基于改进量子旋转门的量子进化数据聚类

刘 芳[1,2],王 爽[2],柳莹莹[1,2],戚玉涛[1,2]
(1.西安电子科技大学计算机学院,陕西西安 710071; 2.智能感知与图像理解教育部重点实验室,陕西西安 710071)

摘 要: 在量子进化计算中,量子旋转门是种群进化的主要算子,但是该算子旋转角度的选取是离散且固定的,使问题的搜索容易陷入局部最优.因此,本文提出了一种改进的量子旋转门算子,它能够自适应地计算旋转角度,使种群能够具有比较好的全局搜索能力;同时为了避免陷入局部最优,本文对旋转后的概率幅进行了修正操作.针对数据聚类问题,本文提出了一种基于改进量子旋转门的量子进化数据聚类方法.仿真对比实验表明:与采用常规的量子旋转门的算法及一些其他的进化算法相比,本文方法在聚类正确率上有了很大的改善;同时,针对具有对称分布的数据集,在统一采用对称距离测度后,本文的方法也取得了较好的效果.

关键词: 量子进化计算; 数据聚类; 量子旋转门
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2011) 09-2008-06

A Quantum-Inspired Evolutionary Algorithm Based on a Modified Quantum Rotate Gate for Data Clustering

LIU Fang[1,2], WANG Shuang[2], LIU Ying-ying[1,2], QI Yu-tao[1,2]
(1. *School of Computer Science & Technology, Xidian University, Xi'an, Shaanxi* 710071, *China*;
2. *Key Lab of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an, Shaanxi* 710071, *China*)

Abstract: In traditional quantum-inspired evolutionary algorithm (QEA), a quantum rotate gate is the main operator in a quantum population evolution. However, the choice of rotate angle is also discrete and constant, which makes the search of the problem easy to fall into local optimum. Therefore, a modified quantum rotate gate is proposed in this paper. The new gate uses adaptive method of calculation of rotation, which makes the population have a relatively good global search capability. At the same time, the probability amplitude is modified after rotation to enable population to jump out of local optimum. For data clustering problem, a quantum-inspired evolutionary algorithm based on the modified quantum rotate gate is proposed. The simulation experiment results show, compared with the algorithm based on a normal quantum rotate gate and some of other evolutional algorithms, the proposed algorithm increases correct rate of data clustering. At the same time, the simulation experiment results on the data sets of symmetrical distribution property show, compared with the algorithms adopting a symmetrical distance measure, our algorithm also achieves better results.

Key words: quantum-inspired evolutionary algorithm; data clustering; quantum rotate gat

1 引言

聚类[1~3]是一个重要的无监督分类技术,已经被广泛应用于计算机视觉、信息检索、数据挖掘和模式识别等领域.在现有的聚类方法中,基于目标函数的聚类算法由于把聚类问题归结为一个优化问题,是聚类算法研究的重要分支之一,K-均值算法就是其中最典型的方法[4,5].由于 K-均值算法用梯度下降法优化目标函数,因此对初始值敏感,容易陷入局部最优[5].作为一类有效的全局优化技术,进化算法已经被很多学者用于聚类问题[6,7].在设计进化聚类算法时,最核心的两个问题是进化个体的编码及相似度度量.针对聚类问题的个体编码方式很多,使用较多的是对 K-均值算法的聚类中心进行编码,然后对数据样本按照其与聚类中心的相似

收稿日期:2010-05-28;修回日期:2010-11-29
基金项目:国家教育部博士点基金(No.200807010003,No.20090203120016);国家863高技术研究发展计划(No.2008AA01Z125,No.2009AA12Z210);陕西省"13115"科技创新工程重大科技专项(No.2008ZDKG-37);国家自然科学基金(No.60703107,No.60703108,No.60803098,No.60803706,No.60872135);中国博士后科学基金特别资助(No.200801426);中国博士后科学基金资助(No.20080431228,No.20090461283);中央高校基本科研业务费专项资金资助(No.JY10000903007,No.JY10000902040)

度进行类别划分.传统的进化算法虽然是一种全局搜索算法,但是仍然存在着因多样性损失而容易陷入早熟收敛的问题.为了使进化算法保持较好的种群多样性,提高求解质量,本文引入量子进化算法[8,9]求解数据聚类问题.

量子进化算法将量子计算的并行性引入进化计算之中,使用量子比特编码染色体,这种概率表示能够使一个量子染色体同时表征多个状态的信息,带来了丰富的种群,而且当前最优个体的信息能够很容易的用来引导变异,使得种群以大概率向着优良模式进化,加快收敛速度[10].因此,本文将量子进化算法用于聚类,能够以比较小的种群得到比较大的搜索空间,从而更快更好地收敛到正确的聚类结果.

2　聚类问题描述

聚类(Clustering)是一种常见的数据分析工具[1~3],聚类问题的数学描述如下:

被研究的样本集为 E,类 C 定义为 E 的一个非空子集,即 $C\subset E$ 且 $C\neq\varnothing$,聚类就是满足下列两个条件的类 $C_1,C_2,\cdots,C_k$ 的集合:

$$\begin{cases}\bigcup_{i=1}^{k} C_i = E\\ C_i\cap C_j=\varnothing,\ i=1,\cdots,k;j=1,\cdots,k;i\neq j\end{cases}\tag{1}$$

基于目标函数的聚类方法把聚类问题归结为一个带约束的非线性规划问题,通过优化求解获得数据集的划分和聚类.目前进化计算已被广泛应用于聚类技术目标函数的优化.U.Maulik 和 S.Band-yopadhyay 提出了基于遗传算法的聚类技术[6],Hall,Ozyurt 和 Bezdek 提出了用遗传的优化方法求解聚类问题[7]等.K-均值算法以 K 为参数,把 n 个对象分成 K 个簇,使簇内具有较高的相似度,簇间的相似度比较低,通常采用如下目标函数:

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n} d^2(x_j,p_i)\tag{2}$$

其中,p_i 为第 i 个聚类中心,x_j 为第 j 个数据,$d(x_j,p_i)$ 为两点之间的距离,一般取欧氏距离,目标函数即为所有点到所有对应聚类中心的距离和,该距离和越小,说明聚类的结果越好.

3　量子进化算法

进化算法(Evolutionary Algorithm,EA)是一类模拟生物进化过程和机制求解问题的自组织、自适应人工智能技术[11~13].分析 EA 可以发现:EA 容易因种群多样性损失而陷入早熟收敛[12~14].

量子力学是 20 世纪物理学最惊心动魄的发现之一,以量子力学原理为基础的量子信息学为信息科学的发展提供了新的研究思路[15,16].量子进化算法(QEA)[8~10]是将量子计算的机理和特性引入到进化算法中,充分利用量子计算的并行性和随机性,在保持良好的种群多样性的同时提高搜索效率.

3.1　量子比特

在 QEA 中,最小的信息单元为一个量子位——量子比特[17~19].一个量子比特的状态可取 0(记为$|0\rangle$)或 1(记为$|1\rangle$),或者处于两者之间的中间态,即其状态$|0\rangle$和$|1\rangle$的不同叠加态,所以一个量子位记为$|\psi\rangle$,可以表示为:

$$|\psi\rangle=\alpha|0\rangle+\beta|1\rangle\tag{3}$$

且满足归一化条件:

$$|\alpha|^2+|\beta|^2=1\tag{4}$$

其中,α 和 β 分别是$|0\rangle$和$|1\rangle$出现的概率幅,$|\alpha|^2$,$|\beta|^2$ 分别表示量子位处于状态 0 和状态 1 的概率.

3.2　量子位的相位

用一对满足式(4)在[-1,1]之间的实数(α,β)来描述一个量子位的概率幅,那么该量子位的相位 ω 表示如下:

$$\omega=\arctan(\beta/\alpha)\tag{5}$$

用符号 d 来表示 α 和 β 的乘积,即

$$d=\alpha\times\beta\tag{6}$$

其中,d 的正负值代表此量子位的相位 ω 在平面坐标中所处的象限,如果 d 的值为正,则表示 ω 处于第一、三象限,否则处于第二、四象限.

3.3　量子染色体编码

EA 的常用编码方式有二进制编码、十进制编码和符号编码.在 QEA 中,使用一种基于量子比特的编码方式[17,18],即用一对实数来定义一个量子比特位.一个具有 m 个量子比特位的系统可以描述为:

$$\begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_m\\ \beta_1 & \beta_2 & \cdots & \beta_m\end{bmatrix}\tag{7}$$

其中,$|\alpha_i|^2+|\beta_i|^2=1(i=1,2,\cdots,m)$.这种表示方法可以表征任意的线性叠加态.

例如,一个具有如下概率幅的 3 量子比特系统:

$$\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{2}\\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{\sqrt{3}}{2}\end{bmatrix}\tag{8}$$

则系统的状态可以表示为:

$$\begin{aligned}&\frac{1}{4}\left|000\right\rangle+\frac{\sqrt{3}}{4}\left|001\right\rangle-\frac{1}{4}\left|010\right\rangle-\frac{\sqrt{3}}{4}\left|011\right\rangle\\&+\frac{1}{4}\left|100\right\rangle+\frac{\sqrt{3}}{4}\left|101\right\rangle-\frac{1}{4}\left|110\right\rangle-\frac{\sqrt{3}}{4}\left|111\right\rangle\end{aligned}\tag{9}$$

4　基于量子进化计算的数据聚类算法

本文将量子进化算法用于数据聚类问题,提出了

基于改进量子旋转门的量子进化数据聚类算法(IQEAC).量子旋转门是量子进化算法的核心算子,在现有的量子旋转门操作中,量子相位的旋转角度是固定的离散值.这种有限状态的更新方式制约了种群多样性的增加,使得算法容易因多样性不足而导致早熟收敛.为了防止算法陷入局部最优,本文提出了一种改进的量子旋转门算子,能够自适应地计算量子相位的旋转角度并对旋转后的概率幅进行修正,从而提高量子进化算法的多样性保持能力.

4.1 量子编码和亲和度定义

本文对聚类中心进行量子编码.每个量子比特(Qubit)由 α 和 β 来表示,并且满足 $|\alpha|^2+|\beta|^2=1$.经量子观测后,生成二进制编码的普通种群.假设量子种群大小为 pop,数据集类别数为 K,数据维数为 D,每一维数据用 b 位二进制来表示,则每一个量子染色体的长度 $L=K\times D\times b$.则种群 $\boldsymbol{Q}(t)=\{q_1^t,q_2^t,\cdots,q_{pop}^t\}$ 中第 i 个个体的编码形式为:

$$\boldsymbol{q}_i^t=\begin{bmatrix}\alpha_1^t & \alpha_2^t & \cdots & \alpha_L^t\\ \beta_1^t & \beta_2^t & \cdots & \beta_L^t\end{bmatrix},i=1,2,\cdots,pop \tag{10}$$

为了保证所有可能的线性叠加态以相同的概率出现,本文把种群中所有的 α 和 β 都初始化为 $1/\sqrt{2}$,用 $b=8$ 位二进制位来表示一个实数.

在数据聚类问题中,本文取适应度函数为:所有点到所有对应聚类中心的距离和的倒数,如式(11)所示,其中 J 在式(1)给出;量子进化的目的就是寻找适应度函数的最大值.

$$f=\frac{1}{J} \tag{11}$$

4.2 IQEAC 算法流程

IQEAC 算法具体步骤如下:

Step 1 令 $t=0$,初始化量子种群 $\boldsymbol{Q}(t)=\{q_1^t,q_2^t,\cdots,q_{pop}^t\}$;

Step 2 对量子种群 $\boldsymbol{Q}(t)$进行观测,得到普通种群 $\boldsymbol{P}(t)=\{p_1^t,p_2^t,\cdots,p_{pop}^t\}$.其中,每个个体 $p_i^t=\{a_1^t,a_2^t,\cdots,a_i^t,\cdots,a_L^t\}$,$a_i^t$ 为表示 0 或 1 的二进制位;

Step 3 对普通种群 $\boldsymbol{P}(t)$进行评价,计算种群中每个个体的适应度,得到数组 $\boldsymbol{Fit}=\{f_1,f_2,\cdots,f_{pop}\}$,并保留适应度最大的个体 $best$;

Step 4 对 $\boldsymbol{Q}(t)$采用改进的量子旋转门操作和修正操作进行更新,生成新的量子种群 $\boldsymbol{Q}(t)'$;

Step 5 重新观测量子种群 $\boldsymbol{Q}(t)'$,得到普通种群 $\boldsymbol{P}(t)'$;

Step 6 对 $\boldsymbol{P}(t)'$进行量子全干扰交叉得到 $\boldsymbol{P}(t)''$;

Step 7 把 $\boldsymbol{P}(t)''$中的每个个体解码后的聚类中心作为 K-均值的初始中心,用 K-均值算法迭代 3 次,更新 $\boldsymbol{P}(t)''$,得到新的种群 $\boldsymbol{P}(t)'''$;

Step 8 对种群 $\boldsymbol{P}(t)'''$进行评价,计算每个个体的适应度 Fit,并保留最佳个体 $best$;

Step 9 若连续 10 次相邻两次最佳个体的适应度值之差小于 1e-10,则退出,否则 $t=t+1$,转到Step 4.

4.3 改进的量子旋转门操作和修正操作

在量子进化算法中,使用量子旋转门操作对量子染色体基因位进行改变,如式(12)所示.$\boldsymbol{U}(\theta)$为量子旋转门,θ 为旋转角度,量子染色体基因位$[\alpha,\beta]^{\mathrm{T}}$经过量子旋转门操作后变异为$[\alpha',\beta']^{\mathrm{T}}$.

$$\begin{bmatrix}\alpha'\\ \beta'\end{bmatrix}=\boldsymbol{U}(\theta)\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} \tag{12}$$

旋转角度 θ 一般是从事先定义的旋转角度表(如文献[9,10,17~19])中取值的,该表中的角度值是离散的且固定的.这种查表选择旋转角度的方式对量子染色体基因的变异操作是有限固定状态的,当种群多样性不足时,这种有限状态的变异操作无法带来种群多样性的增加,从而使搜索容易陷入局部最优.本文提出了一种改进的量子旋转门操作和对旋转后的概率幅进行修正的操作.改进的量子旋转门操作能够根据当前量子基因位和染色体适应度的值自适应地计算合适的旋转角度,从而实现在种群多样性不足的情况下增加种群多样性,避免算法早熟收敛的目的.改进的量子旋转门操作和修正操作的具体实现方式如下:

假设 x_i 和 b_i 分别为运行量子观测操作后个体 x 和最佳个体 b 中的第 i 位上的值,Fit_x 和 Fit_b 分别为两个个体的适应度函数值,引入如下定义:

$$S_1=\mathrm{sgn}((x_i-0.5)\alpha_i\beta_i) \tag{13}$$

$$S_2=\mathrm{sgn}(f_x-f_b) \tag{14}$$

其中,$\mathrm{sgn}(y)=\begin{cases}1, & y\geqslant 0\\ -1, & y<0\end{cases}$为符号函数.

改进的量子旋转门操作中,旋转角度 θ_i 可以通过式(15)计算得到:

$$\theta_i=S_1S_2\theta_0\mathrm{e}^{-(f_x-f_b)} \tag{15}$$

其中 θ_0 为初始旋转角度,实验中取 θ_0 为 0.05π,S_1 和 S_2 用来控制旋转方向,Fit_x 和 Fit_b 用来自适应地调整旋转角度的大小.

如果量子比特的 α 或 β 的值过早的收敛到 0 或 1 附近,则对该量子基因位的观测结果只能为接近 0 或 1 的值,此时量子旋转门操作会进一步加快量子比特的收敛,导致多样性进一步损失,从而陷入局部最优.为了克服量子旋转门操作的这一缺陷,本文对旋转后的概率幅采用 H 门[8]进行修正.设定一个阈值 $0<e\ll 1$,当 $|\alpha|^2$ 和 $|\beta|^2$ 小于阈值 e 时,采用式(16)的操作对旋转后的量子基因位进行修正:

$$[\alpha_i^{t+1}\beta_i^{t+1}]^{\mathrm{T}}=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}[\alpha_i^{t}\beta_i^{t}]^{\mathrm{T}},$$

$$\begin{aligned}&①\text{if } |\alpha_i^{t+1}|^2\leqslant e \text{ and } |\beta_i^{t+1}|^2\geqslant 1-e,\\ &\quad [\alpha_i^{t+1}\beta_i^{t+1}]^{\mathrm{T}}=[\sqrt{e}\ \sqrt{1-e}]^{\mathrm{T}}\\ &②\text{if } |\beta_i^{t+1}|^2\leqslant e \text{ and } |\alpha_i^{t+1}|^2\geqslant 1-e,\\ &\quad [\alpha_i^{t+1}\beta_i^{t+1}]^{\mathrm{T}}=[\sqrt{1-e}\ \sqrt{e}]^{\mathrm{T}}\\ &③\text{ otherwise},\\ &\quad [\alpha_i^{t+1}\beta_i^{t+1}]^{\mathrm{T}}=[\alpha_i^{t+1}\beta_i^{t+1}]^{\mathrm{T}}\end{aligned}\tag{16}$$

从式(16)可以看出,本文提出的修正操作在量子比特的 α 或 β 的值过于接近 0 或 1 时对其进行修正,从而增加观测种群的多样性,避免算法陷入早熟收敛.

4.4　量子观测算子

观测算子采用随机数的方法把量子种群观测成普通种群.对于每一个量子位,随机产生一个数 $r\in[0,1]$,如果它比 $|\alpha_i|^2$ 小,则普通染色体 $P(t)$ 中相应的位取'0',否则取'1'.具体观测流程如图 1 所示.

```
Procedure:
begin
    i: = 0
    while (i < L) do
        begin
            i: = i + 1
            if random[0,1] < |α_i^t|^2
            then p_i = 0
            else p_i = 1
            end
        end
end
```

图 1　量子观测算子

4.5　量子交叉算子

本文算法作采用了文献[9,10,17,18]中的量子全干扰交叉算子对观测后的种群进行交叉操作.这种量子交叉操作模拟了量子干涉特性,充分利用种群中的尽可能多的染色体的信息,有利于增加种群多样性,避免算法早熟收敛.

4.6　算法复杂度分析

假设种群大小为 pop,每个个体的长度为 L,迭代次数为 G,则量子更新算子的时间复杂度为:$O(pop\times L)$,量子全干扰交叉的时间复杂度为 $O(pop\times L)$,所以本文算法的时间复杂度为 $O(pop\times L\times G)$.

5　仿真实验及结果分析

为了测试算法的性能,本文对三类共 13 个数据集进行了测试,并设计了三组实验了测试 IQEAC 算法的性能.实验中,距离测度采用欧氏距离.IQEAC 算法与改进旋转门前的 QEAC 算法,*K*-均值算法,免疫克隆选择 ICSA 算法进行了比较.

5.1　实验数据集

Group1 数据集:该数据集为二维数据,主要呈现球型分布[20].Data_3_2 为三类 76 个数据点;Data_5_2 为五类 250 个数据点;Data_6_2 为六类 300 个数据点.

Group2 数据集:该数据集簇内呈对称、簇间不对称分布[21].X2line 是两类条形分布数据,每类 200 个数据;X2ring 为两类环形分布数据,每类 200 个数据;X3mixed-clusters 为三类不同分布的数据,共 400 个数据点;X577 为三类 577 个数据点.

Group3 数据集:该数据来自 UCI 数据集[22].Iris 为 150 个 4 维 3 类数据集;Wine 为 178 个 13 维 3 类数据集;glass 为 214 个 9 维 6 类数据集;lungCancer 为 32 个 56 维 3 类数据集;liver_disorder 为 345 个 6 维 2 类数据集;new_thyroid 为 215 个 5 维 3 类数据集.

5.2　实验结果分析

实验中,改进旋转门的量子进化聚类算法 IQEAC,种群规模为 10,全干扰交叉概率 $p_c=0.2$,初始旋转角度 θ_0 为 0.05π,H 门的阈值 e 为 10^{-5};量子进化聚类算法 QEAC,采用量子进化算法[9,10,18]进行数据聚类,种群规模为 10,交叉概率 $p_c=0.2$,旋转角度为 0.05π;免疫克隆选择算法 ICSA 采用实数编码,初始种群规模为 20,采用自适应克隆比例的方法,克隆个数为 $(f_i/SumFit)\times 100$,*Fit* 为适应度,*SumFit* 为适应度总和,变异概率为 $2/(k*t)$的高斯变异,k 为类别数,t 为数据的维数,采用交叉概率 $p_c=0.8$ 单点交叉.以上三种算法的终止条件都为:如果满足连续 10 次相邻两代种群的最佳适应度之差小于 10^{-10},则结束,否则达到最大迭代次数结束;*K*-均值算法终止条件为达到最大迭代次数.所有算法的最大迭代次数都为 100.

为了评价算法的性能,本文把聚类正确率作为评价标准.聚类正确率定义为:

$$\text{聚类正确率}=\frac{\text{正确分类的样本个数}}{\text{样本总数}}\tag{17}$$

实验 1　量子旋转门改进前后聚类效果比较

为了测试本文改进的量子旋转门的性能,本文把 IQEAC 算法与改进前的量子进化聚类算法 QEAC,*K*-均值,免疫克隆选择算法 ICSA 在三组数据集上进行了比较,表 1 为 30 次独立运行结果的统计值.从实验结果可以看出,对于 Group1 数据集,聚类正确率都不错,但是 IQEAC 算法有两个数据集都达到了 100% 的正确率;而 Group2 和 Group3 数据集绝大部分的数据,本文的 IQEAC 算法都比其他算法在聚类正确率上都有提高,说明在算法相同的情况下,算子的设计的好坏也决定着算法实验效果的好坏.本文改进的旋转门算子,增加了 IQEAC 算法的全局搜索能力,增加了种群的多样性,更容易找到较好的解.

ICSA是基于免疫克隆选择的聚类算法.研究表明[12,14,17,19],免疫克隆选择算法与进化算法相比能够更好地保持种群多样性,因此在分类精度上 ICSA 同样表现出了较好的性能.但是由于 ICSA 采用了克隆选择的进化流程,在收敛速度上要比 IQEAC 慢.下面的实验将对比算法在运行时间上的性能.

表 1 *K*-均值,ICSA,QEAC 和 IQEAC 四种算法聚类正确率表

数据 \ 算法 \ 正确率		聚类正确率(%)			
		K-均值	ICSA	QEAC	IQEAC
Group1	Data_3_2	100	100	96.40	100
	Data_5_2	94.47	94.01	81.88	95.84
	Data_6_2	92.22	99.33	86.94	100
Group2	X2line	77.37	82.25	82.52	83.09
	X2ring	83.09	87.62	86.08	87.73
Group3	X3mixedclusters	81.82	81.60	80.58	82.47
	X577	95.74	96.71	94.74	97.20
	Iris	88.67	88.67	78.89	88.82
	Wine	91.16	95.24	60.43	95.49
	glass	51.37	54.60	43.57	55.95
	lungCancer	52.19	56.12	50.02	50.21
	liver_disorder	57.97	57.97	57.97	57.97
	new_thyroid	75.35	89.30	82.26	90.50

实验 2 算法运行时间比较

为了测试 IQEAC 算法的运行效率,本文把 IQEAC 算法与 ICSA 算法在运行时间上做了一个比较,运行环境为 Pentium(R) 4 CPU,2GB 内存,取种群大小都为 10,算法运行到最大迭代次数 100 代终止,表 2 为 30 次运行的平均时间.

本文分别抽取三组数据中两种算法聚类结果相差不大的数据集进行了时间测试,从表 2 可以看出,在聚类正确率相当或者 IQEAC 算法稍好的情况下,本文算法所用的时间远远小于 ICSA 的运行时间,说明本文的算法的搜索效率高.

表 2 IQEAC 和 ICSA 算法运行时间比较表

数据		聚类算法	聚类正确率(%)	所用时间(s)
Group1	Data_6_2	ICSA	99.33	62.94
		IQEAC	100	13.67
Group2	X2line	ICSA	82.25	38.77
		IQEAC	83.09	6.50
Group3	Iris	ICSA	88.67	15.57
		IQEAC	88.82	10.54
	liver_disorder	ICSA	57.97	31.81
		IQEAC	57.97	11.78
	new_thyroid	ICSA	89.30	21.19
		IQEAC	90.50	13.39

6 基于点对称距离的量子进化聚类算法

本文对量子进化算法进行改进,并用于数据聚类,取得了良好的效果.但是,相似性度量对于算法的实验结果也起着很重要的作用.为了测试算法的性能,本文引入点对称距离[20,21],对算法进行了测试,并与同类算法进行了比较.

本文定义阈值 θ 为每个数据点的最近邻中距离最大的值,即:

$$\theta = \max_{i=1\cdots n} d_{NN}(\bar{x}_i) \tag{20}$$

其中 $d_{NN}(\bar{x}_i)$为 $\bar{x}_i$ 到它的最近邻的距离,如果计算的点对称距离与欧式距离之比小于该阈值,则按照点对称距离来对数据聚类,否则按照欧氏距离来聚类.

对比实验及结果分析

在距离测度采用对称距离后,本文将基于点对称距离的量子进化聚类算法(PSQEAC)与采用对称距离的 *K*-均值算法 SBKM[21],采用改进的点对称距离的遗传算法 GAPS[20]进行了比较,同样对 5.1 节的三组数据集进行了测试.

采用点对称距离的量子进化聚类算法 PSQEAC,种群规模为 10,全干扰交叉概率 $p_c = 0.2$,初始旋转角度 θ_0 为 0.05π,最大迭代次数 100 次,终止条件与 IQEAC 算法相同;采用点对称距离的 *K*-均值聚类算法 SBKM[21],距离阈值取 0.18,算法终止条件为达到最大迭代次数;采用点对称距离的遗传算法 GAPS 使用文献[20]中实验设置.本文比较了改进距离测度后,用点对称距离的 SBKM,GAPS 和本文的 PSQEAC 算法的性能,每种算法在每个数据集上独立运行 30 次,取平均正确率,实验结果如表 3 所示.

表 3 SBKM,GAPS 和 PSQEAC 三种算法聚类正确率表

数据 \ 算法 \ 正确率		聚类正确率(%)		
		SBKM	GAPS	PSQEAC
Group1	Data_3_2	86.67	100	100
	Data_5_2	26.40	94.01	94.39
	Data_6_2	63.66	99.33	99.33
Group2	X2line	90.80	95.48	94.52
	X2ring	72.65	88.67	90.97
	X3mixedclusters	86.75	95.27	95.54
	X577	65.86	99.13	99.13
Group3	Iris	63.73	90.70	89.33
	Wine	93.73	96.15	95.39
	glass	53.58	57.78	55.14
	lungCancer	54.69	56.12	55.95
	Liver_disorder	57.97	57.97	57.97
	New_thyroid	77.67	82.31	82.03

从表 3 的实验结果可看出,采用点对称距离后,本文的基于量子进化算法的聚类算法 PSQEAC 的性能比 GAPS 算法在大部分数据上都有提高.说明采用点对称距离后,本文改进的量子进化算法的性能并不受影响,针对不同类型的数据集,都能够得到不错的聚类效果.

7　总结与展望

本文提出的基于量子进化的数据聚类算法,用改进量子旋转门的量子进化算法来对聚类中心进行全局搜索,与改进前的算法相比,取得了比较好的聚类正确率和收敛速度.然后,在采用新的距离测度:对称距离后,本文的算法性能并没有下降,也取得了较好的聚类正确率.通过对三组数据的聚类实验表明,本文的算法比其他同类算法在聚类正确率和速度上都有明显的提高.

参考文献

[1] 李洁.基于自然计算的模糊聚类算法研究[D].西安:西安电子科技大学,2004.

[2] Anderberg M R. Cluster Analysis for Applications[M]. New York: Academic Press, 1973.

[3] 焦李成,刘芳,缑水平,等.智能数据挖掘与知识发现[M].西安:西安电子科技大学出版社,2006.

[4] Jain A K, Dubes R C. Algorithms for Clustering Data[M]. New Jersey: Prentice-Hall, 1988.

[5] Bezdek J C. A convergence theorem for the fuzzy ISODATA clustering algorithms[J]. IEEE Transactions on Pattern Anal Machine Intell, 1980, PAMI-2(1): 1 – 8.

[6] Ujjwal Maulik, Sanghmitra Bandyopadhyay. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2000, 33(9): 1455 – 1465.

[7] Lawrence O Hall, Ibrahim Burak Ozyurt, James C Bezdek. Clustering with a genetically optimized approach[J]. IEEE Transaction on Evolutionary Computation, 1999, 3(2): 103 – 112.

[8] K H Han, J H Kim. Quantum-inspired evolutionary algorithms with a new termination criterion, H_ε gate, and two-phase scheme[J]. IEEE Transaction on Evolutionary Computation, 2004, 8(4): 156 – 169.

[9] Yangyang Li, Jingjing Zhao, Licheng Jiao. Quantum-Inspired evolutionary multicast algorithm[A]. Proceeding of the 2009 IEEE International Conference on Systems, Man, and Cybernetics[C]. USA: IEEE press, 2009. 1496 – 1501.

[10] L C Jiao, Y Y Li, M G Gong, X R Zhang. Quantum-inspired immune clonal algorithm for global optimization[J]. IEEE Transaction on Systems, Man, and Cybernetics, Part B, 2008, 38(5): 1234 – 1253.

[11] 陈国良,王煦法,等.遗传算法及其应用[M].北京:人民邮电出版社,1996.

[12] 焦李成,杜海峰,刘芳,公茂果,等.免疫优化计算、学习与识别[M].北京:科学出版社,2006.

[13] 焦李成,刘静,钟伟才.协同进化计算与多智能体系统[M].北京:科学出版社,2006.

[14] Jiao Licheng, Wang Lei. A novel genetic algorithm based on immune[J]. IEEE Trans on System, Man, and Cybernetics – Part A, 2000, 30(5): 1 – 10.

[15] Narayanan A. Quantum computing for beginners[A]. Proceedings of the 1999 Congress on Evolutionary Computation[C]. Piscataway, New Jersey: IEEE Press, 1999, 2231 – 2238.

[16] Abrams D S, Lloyd S. Nonlinear quantum mechanism implies polynomial-time solution for NP-complete and NP problems[J]. Physical Review Letters, 1998, 81(18): 3992 – 3995.

[17] Ronghua Shang, Licheng Jiao, Yangyang Li, Jianshe Wu. Quantum immune clonal algorithm for multi-objective 0/1 knapsack problems[J]. Chinese Physics Letters, 2010, 27(1): 010308.

[18] 焦李成,等.多目标优化免疫算法、理论和应用[M].北京:科学出版社,2010.1.

[19] 马文萍,焦李成,张向荣,李阳阳.基于量子克隆优化的SAR图像分类[J].电子学报,2007,35(12):2241 – 2246.
Ma Wen-ping, Jiao Li-cheng, Zhang Xiang-rong, Li Yang-yang. SAR image classification based on quantum clonal optimization[J]. Acta Electronica Sinica, 2007, 35(12): 2241 – 2246. (in Chinese)

[20] Sanghamitra Bandyopadhyay, Sriparna Saha. GAPS: A clustering method using a new point symmetry-based distance measure[J]. Machine Intelligence Unit, Indian Statistical Institue. 2007, 40(12): 3430 – 3451.

[21] Mu Chun Su, Chien Hsing Chou. A modified version of the K-means algorithm with a distance based on cluster symmetry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 674 – 680.

[22] David Aha. The UCI Machine Learning Repository[EB/OL]. http://archive.ics.uci.edu/ml/datasets.html.

作者简介

刘　芳　女,1963年2月出生于北京,湖南华容人.西安电子科技大学计算机学院教授,博士生导师,学科带头人.主要研究方向包括:人工智能、图像处理、机器学习、进化计算等.
E-mail: lf204310@163.com