# 基于流形距离的量子进化聚类算法

李阳阳,石洪竺,焦李成,马文萍
(西安电子科技大学智能感知与图像理解教育部重点实验室,陕西西安 710071)

**摘 要:** 基于量子计算的机理和特性,并结合进化计算,本文提出了一种新颖的量子进化聚类算法(QEAM),在该聚类算法中引入了一种新的距离测度函数——流形距离.新方法将聚类归属为优化问题,通过运用量子进化的机理更快地搜索到最优聚类中心,从而得到最优隶属度矩阵划分;同时,通过基于流形距离的相似性度量,有效利用样本所具有的全局一致性信息,充分挖掘样本的空间分布信息,对样本进行正确的类别划分.将本文算法(QEAM)与基于流形距离的免疫进化算法(IEAM),遗传聚类算法(GAC)以及模糊 C-均值算法(FCM)进行了性能比较,对 6 个人工数据集和 3 个 UCI 数据集的仿真实验结果显示,QEAM 对样本空间分布复杂的聚类问题具有较高的准确率和较好的鲁棒性.

**关键词:** 量子计算; 量子进化算法; 数据聚类; 流形距离
**中图分类号:** TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 10-2343-05

# Quantum-Inspired Evolutionary Clustering Algorithm Based on Manifold Distance

LI Yang-yang, SHI Hong-zhu, JIAO Li-cheng, MA Wen-ping
(*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an* 710071, *China*)

**Abstract:** Based on the concepts and principles of quantum computing, a novel quantum-inspired evolutionary algorithm for data clustering (QEAM) is proposed in this paper by using a novel distance measurement index called manifold distance which can measure the geodesic distance along with the manifold. The clustering problem is viewed as an optimization problem. Our main motives of using QEAM consist in searching for appropriate cluster center by using the principles of quantum evolutionary computation, so that a similarity metric of clusters are optimized more quickly and effectively. The experimental results on six artificial datasets and three UCI datasets show the superiority of QEAM over an immune evolutionary clustering algorithm with manifold distance (IEAM), a genetic algorithm for clustering (GAC) and fuzzy c-means algorithm (FCM).

**Key words:** quantum computation; quantum-inspired evolutionary algorithm; data clustering; manifold distance

## 1 引言

在现实生活中,数据聚类已广泛应用在数据挖掘、计算机视觉、信息检索以及模式识别等领域[1,2].聚类就是按照一定的要求和规律对事物进行区分和分类的过程.在这个过程中没有任何关于分类的先验知识,仅靠事物间的相似性作为类别划分的准则.在现有的聚类算法中,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,具有深厚的泛函基础,是聚类算法研究的重要分支之一,其中模糊 C-均值算法(fuzzy c-means algorithm,简称 FCM)[3]是应用最为广泛的聚类算法之一.针对经典算法对初始化敏感、容易陷入局部最优解由此而产生各种错误分类的缺点,可以利用现有的模拟退火(simulated annealing,简称 SA)、遗传算法(genetic algorithm,简称 GA)等先进的优化算法对目标函数优化,从而使聚类算法得到全局最优解的概率大大增加.但现有的进化算法也存在耗时以及波动性大等问题,因此,迫切需要构造性能优良的聚类算法.

本文借鉴量子计算的并行特性,采用量子位编码种群中的染色体,这种表达方式使种群染色体携带了丰富的信息,并结合进化算法,由此提出了一种新颖的聚类算法—基于流形距离的量子进化聚类算法(QEAM).算

收稿日期:2009-04-02;修回日期:2011-07-20
基金项目:国家自然科学基金(No.60803098,No.61001202,No.61003199);陕西省"13115"科技创新工程重大科技专项项目(No.2008ZDKG-37);高等学校学科创新引智计划(111 计划)(No.B07048);中国博士后科学基金特别资助(No.200801426,No.201104618);中国博士后科学基金(No.20080431228,No.20090451369,No.20090461283);陕西省自然科学基金(No.2009JQ8015,No.2010JM8030,No.2010JQ8023);中央高校基本科研业务费专项资金(No.JY10000902039,No.JY10000902040,No.JY10000903007,No.K50510020011)

法中采用量子位表示染色体,使得作用在量子编码染色体上的操作具有高效的并行性,为防止盲目的搜索,利用当前最优染色体的信息来控制更新,使种群以大概率向着优良模式进化来加速收敛.但随着问题的复杂求解能力不尽人意,因此我们在本文算法中考虑在各个子群体间采用量子交叉操作增强信息交流,在各个子群体内部采用量子旋转门对染色体进行进化,并动态调整旋转角度,使算法在全局搜索的同时兼顾局部.在文献[4]中将量子免疫克隆算法应用到数值优化问题上,取得了良好的效果.本文将改进后的量子进化算法用于求解聚类问题并结合流形聚类测度函数,并与遗传聚类算法(GAC)[5],基于流形距离的免疫进化聚类算法(IEAM)[6]以及模糊 C-均值算法(FCM)[3]相比较.实验结果表明 QEAM 性能优越,具有较强的实用价值.

## 2 量子进化聚类算法

### 2.1 算法描述

**step 1** 设定一个较小正整数,随机产生初始种群 $Q(t)$,中 $\alpha_i^t,\beta_i^t(i=1,2,\cdots,m)$都以等概率 $1/\sqrt{2}$初始化;

**step 2** 将量子染色体 $Q(t)$观测成为二进制染色体 $P(t)$;

**step 3** 计算个体适应度函数 $f_k$,保留当前群体中的最优个体;

**step 4** 更新 $Q(t)$:量子旋转门操作得到 $Q_m(t)$;

**step 5** 将量子染色体 $Q_m(t)$观测成为二进制染色体 $P_m(t)$;

**step 6** 更新 $P_m(t)$:进行量子交叉操作得到,并且计算每个个体适应度,保留所有种群中的最优个体;

**step 7** 选择操作,得到 $P(t+1)$;

**step 8** 如果满足终止条件 $s_c$ 则转向 step9,否则转向 step4;

**step 9** 对最好的个体进行译码,计算出聚类原型,再计算出各个样本的分类结果,这个结果就为数据集的聚类结果.

### 2.2 算子的设计

(1)编码方式

本文选择对聚类中心点编码.本文使用一个 10 位的量子染色体来表示一个聚类中心的一维特征.如果数据具有 $n$ 维特征,若将数据聚为 3 类,每个个体将对应 3 个聚类中心,相应的量子进化聚类算法中的每个个体将是 $30^n$ 个量子比特组成的串.

(2)观测操作

通过观察 $Q(t)$的状态,产生一组普通解 $P(t)$,其中在第 $t$ 代中 $P(t)=\{x_1^t,x_2^t,\cdots,x_n^t\}$,每个 $x_j^t(j=1,2,\cdots,n)$是长度为 $m$ 的串$(x_1x_2,\cdots,x_m)$,它是由量子比特幅度 $|\alpha_i^t|^2$ 或 $|\beta_i^t|^2(i=1,2,\cdots,m)$得到的,如在二进制情况下的过程是:随机产生一个[0,1]数,若它大于 $|\alpha_i^t|^2$,取 1,否则取 0.

(3)相似性度量

在面对某些实际的聚类问题,数据的分布往往具有不可预见的复杂分布,导致了基于欧氏距离的相似性度量无法反映出聚类的全局一致性[7].

将数据点看作是无向图 $G=(V,E)$的顶点,令 $p\in V^l$ 表示图上一个长度为 $l=|p|$的连接点 $p_1$ 和 $p_{|p|}$的路径,其中边$(p_k,p_{k+1})\in E,1\leqslant k\leqslant p_{|p|}$.令 $p_{ij}$表示连接数据点 $x_i,x_j$ 的所有路径的集合,流形距离的定义如下[8]:

$$D_{ij}=\min\sum_{k=1}^{|p|-1}L(p_k,p_{k+1}) \tag{1}$$

$$L(x_i,x_j)=\rho^{dist(x_i,x_j)}-1 \tag{2}$$

其中 $\rho$ 表示伸缩因子,且 $\rho>1$, $dist(x_i,x_j)$表示两点间的欧氏距离.

(4)量子变异

在量子理论中,各个状态间的转移是通过量子门变换矩阵实现的,我们发现:用量子旋转门的旋转角度同样也可表征量子染色体中的变异操作,进而方便的在变异中加入最优个体的信息,加快算法收敛[9].在 0、1 编码的问题中,我们可以设计下面这种量子变异算子来加速进化求优:

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

其中 $U(\Delta\theta)$表示量子旋转门,旋转变异的角度 $\Delta\theta$ 可由表 1 得到,具体有如下定义:

$$\Delta\theta_i=\delta\times s(\alpha_i,\beta_i) \tag{4}$$

**表 1 变异角 $\Delta\theta$ 查询表**

| $x_i$ | $best_i$ | $f(x)\geqslant f(best)$ | $\Delta\theta_i$ | $s(\alpha_i\beta_i)$ | | | |
|---|---|---|---|---|---|---|---|
| | | | | $\alpha_i\beta_i>0$ | $\alpha_i\beta_i<0$ | $\alpha_i=0$ | $\beta_i=0$ |
| 0 | 0 | 假 | $\delta$ | 0 | 0 | 0 | 0 |
| 0 | 0 | 真 | $\delta$ | 0 | 0 | 0 | 0 |
| 0 | 1 | 假 | $\delta$ | 0 | 0 | 0 | 0 |
| 0 | 1 | 真 | $\delta$ | −1 | +1 | ±1 | 0 |
| 1 | 0 | 假 | $\delta$ | −1 | +1 | ±1 | 0 |
| 1 | 0 | 真 | $\delta$ | +1 | −1 | 0 | ±1 |
| 1 | 1 | 假 | $\delta$ | +1 | −1 | 0 | ±1 |
| 1 | 1 | 真 | $\delta$ | +1 | −1 | 0 | ±1 |

其中 $x_i$ 为当前染色体的第 $i$ 位; $best_i$ 为当前的最优染色体的第 $i$ 位; $f(x)$为适应度函数,显然 $\delta$ 为旋转角度的大小,控制算法收敛的速度,在本文算法中我们采用动态调整策略,将 $\delta$ 根据进化代数控制在 $0.005\pi$ 到 $0.1\pi$ 之间; $s(\alpha_i\beta_i)$为旋转角度的方向,保证算法的收敛.

(5)量子交叉

交叉是 GA 的另一种搜索最优解的手段,通常采用的交叉操作如单点交叉、多点交叉、均匀交叉、算术交叉等,它们的共同点是限制在两个个体之间,当交叉的两个个体相同时,它们都不再奏效.在这里,我们使用量子的相干特性构造一种新的交叉操作—"全干扰交叉"[4].在这种交叉操作中,种群中的所有染色体均参与交叉,其方式详见文献[4].这种量子交叉可以充分利用种群中的尽可能多的染色体的信息,改进普通交叉的局部性与片面性,在种群进化出现早熟时,它能够产生新的个体,给进化过程注入新的动力.这种交叉操作借鉴的是量子的相干特性,可以克服普通染色体在进化后期的早熟现象.

(6)选择操作

算法中采用赌轮选择策略,除此之外,我们还加入了精英选择策略.通常的选择方法就是高适应度的个体被选择保留下的几率会很大.采用精英选择策略可以保证某一代的最优解在整个进化过程中可以毫发无损地被保留下来.即就是说,在某一代中的最优解的适应度函数值优于当前最优解的适应度函数值,那么当前最优解就被该最优解所代替.

(7)停机条件

为了得到好的聚类结果,适当的停机条件是很必要的.遗传算法常常以设定最大迭代次数为停机条件.在我们的算法中以设定连续无改进次数为停机条件 $e$.例如:我们设定 $\varepsilon = 10^{-5}$ 为停止阈值,当前个体的适应度值与之前个体的适应度值的改变量小于这个阈值就称为无改进,那么连续无改进次数就是 $e = 1$,反之就是有改进.$e$ 置零.当 $e = 10$ 时,就是连续 10 次无改进,算法停止.

## 3　时间复杂度分析

设种群规模 $M$ 以及编码长度为 $N$,则算法每迭代一次的时间复杂性可按以下计算:量子变异操作的时间复杂度为 $O(M \times N)$;量子交叉操作的时间复杂度为 $O(M \times N)$.因此总的时间复杂度最差为:$O(M \times N) + O(M \times N)$.根据符号 $O$ 的运算规则并化简,QEAM 每迭代一次的时间复杂度最差为:$O(M \times N)$.

## 4　实验分析

为了能够直观的考察本文所提算法的性能,我们将该算法 QEAM 应用于 6 个人工数据和 3 个 UCI 数据的聚类问题,这 6 个人工数据分别是 Three circles, Spiral, Squar4, Sticks, Long1, Line-blobs,它们分别具有不同的流形结构(如图 1 所示);3 个 UCI 数据集分别是 Iris 数据集,Wine 数据集以及 Glass 数据集.本文将 QEAM 与 IEAM, FCM, 以及 GAC 进行性能比较,其中 QEAM 和 IEAM 采用流形距离测度,FCM 和 GAC 采用欧氏距离测度.

对于聚类结果,本实验采用了聚类正确率这个指标来衡量;对于算法的聚类性能,本实验采用了指标 Adjusted Rand index[10] 来度量.下面给出 Ajusted Rand Index 的计算式:

$$R(U,V) = \frac{\sum_{lk}\binom{n_{lk}}{2} - \left[\sum_{l}\binom{n_{l\cdot}}{2}\sum_{k}\binom{n_{\cdot k}}{2}\right]\Big/\binom{n}{2}}{0.5\left[\sum_{l}\binom{n_{l\cdot}}{2}+\sum_{k}\binom{n_{\cdot k}}{2}\right]-\left[\sum_{l}\binom{n_{l\cdot}}{2}\sum_{k}\binom{n_{\cdot k}}{2}\right]\Big/\binom{n}{2}} \tag{5}$$

其中 $n_{lk}$ 表示那些既属于类属 $l$ 又属于类属 $k$ 的数据个数,$R(U,V) \in [0,1]$.其数值越大说明划分的正确性越高.

QEAM 的参数设置如下:种群规模 $N = 20$,IEAM 的参数设置如下:种群规模 $N = 20$,疫苗长度 $v = 10$,交叉概率 $p_c = 0.75$,变异概率 $p_m = 0.1$,疫苗接种概率 $p_v = 0.3$,GAC 的参数设置如下:种群规模 $N = 20$,交叉概率 $p_c = 0.75$,变异概率 $p_m = 0.1$,FCM 的参数设置如下:模糊指数 $m = 2.0$,以上 4 种算法的更新阈值均是:$\varepsilon = 10^{-5}$,连续无改进次数 $e = 10$,收缩因子 $\rho = e^2$.我们对每一个数据集独立运行 20 次,各算法在求解以上 9 个问题时得到的平均结果如表 2 所示.

**表 2　各个数据集在 4 种聚类算法的结果比较**

| Dataset | 聚类正确率 | | | | Adjusted Rand index | | | |
|---|---|---|---|---|---|---|---|---|
| | QEAM | IEAM | GAC | FCM | QEAM | IEAM | GAC | FCM |
| Three circles | 1 | 1 | 0.428 | 0.414 | 1 | 1 | 0.026 | 0.024 |
| Spiral | 1 | 1 | 0.642 | 0.639 | 1 | 1 | 0.080 | 0.076 |
| Square4 | 0.923 | 0.916 | 0.934 | 0.933 | 0.806 | 0.789 | 0.832 | 0.830 |
| Sticks | 1 | 1 | 0.845 | 0.849 | 1 | 1 | 0.664 | 0.676 |
| Long1 | 1 | 0.967 | 0.464 | 0.486 | 1 | 0.946 | 0.014 | 0.016 |
| Line-blobs | 1 | 1 | 0.797 | 0.801 | 1 | 1 | 0.487 | 0.494 |
| Iris | 0.967 | 0.96 | 0.893 | 0.887 | 0.904 | 0.886 | 0.728 | 0.715 |
| Glass | 0.458 | 0.379 | 0.318 | 0.322 | 0.213 | 0.182 | 0.161 | 0.166 |
| Wine | 0.938 | 0.826 | 0.949 | 0.949 | 0.819 | 0.556 | 0.85 | 0.85 |

从表 2 中的统计数据以及图 1 可以明显看出,对流形结构明显、非球形分布的 Three-circles, Spiral, Sticks, Line-blobs 和 Long1 5 个问题,以流形距离作为测度的 QEAM 和 IEAM 能够正确地对类别进行划分,但是以欧式距离作为相似性度量的 GAC 和 FCM 都很差.对于流形结构不明显、呈球状分布的 Square4 这个问题,4 种算法均没有获得完全正确的划分.对于 3 个 UCI 数据集,除了 Wine 问题外,QEAM 也能明显地体现出其优势.

同时采用流形距离作为相似性度量的 QEAM 以及 IEAM 算法,为了凸显 QEAM 算法较 IEAM 算法的优越性,对每个数据集 20 次的平均运行时间逐一进行统计,

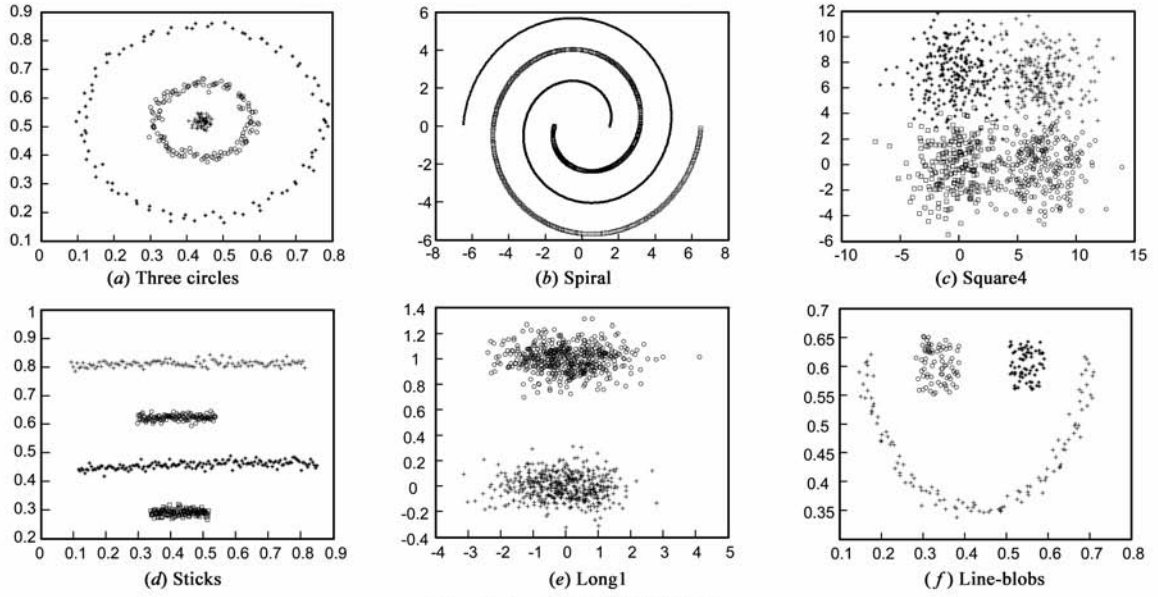


图1 6个人工数据集空间结构图

且所有实验运行在配置为如下的 PC 机上：CPU 为 2.33GHz Pentium IV、内存为 2G、实用的程序语言为 Matlab 6.0，见表 3.

表 3 基于流形测度的两种算法时间比较

| Dataset | Time | |
|---|---|---|
| | QEAM | IEAM |
| Three circles | 68.81s | 110.28s |
| Spiral | 1814.28 | 3295.54 |
| Square4 | 1524.35s | 3747.82s |
| Sticks | 488.47s | 712.76s |
| Long1 | 1364.71s | 3659.04s |
| Line-blobs | 52.84s | 99.92s |
| Iris | 21.14s | 38.24s |
| Glass | 44.37s | 75.29s |
| Wine | 24.79s | 42.90s |

从上述实验结果中我们可以得出：较 IEAM 算法，QEAM 参数设置少，简单易于实现，从表 2 中可以得到 QEAM 比 IEAM 算法在时间上也有明显优势.

## 5 鲁棒性分析

为了考察以上 4 种算法的鲁棒性，我们将 4 种算法在求解这 9 个数据集时的鲁棒性进行分析与比较. 系统地说，算法 $m$ 在某个数据集上的相对性能用该算法所获得的 Adjusted Rand Index 的值与所有算法在求解该问题时得到的最大 Adjusted Rand Index 的值来衡量[11]. 因此，在某个数据集上表现最好的算法的相对性能为 1，其他算法相对性能 $b_m \leqslant 1$. 算法 $m$ 在所有数据集上的鲁棒性总和可以用于评价算法鲁棒性. 总和越大鲁棒性越好. 见图 2 所示，4 种算法的鲁棒性比较.

从上图可以看出，基于流形距离的测度函数的两个算法明显优于基于欧氏距离测度的两个算法. QEAM 的总和值达到了 8.933，IEAM 的总和值也达到了 8.382，但是基于欧氏距离的两个算法的总和只有 5.552 和 4.854. 从以上分析可以说明，本文提出的算法具有优越的鲁棒性.

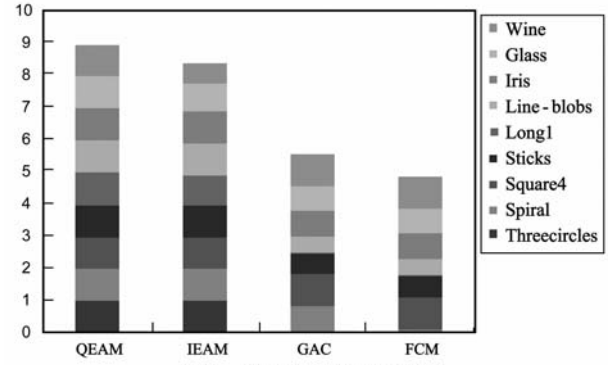


图2 4种算法的鲁棒性分析

## 6 结论

基于量子计算特性，结合进化算法，通过引入流形距离测度，提出了一种新的用于解决复杂聚类问题的量子进化聚类算法. 算法中我们将量子叠加性应用于传统进化算法中去，较 IEAM 算法，我们提出的方法参数设置少，简单易于实现，而量子旋转门和量子交叉的使用有效的避免了传统遗传算法早熟和收敛速慢的瓶颈问题，而相较于 FCM 算法，本文算法对初始值不敏感，稳定且多样性强易于跳出局部最优解. 实验结果表明，与 FCM，IEAM 和 GAC 方法相比，无论从聚类精度还是鲁棒性以及处理时间上，都显示出 QEAM 要优于上述三种聚类方法. 接下来我们的研究将致力于本文算法的实际应用，如 SAR 图像分割和识别问题.

## 参考文献


[1] 曹芳,洪文,吴一戎.基于 Cloude-Pottier 目标分解和聚合的层次聚类算法的全极化 SAR 数据的非监督分类算法研究[J].电子学报,2008,36(3):543 – 546.
Cao Fang, Hong Wen, Wu Yi-rong. An unsupervised classification for fully polarimetric SAR data using cloude-pottier decomposition and agglomerative hierarchical clustering algorithm[J]. Acta Electronica Sinica, 2008, 36(3): 543 – 546. (in Chinese)

[2] 刘若辰,沈正春,贾建,焦李成.基于免疫优势的克隆选择聚类算法[J].电子学报,2010,38(4):960 – 965.
Liu Ruo-chen, Shen Zheng-chun, et al. Immunodomaince based on clonal selection clustering algorithm[J]. Acta Electronica Sinica, 2010, 38(4): 960 – 965. (in Chinese)

[3] 西奥多里德斯,等.模式识别[M].北京:电子工业出版社,2006.
Sergios Theodoridis, et al. Pattern Recognition[M]. Beijing: Publishing House of Electronics Industry, 2006.

[4] Jiao L C, Li Y Y, Gong M G, et al. Quantum-inspired immune clonal algorithm for global optimization[J]. IEEE Transactions on System, Man, and Cybernetics, Part B, 2008, 38(5): 1234 – 1253.

[5] Ujjwal Maulik, Sanghamitra Bandyopadhyay. Genetic algorithm-based clustering technique[J]. Pattern Recognition 2000, 33: 1455 – 1465.

[6] 刘静,钟伟才,刘芳,焦李成.免疫进化聚类算法[J].电子学报,2001,29(12A):1868 – 1872.
Liu Jing, Zhong Wei-cai, et al. A novel clustering based on the immune evolutionary algorithm. Acta Electronica Sinica, 2001, 29(12A): 1868 – 1872. (in Chinese)

[7] Zhou D, Bousquet O, La T N, Weston J., Scholkopf B. Learning with local and global consistency[J]. Advances in Neural Information Processing System. UAS: MIT Press, 2004, 16: 321 – 328.

[8] 公茂果,焦李成,马文萍,张向荣.基于流形距离的人工免疫无监督分类与识别算法[J].自动化学报,2008,34(3):368 – 376.
Gong Mao-guo, Jiao Li-cheng, et al. Unsupervised classification and recognition using an artificial immune system based on manifold distance[J]. Acta Automatic Sinica, 2008, 34(3): 368 – 376. (in Chinese)

[9] Han K-H, Kim J-H. Quantum-inspired evolutionary algorithms with a new termination criterion, $H_\varepsilon$ gate, and two-phase scheme [J]. IEEE Transactions on Evolutionary Computation, 2004, 8(6): 156 – 169.

[10] Handl J, Knowles J. An evolutionary approach to multiobjective clustering[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(1): 56 – 76.

[11] Geng X, Zhan D C, Zhou Z H. Supervised nonlinear dimensionality reduction for visualization and classification[J]. IEEE Transactions on System, Man, and Cybernetics, Part B, 2005 35(6): 1098 – 1107.


## 作者简介


**李阳阳**　女,1979 年出生于河南开封,博士,西安电子科技大学副教授,电子学会会员,主要研究领域为量子计算智能和模式识别.
E-mail: yyli@xidian.edu.cn

**焦李成**　男,1959 出生于陕西白水,西安电子科技大学教授,博士生导师,电子学会高级会员,主要研究领域为智能信息处理和图像处理.
E-mail: jlcxidian@163.com

**石洪竺**　女,1986 年出生于贵州思南,硕士,主要研究领域为量子计算智能和和模式识别.
E-mail: shihongzhu1985@163.com

**马文萍**　女,1981 年出生于陕西铜川,博士,西安电子科技大学副教授,主要研究领域为自然计算和图像处理.
E-mail: wpma@mail.xidian.edu.cn