A Novel Quantum Genetic Clustering Algorithm for Data Segmentation

Ming-an Zhang
Institute of Software, Chinese Academy of Sciences
4# South Fourth Street, Zhongguancun, Beijing 100190, PR.China
86 010 62661154
171166789@qq.com

Yong Deng*
Institute of Software, Chinese Academy of Sciences
4# South Fourth Street, Zhongguancun, Beijing 100190, PR.China
86 010 62661154
dengyong@iscas.ac.cn

Dong-xia Chang
Institute of Information Science, Beijing jiaotong University
No.3 Shang Yuan Cun,Hai Dian District Beijing 100040, PR.China
86 010 51684108
dxchang@bjtu.edu.cn

ABSTRACT
Based on the concept and principles of quantum computing, a novel genetic clustering algorithm is proposed, which can automatically clustering a data set into clusters, and evolve the optimal number of clusters as well as the cluster centers of a data set. A Q-gate with adaptive selection of the angle for every niche is introduced as a variation operator to drive individuals toward better solutions. Experiments show that the algorithm proposed is better than simple clustering algorithms.

Categories and Subject Descriptors
I.4.6 Segmentation: pixel classification

General Terms
Algorithms

Keywords
quantum genetic algorithm; quantum rotation; data segmentation

1. INTRODUCTION
Clustering algorithms have been extensively used to data segmentation[1]. But the traditional ones need to obtain the cluster number and the initial cluster centers in advance. Stochastic clustering algorithms based on simple genetic algorithm have been proposed to overcome these problems[2]. Some concepts[3] of quantum computing are adopted in the proposed algorithm (QGCA). A simpler representation with real-coded is adopted, whereby each individual represents a single cluster center.

2. QGCA DESCRIPTION
2.1 Quantum Concepts
In quantum computation, the smallest unit of information is called a quantum bit, or a qubit. A qubit is generally expressed as

$$|\varphi\rangle = \alpha\,|\,0\rangle + \beta\,|\,1\rangle \tag{1}$$

Where $|\alpha|^2+|\beta|^2=1$, and complex α and β, known as the probability amplitude of the states, 0 and 1, respectively. The binary quantum coding method is commonly used, and a qubit can be represented by a plurality $[\alpha,\beta]^T$. The chromosome can be as follows:

$$v=\begin{bmatrix}\alpha_{11} & \cdots & \alpha_{1k} & \alpha_{21} & \cdots & \alpha_{2k} & \alpha_{m1} & \cdots & \alpha_{mk}\\ \beta_{11} & \cdots & \beta_{1k} & \beta_{21} & \cdots & \beta_{2k} & \beta_{m1} & \cdots & \beta_{mk}\end{bmatrix} \tag{2}$$

Here, m: the number of genes, k: the number of quantum bit in genes, and $|\alpha_{pq}|^2+|\beta_{pq}|^2=1$,The traditional genetic algorithm keeps the population diversity with crossover, mutation operators, while quantum genetic algorithm with quantum probability amplitude method by quantum gates. The single qubit gate3 is used:

$$U(\theta)=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix} \tag{3}$$

Here, θ is the rotation angle of quantum rotation gate. There are many schemes of quantum rotation gate adjustment strategy4. Its core idea is: the current solution convergence to a higher fitness individual. If the *i*-th bit of the chromosome is 0, make $|\alpha|^2$ bigger. If *i*-th of the chromosome is 1, make $|\beta|^2$ bigger.

2.2 Algorithm Flow
1. Input the max generation, the probability of crossover and mutation. 2. Population initialization. 3. For each chromosome in the population, decoding the chromosome into real-valued; compute the fitness. 4. Dynamic niching[4]. 5. Quantum update by the quantum rotation gate. 6.If the generation is satisfied, output the niching masters, or continue the evolution process.

2.3 Chromosome Representation
The quantum chromosome can be represented as

$$|\varphi\rangle = \alpha\,|\,x^l\rangle + \beta\,|\,x^u\rangle \tag{4}$$

where the feature vector x^l and x^u are the lower and the upper bound of x, respectively. Obviously, a qubit may be in the x^l state, in the x^u state, or in any superposition of the two. Then the chromosome can be represented by qubit as follows:

$$q_k^t=\left[\begin{array}{cc|cc|c|cc}\alpha_{k,11}^t & \alpha_{k,12}^t & \alpha_{k,21}^t & \alpha_{k,22}^t & \cdots\ \cdots & \alpha_{k,N1}^t & \alpha_{k,N2}^t\\ \beta_{k,11}^t & \beta_{k,12}^t & \beta_{k,21}^t & \beta_{k,22}^t & \cdots\ \cdots & \beta_{k,N1}^t & \beta_{k,N2}^t\end{array}\right] \tag{5}$$

The length of the quantum chromosome is 2N. We obtain a real-valued encoded mode. For any qubit $\left[\alpha_{i,k}^t,\beta_{i,k}^t\right]^T$, k=1,2, we generate a random number $r_{i,k}\in[0,1]$. If $r_{i,k}<\left|\alpha_{i,k}^t\right|^2$, the qubit will be in x^u state. Therefore, the qubit chromosome collapses into $[x^i,x^j]$, where $\{i,j\}\in\{l,u\}$. Any each dimension of the qubit chromosome will be in the 4 states: [x^l, x^l], [x^l, x^u], [x^u, x^l], [x^u, x^u]. To decode

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.

the chromosome into real value, a decoding criterion is introduced as Tab.1. Here $\Delta x_i=(x^u- x^l)/4$ and r is a random number between 0 and 1. If the chromosome is [x^l, x^l], the x_i will take a small value inclining to the lower bound. If the chromosome is [x^u, x^u], the x_i will take a small value inclining to the upper bound. So the chromosome can be transformed into a real-valued one.

Table 1. Encoded rules of the chromosome

Qubit	Encoded mode
[x^l, x^l]	$x_i = x^l + r\Delta x_i / 4$
[x^l, x^u]	$x_i = x^l + (1+r)\Delta x_i / 4$
[x^u, x^l]	$x_i = x^u - (1+r)\Delta x_i / 4$
[x^u, x^u]	$x_i = x^u - r\Delta x_i / 4$

The population size is P, and the initial population is generated randomly. $\alpha^0_{i,j}$ and $\beta^0_{i,j}$ are initialized to 1/2, i=1,…, P, j=1,2.

2.4 Fitness Function

The fitness function is defined as

$$f(\mathrm{c}) = \tilde{J}_s(\mathrm{c}) = \sum_{j=1}^{n}\left(\exp - \frac{\|\mathrm{x}_j\text{-}\mathrm{c}\|^2}{\beta}\right)^{\gamma}, j = 1, \cdots, n \tag{6}$$

where, , j=1,…,n, are all points in the data set to be clustered.

2.5 Adaptive Quantum Rotation Strategy

We adopt a dynamic niching method[4] to divide the individuals in evolution into some small populations.

$$Pop = \bigcup_{i\in\{1,2,\cdots v(t)\}} S_t^i \cup S_t^* \tag{7}$$

After the population is divided into multiple sub ones, in order to make the qubit chromosomes effectively converge to the better states, we put forward an adaptive rotation angles computing method. For the real individuals, the rotation angle is as

$$\theta_{k,i}^{t,q} = sign\{\alpha_{k,i}^{t,q} \cdot \beta_{k,i}^{t,q} \cdot (f(c_l^t) - \bar{f}^q)\} \cdot \frac{f(M_i^{t,q}) - f(c_i^t)}{f(M_i^{t,q} - \bar{f}^q)} \times 0.05\pi \tag{8}$$

Where, $\mathrm{M}^{t,q}$ is the master of the q-th niche, c_l^t is the individual in the q-th niche, $f(\mathrm{M}^{t,q})$, $f(\mathrm{c}_l^t)$ and $\bar{f}^q$ are the fitness of $\mathrm{M}^{t,q}$, c_l^t and the average fitness of q-th sub population, sign(.) seen in Ref.6. When the performance is poor, the evolution of individual choice angle will increase. When good, we will choose small $\theta_{k,i}^{t,q}$ so as to avoid the better individual damaged. For the individuals in the isolated set S_t^*, which is not belong to any real one, the rotation angle is defined as:

$$\theta_{k,i} = \begin{cases} 0 & if \quad f(c_l) = f(c_{best}^t) \\ 0.02\pi & \quad others \end{cases} \tag{9}$$

3. SIMULATION EXPERIMENTS

We use the UCI machine learning repository Iris, Breast Cancer and Wine data sets on GAC[1], KGAC[2], GAGR[5] and QGNC in the experiments. GAC, KGAC, GAGR algorithm can't automatically determine the number of clusters. We should set several different clusters. Then determine the appropriate clusters according to the clustering performance under the condition of different values. The initial populations are set to 50. The crossover probability and mutation probability of GAC, KGAC, GAGR are 0.8 and 0.001. QGNC can divide the population into different niche, and the number of niche is same as the one of clusters. Therefore, we first validate QGNC's performance of automatically determining clusters. We made experiments for each data independently 100 times. Results are in Tab.2.

Table 2. Mean, variance and correct estimate

Data set	Iris	Breast	wine
Actual	3	2	3
Obtained	2.86(0.4494),72	2(0.0408),98	3.0769(0.4130), 65

Tab.2 shows QGNC algorithm can effectively determine the number of clusters. We will further analyze the algorithm efficiency of QGNC, GAC, KGAC and GAGR on Iris data set. We will run the algorithms independently 20 times. Tab.3 gives the running times (MATLAB R2009a, 2.33GHz Xeon (R) CPU). Tab.3 shows QGNC algorithm search the optimal solutions running better time than the other three algorithms. It spent shorter time to obtain the optimal solution.

Table 3.The average running times to four algorithms

Data set	GAC	KGAC	GAGR	QGNC
Iris	45.295	0.4494	0.3741	0.2040

4. CONCLUSIONS

In the paper, each chromosome is encoded a center of a cluster by a real-valued qubit. The dynamic niching is accomplished without assuming any a priori knowledge on the number of niches. An adaptive selection of the rotation angle used by the quantum rotation gate is introduced. The experiment results have shown that our algorithm is effective for data segmentation.

5. ACKNOWLEDGMENTS

This work was supported by Beijing Municipal Science Foundation No. 4133092 and 863 program (No.2012AA011206).

6. REFERENCES

[1] Murthy C A, Chowdhury N, In search of optimal clusters using genetic algorithms. Pattern Recognition Letters, 1996, 17: p 825-832.

[2] Bandyopdhyay S, Maulik U, An evolutionary technique based on K-Means algorithm for optimal clustering in RN. Information Sciences, 2002, 146: 221-237.

[3] Sha Lin-Xiu, He Yu-Yao. A novel self-adaptive quantum genetic algorithm[J]. International Conference on Natural Computation, 2012, p 618-621.

[4] B. L. Miller and M. Shaw. Genetic algorithms with dynamic niche sharing for multimodal function optimization, Proc. 1996 IEEE Trans. Evol. Comput., 786-791.

[5] Chang D X, Zhang X D, Zheng C W, A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recogn, 2009, 42: p 1210-1222.