@163.com

Yin Siqing
North University of China
Software School
Taiyuan City, Shanxi Province, China
yinsq@163.com

***Abstract*—The K-means algorithm is widely used because of its reliable theory, simple algorithm, fast convergence and it can effectively handle large data sets. However, the traditional K-means algorithm is sensitive to the initial cluster centers; make the average of all objects in the same class as cluster centers, so clustering results is largely affected by the isolated points. To address the problems, search the initial cluster centers of K-means algorithm used of genetic algorithms, improve the K-means algorithm to reduce the impact of isolated points, the data showed that it has good results.**

Keywords-K-means algorithm; initial cluster center; genetic algorithms

I. INTRODUCTION

K-means algorithm is a basic partition method in cluster analysis, but it is sensitive to the initial cluster centers, different cluster centers generate quite different clustering results, and it is largely affected by the isolated points [1][2]. Genetic algorithm is an adaptive global optimization search algorithm, formatted by simulating the principle of survival of the fittest, survival of the fittest in natural environment. It mainly include genes coding, fitness calculations, creating the initial population, determine the evolutionary operation etc, which mainly include selection, crossover and mutation [3][4]. This article will use the adaptive search ability of genetic algorithm to find the initial cluster centers of K-means, and improve the K-means algorithm using the method proposed by Li Yeli, Qin Zhen to reduce the impact of isolated point, the experiment showed that the method achieves good results.

II. K-MEANS ALGORITHM ANALYSIS

K-means algorithm is a basic partition method in cluster analysis, but it is sensitive to the initial cluster centers, improper choice of cluster centers will result in cluster failure; K values need to artificially pre-determined, which is very difficult for those who has no experience; affected by the isolated points, each round of calculation of the cluster center has deviation and eventually lead to cluster failure. At present people mainly using genetic algorithm to determine the K value and achieved good results.

For selecting the initial cluster centers, the method proposed in [1] is to select K cluster centers which is the farthest away from each other by calculate similarity between each other[1]. In the experiment, select the 3, 4 attributes of iris data, it reached good results only has six errors. The method proposed in [5] is to determine the farthest distance among data objects, then divide the data set into k segments and calculate the mean within each section as the initial cluster centers [5]. Experiment proved that this method not only can improve the cluster efficiency but also make the cluster result more accurate.

Currently, the method to reduce the impact of isolated points are: one is proposed in [1] that is to filter out m objects which has the biggest sum of distance with other objects in order to rule out isolated points; Second, proposed in [6], the method is to calculate the average value of subset whose object is more close to center as a new round of cluster center.; Third, also proposed in [6], the method is to calculate the maximum distance and the minimum distance between date object and cluster center within same cluster in the k-1 round first, then to calculate mean of subset as cluster center whose object's similarity with the cluster center is greater than (max + min) / 2 [6] .When the data has no isolated points, taking the first approach would exclude the right data objects, resulting in clustering error, while the third method would lead to increase the number of clustering iterations. Experiment proved that the second method is better, it will not only rule out the effects of isolated points but also can calculate a new round of cluster centers more accurate, this method can also get accurate result when the data has no isolated points.

III. K-MEANS CLUSTERING BASED ON GENETIC ALGORITHM

A The basic idea about selecting initial cluster centers using genetic algorithm

In the proposed algorithm, we first use random function to select K data objects as initial cluster centers to form a chromosome, a total of M chromosomes selected, then have K-means operation on each group of cluster center in the initial population to compute fitness, select individuals according to the fitness of each chromosome, select high-fitness chromosomes for the crossover and mutation operation eliminating low fitness chromosomes, format next generation group finally. In this way, within each new generation of groups, the average fitness are rising, each cluster center is closer to the optimal cluster center, and finally select chromosome that have the highest fitness as the initial cluster center.

Algorithm flow chart shown in Figure.1 as follows:

B Chromosome Coding

In this paper, we use real-coded, the value of chromosome gene corresponds to cluster center number,

978-1-4244-7237-6/10/$26.00 ©2010 IEEE

length of the chromosome is the number of cluster, and the specific code form is:

$$X= (X_1, X_2, \ldots\ldots X_k) \quad (1)$$

K is the number of cluster center of a chromosome.

C Population Initialization

The range of M is 20-100. Specific operation is as follows: select K cluster centers randomly to form a chromosome Ran, if the center randomly selected has already exist in the same chromosome, then remove the center and reselect until it reaches K centers, until the population size to M.

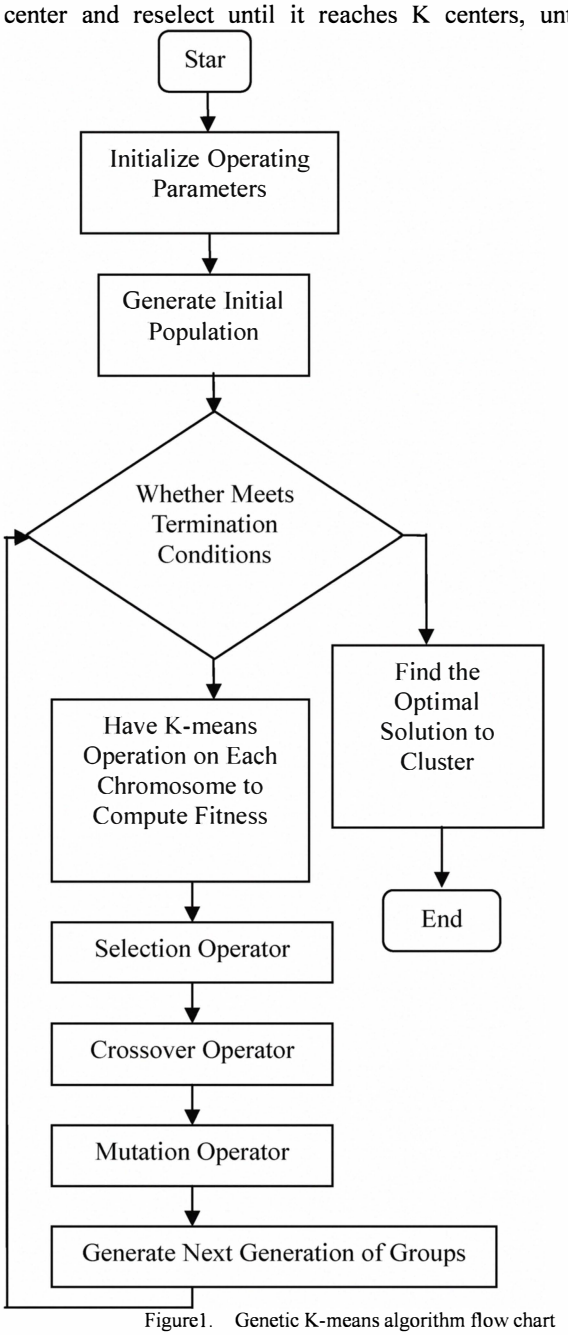

Figure1. Genetic K-means algorithm flow chart

D Select the Fitness Function

This paper use the inverse of objective function J as the fitness function, that is

$$F=1/J \quad (2)$$

The smaller J is, the greater fitness function will become, so the better clustering effect is.

E Genetic Operation

This paper use proportional selection operator, single-point crossover operator and uniform mutation operator. To avoid premature or slow convergence phenomenon using a fixed probability, this paper use self-adaptive genetic operator, that is dynamically adjust the crossover rate and mutation rate. P_c and P_m is calculated as follows:

$$p_c = \begin{cases} p_{c1} - \dfrac{(p_{c1}-p_{c2})(f_{ave}-f')}{f_{\max}-f_{ave}}, & f' > f_{ave} \\ p_{c2}, & f' \le f_{ave} \end{cases} \quad (3)$$

$$p_m = \begin{cases} p_{m1} - \dfrac{(p_{m1}-p_{m2})(f-f_{ave})}{f_{\max}-f_{ave}}, & f \ge f_{ave} \\ p_{m2}, & f < f_{ave} \end{cases} \quad (4)$$

Among them, f_{ave} means average fitness value of each generation group; $f_{\max}$ means the largest individual fitness value in the group; f' means the larger fitness value of the two crossing individuals; f indicates the fitness value of mutating individual. The formula makes individuals with high fitness have lower crossover rate and mutation rate; individuals with small fitness have a higher crossover rate and mutation rate. This helps protect the best individual, but also can make individuals with lower fitness cross and mutate at higher rate, producing excellent model [7].

F Elitist Strategy

The elitist strategy in this article are as follows: if the highest fitness in current group is larger than the best individual's fitness so far, then use the best individual in the current group as the new best individual so far., otherwise, replace the worst individual in current generation with the best individual so far .

G Loop Termination Conditions

In this paper, we use termination algebra T as running end condition of genetic algorithm, which indicate that the genetic algorithm stop running after it runs to the specified evolution algebra, and output the best individual in current group as optimal solution of the problem. Generally range from 100 to 1000.

H Description of the Specific Algorithm

a) Set the parameters: population size M, the maximum number of iteration T, the number of clusters K, etc.

b) Generate m chromosomes randomly,a chromosome represents a set of initial cluster centers, to form the initial population.

c) According to the initial cluster centers showed by

every chromosome, carry out K-means clustering, each chromosome corresponds to once K-means clustering, then calculate chromosome fitness in line with clustering result, and implement the optimal preservation strategy.

d) For the group, to carry out selection, crossover and mutation operator to produce a new generation of group.

e) To determine whether the conditions meet the genetic termination conditions, if meet then withdrawal genetic operation and turn 6, otherwise turn 3.

f) Calculate fitness of the new generation of group; compare the fitness of the best individual in current group with the best individual's fitness so far to find the individual with the highest fitness.

g) Carry out K-means clustering according to the initial cluster center represented by the chromosome with the highest fitness, and then output clustering result.

Ⅳ. EXPERIMENTAL RESULTS AND ANALYSIS

Comparing K-means algorithm based on genetic algorithm (the article) with the original K-means algorithm and two known improved algorithms to verify the effectiveness that selecting initial cluster center using genetic algorithm. Improved algorithm 1 is proposed in [5], the improved algorithm 2 is proposed in [1]. In order to exclude the impact of isolated points, the article use the method proposed in [6] that using the average value of subset whose object is more close to center as a new round of cluster center to improve K-means algorithm, and also apply this method to original K-means algorithm, improved algorithm 1 and improved algorithm 2, having a comprehensive comparison on them.

Experimental data are two groups of data from the UCI database, iris data sets and wine data sets. Before clustering, we first have z-score standardization on data to prevent clustering failure, considering that great value properties affect the distance between the samples.

We added groups of isolated points respectively to the two sets of data above-mentioned. Iris data, adding five isolated points (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); wine data, adding five isolated points (0,0,0,0,0,0,0,0,0,0,0,0,0), (13.34,94,2.36,1.7,110,0.55,0.42,3.17,1.02,1.93,750,5.36,666),(14.34,1.68,2.7,25,98,2.8,31,0.53,2.7,13,0.57,1.96,666),(14.2,1.76,2.45,15.2,1.12,3.27,3.39,0.34,1.97,6.75,1.05,2.85,450),(12.67,0.98,2.24,18,99,2.2,1.94,0.3,1.46,2.62,123,3.16,450).

Experiment parameter settings are as follows: k = 3; pc1 = 0.9; pc2 = 0.6; pm1 = 0.5; pm2 = 0.1; pc = 0.6; pm = 0.1; m (initial population size) = 50, maxgen (the maximum number of iteration) = 100. The results are showed in Table Ⅰand Table Ⅱ, the initial cluster center values are data object label.

In the experimental results, for improvement 1, due to calculating the mean within each separate section as the initial cluster centers it is not marked. Can be seen from the above data, the traditional K-means algorithm is sensitive to the initial cluster centers, different cluster centers have quite different clustering results, and the results sometimes are poor, the algorithm is unstable. For the improvement 1 algorithm and improvement 2 algorithm, due to selecting the ideal initial cluster centers by calculating, so they have better clustering effect, the objective function value is smaller. And the K-means algorithm based on genetic algorithm proposed in the article find out the optimal objective function value through searching initial cluster centers, in the three groups of data, its objective function values are smaller than the other two algorithms', indicating that algorithm proposed in this paper has better clustering effect. During 3 experiments, this algorithm has already found the optimal objective function value, indicating that the algorithm is relatively stable. Can be seen from the from the table ,when the data have apparent isolated points , the algorithm proposed in this paper can significantly reduce the impact of outliers and improve clustering accuracy than the other two improved methods, and when the data have no apparent isolated points, the proposed algorithm still have more accurate clustering results than the other two improved methods, it is proved in the experiment that the method selecting initial cluster centers proposed in the text is not affected when the data have apparent isolated points, while the other the other two improved methods have certain limitations.

Ⅴ. CONCLUSION

The traditional K-means algorithm has many deficiencies., K-means Clustering Based on Genetic Algorithm proposed in this paper effectively overcome the shortcoming that K-means algorithm is sensitive to the initial cluster centers and reduce the impact of isolated points, it fully use the local search ability of K-means algorithm r global search ability of genetic algorithm, effectively avoid the phenomenon of clustering premature, so clustering has higher accuracy.

References

[1] Lian Fengna, Wu Jinlin, Tang Qi. An Improved Algorithm of K-means[J]. Computer and Information Technology, 2008,16 (1) :38-40.

[2] Jim Z.C.Lai, Tsung-JenHuang, Yi-ChingLiaw. A fast k-means clustering algorithm using cluster center displacement[DB]. Pattern Recognition, 2009, 42: 2551-2556.

[3] Zhan Yan, Yang Fang, Wang XiZhao. Learning the Center Number of Clustering Algorithms Using Genetic Algorithms[J]. Computer Engineering and Applications , 2003.16:86-87.

[4] Michael Laszlo,Sumitra Mukherjee.A genetic algorithm that exchanges neighboring centers for k-means clustering[DB]. Pattern Recognition Letters , 2007, 28: 2359–2366.

[5] Bu Yuanyuan, Guan Zhongren. The Study on Clustering Algorithm Based on K-means[N] . Journal of Southwest University for Nationalities•Nature Science Edition, 2009,35 (1) :198-200.

[6] Li Yeli, Qin Zhen. An Improved Algorithm of K-means[N]. Journal of Beijing Institute of Graphic Communication, 2007,15 (2) :63-65.

[7] Jin Yuping, Li Jingmei. An Intelligent Exem-paper Generating Method Based on Genetic Algorithm[D]. M.Eng Dissertation. Harbin: Harbin Engineering University, 2008.

Table Ⅰ Experimental Results

Clustering Algorithm	Iris（no outliers）		Iris（with outliers）	
	Initial cluster centers	*Minimum objective function value*	*Initial cluster centers*	*Minimum objective function value*
K-means 1	(45,31,106)	146.352815	(92,75,110)	123.813801
K-means 2	(29,145,27)	146.337854	(131,112,119)	140.708681
K-means 3	(100,30,26)	129.445793	(56,28,17)	139.178225
Improvement 1	---	129.438591	---	139.846509
Improvement 2	(41,117,15)	129.445801	(152,154,151)	136.897583
The article 1	(67,68,80)	129.348297	(109,67,0)	123.813797
The article 2	(72,111,56)	129.348297	(97,139,5)	123.813797
The article 3	(40,84,110)	129.348297	(125,149,106)	123.813797

Table Ⅱ Experimental Results

Clustering algorithm	Wine（no outliers）		Wine（with outliers）	
	Initial cluster centers	*Minimum objective function value*	*Initial cluster centers*	*Minimum objective function value*
K-means 1	(154,38,115)	450.620022	(0,176,2)	399.554657
K-means 2	(85,76,10)	450.607597	(170,24,89)	399.112285
K-means 3	(46,148,47)	450.620022	(150,163,143)	399.096192
Improvement 一	---	450.607597	---	399.112285
Improvement 二	(59,121,158)	450.619995	(178,182,179)	438.604126
The article 1	(92,131,138)	450.607544	(157,64,9)	399.096191
The article 2	(134,80,142)	450.607544	(116,73,38)	399.096191
The article 3	(68,74,135)	450.607544	(135,141,27)	399.096191