

A Valid Clustering Algorithm for High-dimensional Large Data sets Based on Distributed Method

GUO Xian e¹, YAN Junmei¹

¹Mathmatic and Computer Science Institution, Datong,Shanxi

E-Mail: {scjguo@yahoo.com.cn, junmei689@163.com}

Abstract—Data sets are randomly divided into several subsets,then fuzzy clustering method for A high-dimensional datas based on genetic algorithm is proposed to cluster the subsets,by importing a fuzzy dissimilar matrix to express the dissimilar degree between any two datas,and initializing the high-dimensional samples to two-dimensional plane.Then iteratively optimize the coordinate value of two-dimensional plane using genetic algorithm,which makes the Euclidean distance between the two-dimensional plane approximate to the fuzzy dissimilar degree between samples gradually.At last cluster the two-dimensional datas using FCM algorithm,so avoid dependence of clustering validity on the space distribution of high-dimensional samples.Experimental results show the method has high quality result,and improves the clustering speed greatly.

Index Terms—fuzzy clustering; distributed method; genetic algorithm; fuzzy dissimilar matrix; large data sets; high dimension

I. INTRODUCTION

Clustering is to classify things in terms of some essential arributes,such that similarity between samples

Foundation item:national science and technology office high and new technology plan item(2005EJ000017); Hebei province science and technology research and development plan(02547015D);Hebei province common high school doctor stake foundatton,2002(B2002118).

Biography: Guo xian e(1964-),female,associate professor, research direction:soft computing,fuzzy reasoning, data ming; Yan junmei(1981-),female, research direction:fuzzy clustering,data mining.

from the same class is significant while similarity between samples from different classes is small.It is non-supervised pattern recognition problem.The typical clustering algorithm are c-means clustering algorithm and fuzzy c-means(FCM) algorithm[1].

Efficiency is very slow using the typical algorithm to cluster the ever-increasing size of data sets[2],for this,we present a clustering algorithm based on distributed method.

If large data sets are two dimensional,cluster the subsets directly using typical algorithms.

High dimensional datas refer to space distribution,clustering validation depend on the space distribution of the sample considerably[3].

For example,c-means clustering algorithm is suitable only for hyper-spherical feature space of the samples,but not for randomly distributed ones[4].

FCM algorithm is suitable for ellipsoid[5].In order to get over the dependence,we use a high-dimensional data fuzzy clustering algorithm to cluster each subset of high-dimensional large data sets.

The purpose is to transfer fuzzy dissimilar degree between high-dimensional samples to the Euclidean distance between two-dimensional samples,namely transfer the difference between high-dimensional samples to the difference of two-dimensional samples,and the high dimensional samples are mapped into two-dimensional plane.

At last use FCM algorithm to cluster the datas in two-dimensional plane.

II. HIGH-DIMENSIONAL LARGE DATA SETS FUZZY CLUSTERING ALGORITHM BASED ON DISTRIBUTED METHOD

Algorithm procession description as follows:

- (1) Data sets $A=\{a_1,a_2,\dots,a_n\}$ are given;
- (2) Divided data sets A into substes A_1,A_2,\dots,A_p ;
- (3) If (the dimensional number >2),then use high-dimensional data fuzzy clustering algorithm based on genetic algorithm to cluster the subsets.Namely the high-dimensional samples are randomly initialized to two-dimensional plane,and use genetic algorithm to iteratively optimize two-dimensional samples,and the high-dimensional samples are mapped into two-dimensional plane,at last cluster the two-dimensional samples we gained using FCM algorithm;Else directly cluster them using FCM algorithm.
- (4) We gained the number of clustering center are m_1,m_2,\dots,m_p ,after clustering p subsets respectively;
- (5) If $|m_1+m_2+\dots+m_p|\geq n_0$ (n_0 is the threshold value of problem scale) make the number of $(m_1+m_2+\dots+m_p)$ clustering center A,go to step (2);Else go to step (6);
- (6) Directly cluster $(m_1+m_2+\dots+m_p)$ clustering centers using FCM algorithm;
- (7) Class combination:after step (6) finished,if centers x_1 and x_2 belong to the same class,and after step (4) c_1 is a class and its clustering center is x_1 while c_2 is a class and its clustering center is x_2 ,then combine the classes x_1 and x_2 to a class.
- (8) If the data sets are high dimensional samples, then the clustering result of step (7) return to high dimensional samples;else algorithm is stopped at step (7).

III. TWO-DIMENSIONAL SAMPLES CLUSTERING ANALYSIS BASED ON DISERIBUTED METHOD

A. Experimental Analysis on Distributed Method

Experiment select a part of Iris datas and use the attributes:petal length and pental width.Iris datas have 3 classes.The datas we selected are subsets of each class.

TABLE I

Iris Datas Are Selected

The first	1.3 0.2; 1.7 0.4; 1.4 0.3; 1.4 0.2; 1.5 0.1; 1.5 0.2; 1.6
-----------	---

1.8;6.1 2.5;6.4 2;6.7 2.2;6.9 2.3;6.1 1.9}.

All the clustering centers constitute the set $D=\{4.0133$

subset	0.2; 1.2 0.2; 1.4 0.1; 1.1 0.1
The second subset	4.7 1.4; 4.5 1.5; 4.9 1.5; 4 1.3; 4.6 1.5; 4.5 1.3; 4.7 1.6; 3.3 1; 4.6 1.3; 3.9 1.4; 3.5 1; 4.2 1.5; 4 1; 3.6 1.3
The third subset	5.9 2.1; 5.6 1.8; 5.8 2.2; 6.6 2.1; 6.3 1.8; 5.8 1.8; 6.1 2.5; 6.4 2; 5.3 1.9; 5.5 2.1; 6.1 1.9; 5.3 2.3; 5.5 1.8; 6.7 2.2; 6.9 2.3

TABLE II

The Result of Clustering A

Clustering center	Datas belongs to the class
4.0133 1.2493	3.3 1;4 1.3;4.6 1.3
1.3551 1.0775	1.5 1;1.4 1;1.1 1
1.4302 2.3097	1.3 2;1.7 4;1.4 3;1.4 2;1.5 2;1.6 2;1.2 2

TABLE III

The Result of Clustering B

Clustering center	Datas belongs to the class
4.5828 1.4506	4.7 1.4;4.5 1.5;4.6 1.5;4.5 1.3;4.7 1.6
4.8857 1.5519	4.9 1.5
5.5838 1.9526	5.9 2.1;5.6 1.8;5.8 1.8;5.3 1.9;5.5 2.1;5.3 2.3;5.5 1.8

TABLE IV

The Result of Clustering C

Clustering center	Datas belongs to the class
4.0614 1.3953	3.9 1.4;4.2 1.5;4 1
3.6125 1.1291	3.5 1;3.6 1.3
6.3785 2.1188	5.8 2.2;6.6 2.1;6.3 1.8;6.1 2.5;6.4 2;6.7 2.2;6.9 2.3;6.1 1.9

Data sets are randomly divided into 2 or 3 subsets,we set the number of clustering center 3,and the last once clustering set the number of clustering center 3.following is an example of the experiment.

The datas are randomly divided into 2 or 3 subsets.
 $A=\{1.3 2;1.7 4;1.4 3;1.4 2;1.5 1;1.5 2;1.6 2;1.2 2;1.4 1;1.1 1;3.3 1;4.6 1.3;4 1.3\}$;
 $B=\{4.7 1.4;4.5 1.5;4.9 1.5;4.6 1.5;4.5 1.3;4.7 1.6;5.9 2.1;5.6 1.8;5.8 1.8;5.3 1.9;5.5 2.1;5.3 2.3;5.5 1.8\}$;
 $C=\{3.9 1.4;3.5 1;4.2 1.5;4 1;3.6 1.3;5.8 2.2;6.6 2.1;6.3$ the the procedures.

1.2493; 1.3551 1.0775; 1.4302 2.3097; 4.5828 1.4506;

4.8857 1.5519; 5.5838 1.9526; 4.0614 1.3953; 3.6125 1.1291; 6.3785 2.1188}.

The number of D is 9,so needn't recursion.Directly cluster the 9 clustering centers using FCM algorithm.

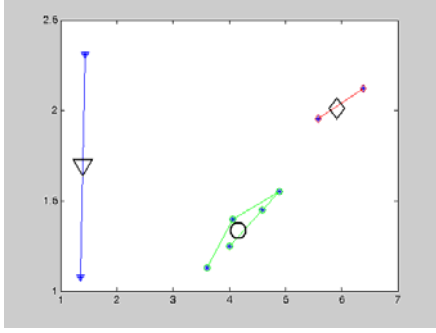


Figure1. Result of clustering D

At last class combination:the result of clustering set D:(1.3551 1.0775) and (1.4302 2.3097) belong to the same class; (5.5838 1.9526) and (6.3785 2.1188) belong to the same class;the rest 5 clustering centers are one class.So the class where clustering center is (1.3551 1.0775) and the class where clustering center is (1.4302 2.3097) should be combined to one class.

And these two classes are just the classes after set A is clustered,uniting these two classes we just get the datas of the first subset in the table 1;In the same way respectively get the datas of the second subset and the third subset in table 1.So using distributed method we get the same result with typical algorithm.

Using this method we do experiments 20 times continually,only once produce the disbut that (4.9 1.5) is divided into the third class,but in fact it is on the spatial border of the second class.So the points on the spatial border are easy to produce disbut.

Hereinbefore is the processing of using distributed method to two-dimensional samples.We proved its validation by experiments.

B.Analysis of TheTime Complexity

We do a comparison to FCM algorithm.FCM needs do the procedures as follows:

$$(1) \omega_i = \frac{\sum_{j=1}^n (\mu_{ij})^m x_j}{\sum_{j=1}^n (\mu_{ij})^m}, i=1,2,\dots,c$$

$$(2) \mu_{ij} = \left(\frac{1}{\|x_j - \omega_i\|_G^2} \right)^{1/(m-1)} / \sum_{k=1}^c \left(\frac{1}{\|x_j - \omega_k\|_G^2} \right)^{1/(m-1)},$$

$$i=1,2,\dots,c;j=1,2,\dots,n$$

If there are n datas to be clustered ,and c classes.In (2) every data need to be computed c times,so in all need to be computed n×c times.in (1) need to compute c times.If need to iterate L times,in all clustering all datas we need time

$$T(n)= o(L \times c \times n + L \times c).$$

In this paper we use distributed method,n datas are divided into s subsets,and every subset have n/m datas.Then $t(n)= o(n^{\log_s m})$.If n datas are divided into s subsets averagely,then every subset have n/s datas and s=m.So

$$t(n)= o(n).$$

$$\text{We know } o(n) \ll o(L \times c \times n + L \times c)$$

And by analyzing the unsupervised distributed clustering algorithm proposed in[6] we know its time complexity also more than $o(n)$.

So using distributed method this paper proposed can improve the clustering speed greatly.

So using distributed method to high-dimensional samples we only need to prove the validation of reducing high-dimensional samples to two-dimensional samples.Namely prove the validation of high-dimensional datas fuzzy clustering algorithm based on genetic algorithm.

IV. HIGH DIMENSIONAL DATAS FUZZY CLUSTERING ALGORITHM BASED ON GENETIC ALGORITHM

A. Fuzzy Dissimilar Matrix

The fuzzy dissimilar matrix stores the dissimilar measurement among the high-dimensional samples,the samples must be normalized in the range of [0,1].Assume that sample space is $X=\{x_1,x_2,\dots,x_n\}$, $\forall x_i \in X$,the feature vector is $x_i=(x_{i1},x_{i2},\dots,x_{ip})$,where x_{ik} denote the k-th attribute of the i-th sample.

The average μ and the mean square deviation of the k-th attribute for n samples are ,respectively

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad (1)$$

$$s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \mu_k)^2} \quad (2)$$

The initial samples are normalized as follows:

$$x'_{ik} = \frac{x_{ik} - \mu_k}{s_k} \quad (3)$$

Using Hamming distance method we get the dissimilarity between i-th and j-th samples

$$r_{ij} = c \sum_{k=1}^p |x'_{ik} - x'_{jk}| \quad (4)$$

The c in the formula(4) is the factor we select to make the r_{ij} in the range of [0,1].Here we set $c=0.01$.Then we get a fuzzy dissimilarity matrix $(r_{ij})_{nn}$ is a $n \times n$ symmetrical matrix with diagonal elements 0. $(r_{ij})_{nn}$:

$$\begin{bmatrix} 0 & & & & & \\ r_{21} & 0 & & & & \\ r_{31} & r_{32} & 0 & & & \\ \dots & \dots & \dots & \dots & \dots & \\ r_{n1} & r_{n2} & \dots & \dots & 0 & \end{bmatrix} \quad (5)$$

B. The Selection of Fitness Function

Firstly the high-dimensional samples are initialized into two-dimensional plane,then use genetic algorithm to iteratively optimize the coordinate value of two-dimensional samples,and make the Euclidean distance among two-dimensional samples approximate to the dissimilarity measurement among high dimensional samples.So,the error function of genetic algorithm is defined as

$$E = \sum_{i=1}^n \sum_{j=i}^n |r'_{ij} - r_{ij}| \quad (6)$$

Where, r'_{ij} is the Euclidean distance between the samples x_i and x_j ;Whose coordinates are (a_i, b_i) , $i=(1,2,\dots,n)$ and (a_j, b_j) , $j=(1,2,\dots,n)$ respectively,and r'_{ij} is defined as

$$r'_{ij} = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2} \quad (7)$$

The smaller the value of the error function is,the greater the fitness of the individual is,and thus the fitness function is defined as

$$f = 1/(1+E) \quad (8)$$

C. High-Dimensional Datas Fuzzy Clustering Algorithm Description Based on Genetic Algorithm

- 1) Initialization.Distribute the samples in a plane randomly,assign randomly coordinate pairs (a_i, b_i) to each sample where $a_i, b_i \in [0,1]$, $i=1,2,\dots,n$.
- 2) Construct the fuzzy dissimilarity matrix $(r_{ij})_{nn}$ using Eqs (1)~(5).
- 3) Form the initial population.Each pair of the coordinates (a_i, b_i) is viewed as a gene and coded to an 8-bit binary.The number of the genes is n..Then the coordinates of all samples are linked into a chromosome(also named as an individual).The length L of a chromosome is 8n bits.According to the different order.N individuals are created to form an initial population S.
- 4) Compute fitness of each individual in S: Compute the fitness of each individual in S using eqs (6)~(8).
- 5) Select parental individuals.The roulette wheel selection and the elitist strategy are adopted.Firstly,the individual with the greatest fitness is chosen as a parental one.Calculate the probability of each residual individual

$$p_k = f_k / \sum_{m=1}^N f_m \text{ and the accumulative probability}$$

$$q_i = \sum_{j=1}^i p_j \text{ .Generate a random even number } r \text{ in the}$$

interval of [0,1).if $r < q_1$,then select the first individual.Otherwise,if k satisfies the condition of $q_{k-1} \leq r < q_k$,then select the individual k.Whirl the roulette wheel for M-1 times;thus M individuals mighe be chosen

- to form a sub-population \mathcal{S}' , $\mathcal{S}' \subset \mathcal{S}$.
- 6) Mate the individual randomly in \mathcal{S}' .
 - 7) Crossover operation. Create a random number r in the interval of $[0,1]$ for each pair of individuals in \mathcal{S}' . If $r < P_c$, where P_c is a given crossover probability, then carry out the crossover operation. Next, generate a random number in the interval of $[1,8n]$ in order to determine the location of crossover, and thus form a sub-population \mathcal{S}'' consisting of new individual.
 - 8) Mutation operation. Create a random number r in the interval of $[0,1]$ for each position of each individual in \mathcal{S}'' . If $r < P_m$, where P_m is a given mutation probability, then carry out mutation operation in this position.
 - 9) Calculate the fitness of all individuals in $\mathcal{S} + \mathcal{S}''$, and eliminate M individuals with smaller fitness to form a new population \mathcal{S} .
 - 10) Temination. If the biggest fitness in the new generation subtract the biggest fitness in the last generation less than ε ($\varepsilon = 0.0005$), then decode; else return 5).
 - 11) use FCM algorithm to the two-dimensional samples, and revert the result to the high-dimensional samples.

D. The Feasible Analysis of Algorithm

The algorithm make the Euclidean distance among two-dimensional samples approximate to the fuzzy dissimilarity among high-dimensional samples using genetic algorithm, and make the error function

$$E = \sum_{i=1}^n \sum_{j=i}^n |r'_{ij} - r_{ij}| \text{ up to minute.}$$

In term of 3.1 and 3.2 we know, $r_{ij} \in [0,1]$,

$$r'_{ij} \in [0,1].$$

If $r'_{ij} \approx 0$, namely the dissimilarity between high-dimensional samples i and j is almost 0, which means classes, where each class include 7 samples.

In the simulations the population size N , the iterative times, the mutation probability P_m and the crossover probability P_c are taken as 100, 60, 0.5, 0.2. The clustering result is shown in fig 2:

sample i and j belong to the same class. And r'_{ij} approximates to r_{ij} , so $r'_{ij} \approx 0$, namely the Euclidean distance between two-dimensional samples are mapped to by the high-dimensional samples is almost 0. Distance between samples from the same class is small while distance between different classes is significant. According to this, we know after FCM this two-dimensional samples belong to the same class.

If $r'_{ij} \approx 1$, namely the dissimilarity between high-dimensional samples i and j is almost 1, which means samples i and j belong to different classes. And r'_{ij} approximate to r_{ij} , so $r'_{ij} \approx 1$, so after FCM algorithm this two-dimensional samples belong to different classes.

In the same way, any two high-dimensional samples, the more dissimilarity is between them, the more the Euclidean distance is between two-dimensional samples that are mapped to by the high-dimensional samples. And the more the Euclidean distance is, the more the dissimilarity is between the two-dimensional samples. So the difference measurement between high-dimensional samples return to the difference measurement between two-dimensional samples. So cluster the two-dimensional samples means to cluster the initial high-dimensional samples, which proves its feasibility.

E. Numerical Simulation

In the experiments we select a part of Iris datas. The test data set consist of 21 instances, 4 attributes and 3 following is the references.

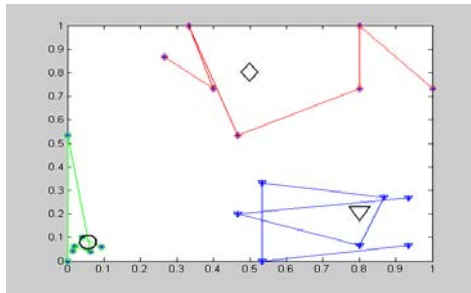


Fig2 result of clustering based on genetic algorithm

Experimental results show this method has better clustering effect, namely we prove the validation of high-dimensional samples are reduced to two-dimensional to cluster.

V. CONCLUSION

In this paper we present a high-dimensional large data sets fuzzy clustering algorithm based on distributed method. Large data sets are randomly divided into several subsets, clustering each subsets we use a method based on genetic algorithm, at last unite the classes.

Firstly in 3.1 we do the experiments using two-dimensional samples, the experimental results show that most of time the distributed method has the same result with once clustering, but the points on the spatial border are easy to produce disbut.

In 3.2 by analyzing time complexity we know this method can improve clustering speed greatly. To the high-dimensional samples in 4.5 by experiments we prove the validation of high-dimensional samples are reduced to two-dimensional samples to cluster, and in 4.4 analyze its feasibility.

So the method we proposed is proper to cluster high-dimensional large data sets.

It can improve clustering efficiency, and has validation.

REFERENCES

- [1] Yuanquan Zhang, Luis Rueda. "A geometric framework to visualize fuzzy-clustered data". IEEE. Proceedings of the XXV International Conference of the Chilean Computer Science Society, 2005, 5(8):8-13.
- [2] Davidson, Ashwin Satyanarayana. "Speeding up k-means clustering by bootstrap averaging. the workshop on

clustering large data sets." IEEE International Conference on Data Mining, 2004, 5(3):98-102.

- [3] R.J.Hathaway, J.C.Bezdek. fuzzy c-means clustering of incomplete data. IEEE Trans. syst. 2001, 31(6):735-744.
- [4] Aggarwal C, Yu P. "Finding generalized projected clusters in dimensional spaces". In ACM SIGMOD Conference, 2000, 5(4):78-52.
- [5] Cheng-Fa Tsai, Chun-Wei Tsai, Chi-ping chen. "A novel multiple searching genetic algorithm for multimedia multicast routing". IEEE Congress on Evolutionary Computation. 2002, 9(10):7065-7068.
- [6] D.K.Tasoulis, M.N.Vrahatis. Unsupervised Distributed Clustering. In Knowledge Discovery and Data Mining, pages 91-95, 2002.