基于广义量子粒子模型的聚类算法及收敛性研究

黄良军[1] 帅典勋[1,2] 张 彬[1]
(华东理工大学计算机科学与技术系 上海 200237)[1]
(清华大学智能技术与系统国家重点实验室 北京 100084)[2]

摘 要 提出了一种应用广义量子粒子模型进行自组织聚类的新方法。该模型将数据聚类过程转化为一个量子粒子在状态构形空间上的随机自组织过程,由量子粒子之间相互纠缠形成的状态构形随时间不断演化,最终会收敛到一个平稳的概率分布,最优状态空间构形与平稳概率分布中具有最大概率的状态构形相对应。对此自组织过程的收敛性进行了理论上的证明。与传统的适用于大规模数据的聚类方法相比较,该算法具有更快的收敛速度,仿真实验表明了其优越性。

关键词 数据聚类,高维数据,随机过程,马尔可夫链
中图法分类号 TP393 **文献标识码** A

Research on a Clustering Algorithm Based on Generalized Quantum Particle Model and its Convergence

HUANG Liang-jun[1] SHUAI Dian-xun[1,2] ZHANG Bin[1]
(Department of Computer Science and Technology, East China University of Science and Technology, Shanghai 200237, China)[1]
(State Key Laboratory of Intelligence Technology and System, Tsinghua University, Beijing 100084, China)[2]

Abstract A novel generalized quantum particle model (GQPM) was presented for data self-organizing clustering. In this model the data clustering process is transformed into a stochastic self-organizing process of the quantum particles in the state configuration space. The state configuration will evolve to a stationary probability distribution, and thus the optimal state configuration on particles can be obtained from the state configuration which has the highest probability in the stationary probability distribution. The convergence of the self-organizing process was proved in this paper. The GQPM algorithm has much faster clustering speed than the traditional clustering algorithm for the large scale database. Its superiorities were verified by the simulation experiments.

Keywords Data clustering, Multidimensional data, Stochastic process, Markov chain

1 引言

"量子纠缠"的特性在将来是人类可以利用的重要资源。本文结合"量子纠缠"适合群体聚类的特征提出了一种新的数据自组织聚类方法。所提出的广义量子粒子模型利用了物理学中"量子纠缠"的以下特性[1]:其一,当两个或多个粒子发生"量子纠缠"后,永远彼此关联,其中一个粒子发生任何状况,另一个粒子必定同时发生相应改变;其二,当两个或多个粒子发生"量子纠缠"后,构成一个复合的量子纠缠系统,其状态是一个精确已知的纯态,量子熵为0,而其子系统和其中的个体却处于混合态,量子熵大于0。当一个粒子与其它粒子发生纠缠后,其状态由纯态变成了混合态,量子熵由0变成了正数。本文提出的广义量子粒子模型构造了携带数据对象并带有量子状态的粒子,将自组织聚类过程转化为一个粒子在状态构形空间上的随机自组织过程,量子熵的大小决定了状态构形的概率分布。当任意两个粒子进行碰撞时,粒子根据规则由粒子携带的数据对象和粒子本身的量子状态决定两个粒子之间是否纠缠,从而在空间里可以形成多种多样的量子纠缠分布,称为粒子的状态构形,而粒子在空间中的几何分布称为粒子的几何构形。那些携带属于同一数据簇的数据对象的粒子通过合成量子纠缠状态而构成纠缠等价类。随着过程的不断进行,粒子的状态构形将随时间的流逝而不断演化,最终收敛到一个平稳的概率分布。平稳概率分布中具有最大概率的状态构形与最优状态空间构形相对应。与传统的适用于大规模数据聚类的方法相比较,本文提出的算法具有更快的聚类速度和更好的聚类性能,仿真试验说明了算法的优越性。

2 广义量子粒子模型及理论

本文提出的广义量子粒子模型是在广义细胞自动机[2,3]和广义粒子模型[4]的基础上进行改进而得来的。广义细胞自动机借鉴了细胞神经网络、细胞自动机以及蚁群算法,在传统细胞自动机的基础上提出了细胞核、细胞状态、细胞调和函数等概念,制定细胞的演化规则及广义细胞自动机并行算法,使细胞阵列上各细胞按演化规则最终达到平稳概率分布。广义

到稿日期:2009-03-31 返修日期:2009-06-04 本文受国家自然科学基金(60575040,60473044)资助。
黄良军 男,博士生,工程师,主要研究方向为人工智能、分布并行计算,E-mail:ly000393@online.sh.cn;**帅典勋** 教授,博士生导师,主要研究方向为人工智能、知识工程、分布并行计算等;**张 彬** 男,硕士,主要研究方向为人工智能、分布并行计算。

粒子模型是在广义细胞自动机的基础上发展出来的，它与广义细胞自动机的主要区别在于广义粒子模型阵列上的每个粒子均可自由移动，而广义细胞自动机上的每个细胞的位置固定不变，其细胞状态却发生了变化。在广义细胞自动机的基础上，广义粒子模型改进了粒子调和函数、转移概率、演变规则等，进一步完善了模型。本文提出的广义量子粒子模型则将“量子纠缠”的特性引入到了模型中，应用此模型的自组织聚类与之前的广义细胞自动机和广义粒子模型的同类应用又有很大的不同，这些不同之处体现在粒子的构成要素、聚类形式以及并行算法等方面。

广义量子粒子模型使用二维 $N\times N$ 阵列作为粒子的移动空间，$N\times N$ 的数目要大于阵列中的信息数据对象，即样本的数量 n。广义量子粒子模型阵列上的每个粒子都由两个部分组成：粒子所携带的给定数据集中的某个数据对象以及代表量子状态的量子位。所有这些粒子都可以随机地分布在阵列上。当任意两个粒子进行碰撞时，粒子根据规则由粒子携带的数据对象和粒子本身的量子状态决定两粒子之间是否发生“量子纠缠”。

在信息自组织过程中，广义细胞自动机和广义粒子模型的数据对象都会根据各自携带的不同信息分布到不同的聚类区域中，最终每个聚类区域都是由具有相同数据对象的粒子组成的，并且散布在阵列上的每个聚类区域都有着最小的区域周长。但广义量子模型却不是这样的，最终将粒子分配到各个不同的纠缠类中，每一个纠缠类中的粒子在几何位置上并不要求有最小的区域周长，但都具有一个量子纠缠态。并且每一个纠缠类中的粒子都携带相似的数据对象，所得到的纠缠分布可以通过量子测定获得。因此在广义量子粒子模型中，就状态构形而言，同一个纠缠类中的粒子共同构成一个纠缠等价类；但是就几何构形而言，同一个状态构形可以具有许多个不同的几何构形。正是由于“量子纠缠”特性的引入，使得广义量子粒子模型聚类算法的收敛速度与非量子化的广义粒子模型的相比有了进一步的提高。

在广义量子粒子模型中，粒子 C_i 的量子状态表示为 $|\psi_i\rangle$，其值初始设定为：

$$|\psi_i\rangle=q_0|0\rangle+q_1|1\rangle \tag{1}$$

其中，$|0\rangle$和$|1\rangle$是一个量子位(qubit)的标准正交计算基态；q_0 和 q_1 均为复数，并且满足 $|q_0|^2+|q_1|^2=1$。$|\psi_i\rangle$是一个叠加态，当测量量子位$|\psi_i\rangle$时，得到$|0\rangle$的概率是$|q_0|^2$，而得到$|1\rangle$的概率是$|q_1|^2$。

m 个粒子 $C_{i_1},\cdots,C_{i_m}$ 的 m 个量子位的纠缠态表示为：

$$|\psi_{i_1\cdots i_m}\rangle=p_0|0\cdots0\rangle_m+p_1|1\cdots1\rangle_m \tag{2}$$

其中，$|p_0|^2+|p_1|^2=1$，并且$|0\cdots0\rangle_m$和$|1\cdots1\rangle_m$分别代表 m 个状态$|0\rangle$和 m 个状态$|1\rangle$的张量积。在不引起歧义的前提下，为方便起见，以后出现的$|0\cdots0\rangle_m$和$|1\cdots1\rangle_m$中的下标 m 将被省略。

式(1)和式(2)中的状态也可以分别使用密度算子(density operator)的形式来表示：

$$\rho_i=|\psi_i\rangle\langle\psi_i|=(q_0|0\rangle+q_1|1\rangle)(\langle0|q_0^*+\langle1|q_1^*) \tag{3}$$

$$\rho_{i_1\cdots i_m}=|\psi_{i_1\cdots i_m}\rangle\langle\psi_{i_1\cdots i_m}| =(p_0|0\cdots0\rangle+p_1|1\cdots1\rangle)(\langle0\cdots0|p_0^*+\langle1\cdots1|p_1^*) \tag{4}$$

其中，“$*$”表示复共轭。

若对于由子系统 A 和 B 组成的复合系统 AB，我们给出其密度算子 ρ_{AB}，那么其子系统 A 的约化密度算子(reduced density operator)ρ^A 定义为：

$$\rho^A=tr_B(\rho_{AB}) \tag{5}$$

其中，$tr_B(|a_1b_1\rangle\langle a_2b_2|)\equiv_{df}|a_1\rangle\langle a_2|tr(|b_1\rangle\langle b_2|)$，$tr_B$ 是一个算子的映射，称为在系统 B 上的偏迹。式中$|a_1\rangle$和$|a_2\rangle$是子系统 A 中的两个量子状态，$|b_1\rangle$和$|b_2\rangle$是子系统 B 中的两个量子状态。等式右边的迹运算是子系统 B 上的普通迹运算，$tr(|b_1\rangle\langle b_2|)$的值是方阵$|b_1\rangle\langle b_2|$的对角线元素之和。

由密度算子 ρ 表示的量子状态的量子熵定义为：

$$S(\rho)=-tr(\rho\log_2\rho)=-\sum_i\lambda_i\log_2\lambda_i \tag{6}$$

其中，λ_i 是密度算子 ρ 的特征值。

假设由 m 个粒子 $C_{i_1},\cdots,C_{i_m}$ 组成的系统具有如下的纠缠量子状态($p_0=p_1=1/\sqrt{2}$)：

$$\rho_{i_1\cdots i_m}=\left(\frac{|0\cdots0\rangle+|1\cdots1\rangle}{\sqrt{2}}\right)\left(\frac{\langle0\cdots0|+\langle1\cdots1|}{\sqrt{2}}\right) \tag{7}$$

根据式(5)，对于任何 $i\in\{i_1,\cdots,i_m\}$，可以求得：

$$\rho^i=tr_{(i_1\cdots i_m-i)}(\rho_{i_1\cdots i_m})=|\varphi_i\rangle\langle\varphi_i|tr(\rho_{i_1\cdots i_m-i})=\frac{|0\rangle\langle0|+|1\rangle\langle1|}{2} \tag{8}$$

其中，下标$(i_1\cdots i_m-i)$代表了把粒子 C_i 从整个粒子集$\{C_{i_1},\cdots,C_{i_m}\}$除去后的子集。

量子系统具有以下特征[1]：纯态满足 $tr(\rho^2)=1$，而混合态满足 $tr(\rho^2)<1$。对于式(7)和式(8)有：$tr(\rho_{i_1\cdots i_m}{}^2)=1$，$tr((\rho^i)^2)=1/2$。可见粒子 C_i 的表现行为相似于各以 1/2 概率出现$|0\rangle$和$|1\rangle$两种状态的混合态。同时可以得出由粒子集$\{C_{i_1},\cdots,C_{i_m}\}$组成的纠缠量子系统的任一子系统均处于一个混合态。

对于式(7)所示的纠缠量子状态，由于 $\rho_{i_1\cdots i_m}\left(\frac{|0\cdots0\rangle+|1\cdots1\rangle}{\sqrt{2}}\right)=\left(\frac{|0\cdots0\rangle+|1\cdots1\rangle}{\sqrt{2}}\right)$恒成立，可以得出$\left(\frac{|0\cdots0\rangle+|1\cdots1\rangle}{\sqrt{2}}\right)$为 $\rho_{i_1\cdots i_m}$ 的特征向量，其对应特征值 $\lambda=1$。密度算子 ρ(亦称作密度矩阵)具有以下特征：满足 $tr(\rho)=1$。另矩阵具有以下性质：方阵 A 的特征值的和等于方阵 A 的迹 $tr(A)$。于是可以得出：纠缠量子系统的密度算子 $\rho_{i_1\cdots i_m}$ 具有唯一的特征值 $\lambda=1$。那么由式(6)可以得出 $S(\rho_{i_1\cdots i_m})=0$，也就是说由粒子$\{C_{i_1},\cdots,C_{i_m}\}$组成的纠缠量子系统的量子熵为零。

同时，由式(8)可以得到：

$$\rho^i|0\rangle=\left(\frac{|0\rangle\langle0|+|1\rangle\langle1|}{2}\right)|0\rangle=\frac{|0\rangle}{2}$$

$$\rho^i|1\rangle=\left(\frac{|0\rangle\langle0|+|1\rangle\langle1|}{2}\right)|1\rangle=\frac{|1\rangle}{2}$$

可以得出 ρ^i 拥有对应于特征值 1/2 的两个特征向量$|0\rangle$和$|1\rangle$。由式(6)可知 $S(\rho^i)=1$，也就是说粒子 C_i 的量子熵为 1。

对于式(4)表示的一般情况下的量子纠缠状态以及 $i\in\{i_1,\cdots,i_m\}$，同样可以得出：

$$\rho^i=|p_0|^2|0\rangle\langle0|+|p_1|^2|1\rangle\langle1| \tag{9}$$

$$S(\rho_{(i_1\cdots i_m)})=0$$

且

$$S(\rho^i)=-\sum_{j=0,1}|p_j|^2\log_2|p_j|^2>0 \tag{10}$$

这表明由粒子$\{C_{i_1},\cdots,C_{i_m}\}$组成的纠缠量子系统的量子

熵小于其任意一个子系统的熵。由此可知粒子 C_i 的条件熵 $S(\rho_{i_1\cdots i_m-i}|\rho^i)=S(\rho_{i_1\cdots i_m})-S(\rho^i)<0$。这是"量子纠缠现象"的特点之一：一个系统的联合状态是精确已知的纯态，而子系统却处于混合态，即联合系统的相关熵(joint entropy)小于它的子系统或个体的熵，而且子系统的条件熵(conditional entropy)是负的。Shannon 信息论的观点与上述观点正好相反。在 Shannon 信息论中，一个系统的相关熵不应小于其子系统的熵，并且条件熵也不应为负。由于量子纠缠现象的这个特性，使得量子计算的方法比非量子的方法更适用于基于群体智能的数据自组织聚类。当量子粒子 C_i 未与其它粒子发生纠缠时，其量子状态为纯态，量子熵 $S(\rho^i)=0$；当量子粒子 C_i 与其它粒子发生纠缠后，其量子状态为混合态，量子熵 $S(\rho^i)>0$。后面的并行算法便是基于量子熵来计算演化规则中的调和函数和转移概率的，量子熵的大小决定了状态构形的概率分布。

3 用于聚类的广义量子粒子模型并行算法

为了模拟量子粒子在一个广义量子粒子模型阵列上的演化，使用两个完全相同的广义量子粒子模型阵列：粒子阵列 I 和粒子阵列 II 来存贮阵列上随机产生的两个不同的几何构形。在时间段 τ 内给定一个使用 GQPM 方法进行聚类的数据集，根据式(2)和式(4)来选择参数 p_0 和 p_1 的值，且 $p_0, p_1 \neq 0$。算法描述如下：

并行步 1　在 t_0 时刻，将粒子随机分布在阵列 I 和阵列 II 上，产生两个不同的几何构形，每一个粒子都携带一个数据对象并且量子状态初始化为 $|\psi_i\rangle=|0\rangle$。

并行步 2　在 t 时刻，根据下述定义计算每一个粒子的量子状态，同时获得一个在粒子阵列 I 和阵列 II 上的量子粒子的纠缠等价划分。

两粒子互相碰撞，当且仅当一个粒子移动到了另一个粒子的邻域中。一旦两粒子 C_i 和 C_j 碰撞并且携带着相似的数据对象，那么此后它们就发生量子纠缠并且构成量子纠缠状态 $|\psi_{ij}\rangle=p_0|00\rangle+p_1|11\rangle$。同时，为了描述方便又不失准确性，假设无论一个单独的粒子是否与其他粒子发生纠缠，这个粒子总是自我纠缠的。由此将一个在粒子集合上的纠缠关系定义为一个对称的、传递的且自反的二元关系，即一个等价关系。基于粒子的一个量子纠缠划分意味着粒子被划分到了基于纠缠关系的等价类中。换句话说，在一个等价类中的所有粒子都是相互纠缠着的。

并行步 3　在粒子阵列 I 和阵列 II 上进行量子测量并寻找到纠缠的粒子，然后根据下述定义计算粒子阵列 I 和阵列 II 上每一个纠缠等价类的量子熵。

对于一个在所有粒子上的纠缠划分 Ω，Ω 中某个等价类 U 的调和函数 $h(U)$ 被定义为等价类 U 中所有粒子的量子熵和的函数，即：

$$h(U)=\left[\sum_{C_i\in U}S(\rho^i)\right]^2 \tag{11}$$

并行步 4　根据下述定义计算在 t 时刻粒子阵列 I 和阵列 II 上的聚合调和函数。

一个纠缠划分 Ω 的聚合调和函数定义为：

$$H(\Omega)=\sum_{U\in\Omega}w(U)h(U) \tag{12}$$

其中，$w(U)$ 是等价类 U 的权重系数。

并行步 5　根据下述定义计算转移概率，然后从粒子阵列 I 和阵列 II 中根据转移概率随机选择一个作为下一时刻 $t+1$ 的处理对象。

广义量子粒子模型阵列上的所有粒子都应该是随机且并发地移动和碰撞的，所以纠缠划分 Ω_1 转变为纠缠划分 Ω_2 的概率为：

$$p(\Omega_1\Rightarrow\Omega_2)=f(\Delta H)=\frac{1}{(1+e^{\frac{\Delta H}{T}})} \tag{13}$$

其中，$\Delta H=H(\Omega_1)-H(\Omega_2)$，而其中的热力学温度因子 T 用于控制收敛速度以及提高自组织聚类的质量。

并行步 6　对于在并行步 5 中没有被选中的粒子阵列 I 或粒子阵列 II，将粒子随机分布其上，产生一个新的几何构形，并按照并行步 2 中的定义计算新产生的阵列中每一个粒子的量子状态。

并行步 7　重复执行并行步 3 到并行步 6，直到每一种纠缠划分出现的概率不再变化。

注释：

1) 粒子阵列 I 和阵列 II 上的所有粒子都是并行执行的。广义量子粒子模型聚类算法具有超大规模的并行性。可以看到使用式(13)计算转移概率不会影响到个体粒子的独立演化。

2) 文后证明了执行广义量子粒子模型算法将得到一个在纠缠划分空间内具有最大 Shannon 熵的稳定概率分布，即在广义量子粒子模型阵列上的任意两个纠缠划分间的转移概率是一个常数。

3) 文后证明了与最优聚类相对应的是在纠缠划分空间内的稳定概率分布中具有最大概率的纠缠划分。

4 广义量子粒子模型聚类算法的收敛性

引理 1　执行广义量子粒子模型算法得到在纠缠划分空间内的一个有限齐次马尔可夫随机过程 $\{\Omega(t), t=0,1,\cdots\}$，其中 $\Omega(t)$ 表示时刻 t 时所有粒子的纠缠划分。

证明：在有限粒子集上可能的纠缠划分的数量是有限的。根据式(13)，从 $\Omega(t)$ 演化为 $\Omega(t+1)$ 的转移概率仅依赖于 $\Delta H=H(\Omega(t))-H(\Omega(t+1))$，并且与 t 之前的纠缠划分 $\Omega(t-i)$，$i\geqslant 1$ 以及当前时间 t 都无关。

引理 2　执行广义量子粒子模型算法生成的马尔可夫随机过程 $\{\Omega(t), t=0,1,\cdots\}$ 是不可约齐次的，并且任何纠缠划分都会在有限时间段内以概率 1 转变回它自身。

证明：根据式(13)，从一个纠缠划分转变到另一个纠缠划分的转移概率恒大于零，所以任意两个纠缠划分是互为可达的。因此纠缠划分空间中的唯一闭集就是所有可能的纠缠划分组成的集合，这意味着它是不可约分的。由引理 2 知，有限的且互为可达的马尔可夫链的所有纠缠划分是正常返的。

引理 3　执行广义量子粒子模型算法生成的马尔可夫随机过程 $\{\Omega(t), t=0,1,\cdots\}$ 在纠缠划分空间内一定会达到一个稳定的概率分布，即任何纠缠划分最终会具有一个固定概率。此外，稳定分布与初始的纠缠划分无关。

证明：根据马尔可夫过程理论，如果纠缠划分空间内的任意两个随机状态可以以非零的概率一步地从一个转变到另一个，则随机过程必收敛于稳定状态。因为任意两个纠缠划分都具有一个有限的聚合调和函数且转移概率大于零，所以随机纠缠划分序列 $\{\Omega(t), t=0,1,\cdots\}$ 最终必然达到稳定的概率

分布。

定理 1 执行广义量子粒子模型算法可以得到一个具有最大 Shannon 熵的随机概率分布，并且具有最大概率的纠缠划分的聚合量子调和函数也最大。

证明：在纠缠划分空间内的一个稳定概率分布下，假设纠缠划分 Ω 以概率 $p(\Omega)$ 出现，则对 Shannon 熵求最大值如下：

$$\max_{p(\Omega)}\{-\sum_{\Omega}p(\Omega)\ln p(\Omega)\} \tag{14}$$

服从约束条件：

$$\begin{cases}\sum_{\Omega}p(\Omega)=1\\ \sum_{\Omega}p(\Omega)H(\Omega)=\overline{H}(\Omega)\end{cases} \tag{15}$$

其中，$H(\Omega)$是纠缠划分 Ω 的聚合调和函数，$\overline{H}(\Omega)$是一个与 Ω 对应的先验概率相关联的统计平均值。通过

$$\frac{\partial}{\partial p(\Omega)}\{\sum_{\Omega}p(\Omega)\ln p(\Omega)-\xi_1[\sum_{\Omega}p(\Omega)H(\Omega)-\overline{H}(\Omega)]-\xi_2[\sum_{U\in\Omega}p(\Omega)-1]\}=0 \tag{16}$$

可以得出

$$p^*(\Omega)=Z^{-1}\exp[H(\Omega)]=Z^{-1}\exp\{\sum_{U\in\Omega}h(U)\}=Z^{-1}\exp\{\sum_{U\in\Omega}\sum_{C_i\in U}S(\rho^i)^2\} \tag{17}$$

其中，$Z=\sum_{\Omega}\exp[H(\Omega)]$；$\xi_1$ 和 ξ_2 是拉格朗日乘子。

另一方面，当执行广义量子粒子模型算法达到稳定概率分布时，将对应于纠缠划分 Ω 的平稳概率表示为 $\gamma^*(\Omega)$，则：

$$p(\Omega_1\Rightarrow\Omega_2)\gamma^*(\Omega_1)=p(\Omega_2\Rightarrow\Omega_1)\gamma^*(\Omega_2)$$

同时，$H(\Omega_1)-H(\Omega_2)=-(H(\Omega_2)-H(\Omega_1))$

可以得出：

$$\gamma^*(\Omega_1)/\gamma^*(\Omega_2)=\exp[H(\Omega_1)/T]\exp[H(\Omega_2)/T]$$

$$\gamma^*(\Omega)=\sum_{\Omega}\exp[H(\Omega)/T]\exp[H(\Omega)/T] \tag{18}$$

因此，不考虑热力学温度 T 时，$\gamma^*(\Omega)=p^*(\Omega)$，其中的热力学温度 T 是为了改善模型性能而引入的参数。这就证明了执行广义量子粒子模型算法得到的随机过程的稳定概率分布正好是最大 Shannon 熵分布。同时可以得出具有最大调和函数的纠缠划分在所获得的稳定概率分布上具有最大概率。

定理 2 执行广义量子粒子模型算法达到平稳概率分布后，具有最大概率的纠缠划分中的每个纠缠类必须包括给定数据集中所有携带相似数据对象的量子粒子。

证明：将平稳概率分布上的具有最大概率的纠缠划分表示为 Ω^*。使用反证法：假设对于纠缠划分 Ω^* 中两个不同的纠缠类 U 和 U'，与纠缠类 U 相似的数据对象被划分到了纠缠类 U' 中。因为

$$h(U)=[\sum_{C_i\in U}S(\rho^i)]^2,h(U')=\left[\sum_{C_i\in U'}S(\rho^i)\right]^2$$

$$h(U\cup U')=[\sum_{C_k\in U\cup U'}S(\rho^k)]^2>h(U)+h(U')$$

可知 $H(\Omega^*)$并不是最大值，所以根据式(17)可得 Ω^* 并不具有最大概率。这样就与已知条件矛盾，证明成立。

5 仿真实验结果与分析

实际执行广义量子粒子聚类算法时，可以采用两种方案：同步方案和异步方案。对于异步方案，并行步 6 中新阵列上粒子的几何构形是完全随机产生的，而粒子的状态构形可以继承前次演化后的纠缠状态，这样可以大大减少需要搜索的状态构形。不过需要增加一个辅助步骤，将经过量子测量后塌陷为基态的量子粒子重新构造成纠缠态，再随机分布到新阵列上。本文中的仿真实验采用的是异步方案。

由于本文所研究的是适用于大规模数据聚类的算法，因此这里选择对于大规模数据支持较好的 CLARANS 方法(Clustering Large Applications based on RANdomized Search)和 PAM 方法(Partitioning Around Medoids)来做比较。图 1 是 3 种方法之间完成聚类所需演化次数的比较，数据数量分别为 2000，4000 和 6000，数据簇数量为 100。图中的纵坐标轴采用了对数刻度。由图 1 可以看出，GQPM 收敛时间要远远小于 CLARANS 和 PAM 所需要的收敛时间，这主要是因为广义量子粒子聚类算法的并行性。

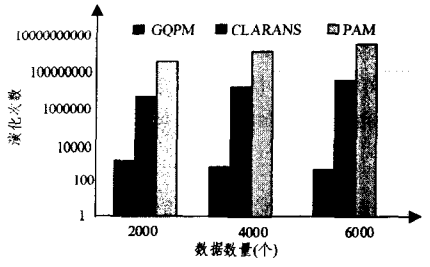

图 1 GQPM、CLARANS 和 PAM 完成聚类所需演化次数的比较

图 2 显示了阵列规模为 100×100、聚类簇数量为 20，50，80，125 时样本数从 2000 到 9000 时完成聚类所需的时间。可以看到随着数据数量的增大，完成聚类所需的演化次数随之减小。当聚类簇的数量相同时，每个聚类簇中的数据对象的数量越多，平均每个粒子与同类粒子相互碰撞到一起需要的演化次数就越少，从而使得聚类收敛所需要的演化次数减少。

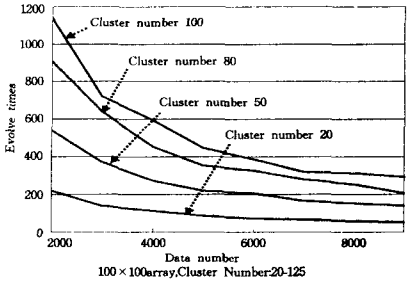

图 2 广义量子聚类模型中不同数量的聚类簇完成聚类的演化次数

图 3 显示了阵列规模为 100×100、数据对象的数量分别为 2000，4000，6000，8000 时，聚类簇数量从 20 到 125 时完成聚类所需的演化次数。可以看到随着聚类簇数量的增大，达到收敛所需的演化次数也随之增大。当数据对象的总数量相同时，聚类簇的数量越大，则每个聚类簇中的数据对象的数量就越少，也就意味着平均每个粒子与同类粒子相互碰撞到一起需要的演化次数就增加，从而使得聚类收敛所需要的演化次数增加。

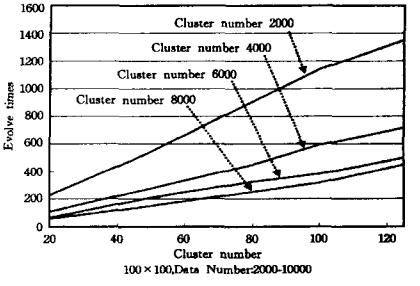

图 3 广义量子聚类模型中不同样本数的聚类完成时间

(下转第 236 页)

模型和CCM模型的输入变量与独立测试实验相同，然后利用CAGC识别测试样本。重复该过程，直到每一个样本作为测试样本时为止。统计所有被正确识别的样本，并计算性能评价指标 *Accu*。上述分类实验重复10次，计算平均性能评价指标。在FFCV中，将数据集平均分成5部分，每次挑选不同的一部分作为测试样本，其余样本作为训练数据集训练GCM模型和CCM模型，然后利用CAGC识别测试样本。重复分类过程5次，直到每部分样本作为测试样本时为止。统计所有被正确识别的样本，并计算性能评价指标 *Accu*，*Prec*，*Sn* 和 *Sp*。上述分类实验重复10次，计算平均性能评价指标。并与加权投票法、SVM和KNN进行比较，其中加权投票法、SVM和KNN的分类参数设置与独立测试实验相同。分类实验同样重复10次，并计算平均性能评价指标。表5给出了交叉测试实验的分类准确度 *Accu*。从表5可以看出在LOOCV中，相对于加权投票法和KNN，SVM同样取得了较好的分类准确度，并且在DLBCL上取得了最好的分类准确度(97.8%)，高于CAGC。然而在Prostate上，*Accu* 低于其它分类器。在所有数据集中，KNN则相对表现了较好的泛化能力。CAGC除了在Prostate上分类准确度略低于SVM外，在其它数据集上都高于加权投票法、KNN和SVM。在FFCV中，SVM同样取得了较好的分类准确度，加权投票法次之。CAGC除了在DLBCL上分类准确度略低于SVM外，在其它数据集上都高于加权投票法、KNN和SVM。

表5 交叉测试实验结果(Accu%)

		ALL-AML Leukemia	Breast Cancer	Prostate Cancer	DLBCL	Colon Tumor	Ovarian Cancer
LOOCV	Weighted Voting	90.3	77.3	70.4	88.6	93.5	63.5
	SVM	95.3	84.6	68.9	97.8	91.9	82.4
	KNN(K=5)	86.1	84.2	80.1	92.8	83.9	73.3
	CAGC	99.3	97.9	91.4	93.4	96.8	92.6
FFCV	Weighted Voting	88.4	72.1	82.9	80.4	90.6	71.4
	SVM	92.1	94.4	80.5	96.3	93.2	84.4
	KNN(K=5)	84.1	80.4	80.6	86.3	84.2	67.9
	CAGC	96.2	93.6	94.3	95.5	97.3	96.3

结束语 针对基因表达谱数据特点提出了两种癌症识别模型(GCM模型和CCM模型)，利用基于权值的投票组合策略提出了一种协同分类方法(CAGC)。在6个数据集上分别进行了交叉测试实验。CAGC有效综合了GCM和CCM识别模型的解决方案，在所有数据集上都取得了很好的分类性能。

参考文献

[1] Kuramochi M, Karypis G. Gene classification using expression profiles: a feasibility study[J]. International Journal on Artificial Intelligence Tools, 2005, 14(4): 641-660

[2] 李霞，张田文，郭政. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报，2004，27(5)：675-682

[3] Lu X G, Lin Y P, Wang H J, et al. A Novel Relative Space Based Gene Feature Extraction and Cancer Recognition[C]// Proc. of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), LNAI 4426. 2007: 712-719

[4] 李颖新，李建更，阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报，2006，26(2)：324-330

[5] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nature Medicine, 2002, 8(1): 68-74

[6] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531-537

[7] Veer L J V T, Dai H, Vijver M J V D, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530-536

[8] Cho S B, Won H H. Machine Learning in DNA Microarray Analysis for Cancer Classification[C]// Proc. of Bioinformatics 2003 First Asia-Pacific Bioinformatics Conference (APBC 2003). 2003: 189-198

[9] 王正群，陈世福，陈兆乾. 优化分类型神经网络线性集成[J]. 软件学报，2005，16(11)：1902-1908

[10] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification[J]. Applied Bioinfomatics, 2003, 2(3 Suppl): S75-S83

[11] Khan J, Wei J S, Ringne M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine, 2001, 7(6): 673-679

[12] O'Neill M C, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect[J]. BMC Bioinformatics, 2003, 4: 13

[13] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning[C]// Proceedings of the 13th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 1993: 1022-1029

[14] Liu B, Cui Q H, Jiang T Z, et al. A combinational feature selection and ensemble neural network method for classification of gene expression data[J]. BMC Bioinformatics, 2004, 5: 136

(上接第228页)

结束语 本文讨论了广义量子粒子模型的理论与性能，并设计了切实可行的将广义量子粒子模型应用于数据自组织聚类的并行算法。所提出的算法收敛速度快，并且在对噪声的不敏感性、聚类数据的强鲁棒性等方面都具有优势，仿真试验表明了所提算法的优越性。接下来要进一步充分挖掘和利用广义量子粒子模型的并行机制，研究其中参数设置的合理性，使得量子计算的优越性能够得到充分的利用。

参考文献

[1] Nielsen M A, Chuang I L. Quantum Computation and Quantum Information[M]. London: Cambridge University Press, 2000

[2] 帅典勋，冯翔，赵宏彬，等. 广义细胞自动机的结构及其硬件实现[J]. 计算机学报，2004，27(11)：1441-1450

[3] 刘燕，帅典勋，柴震川. 基于群体智能的网络信息自组织方式[J]. 计算机科学，2003，30(6)：84-87，125

[4] Shuai D X, Xue F L. Generalized particle model used for data clustering[J]. International Journal of Pattern Recognition, 2006, 20(7): 1001-1028

[5] Ramos R V, Souza R F. Calculation of the quantum entanglement measure of bipartite states, based on relative entropy, using genetic algorithms[J]. Journal of Computational Physics, 2002, 175(2): 576

[6] Grover L K. Quantum mechanics helps in searching for a needle in a haystack[J]. Physical Review Letters, 1997, 79(2): 325

[7] Feng X, Lau F C M, Shuai D X. The coordination generalized particle model-An envolutionary approach to multi-sensor fusion [J]. Information fusion, 2008, 9(4): 450-464

[8] Aczel A D. Entanglement: the greastest Mystery in physics[M]. New York: John Wiley and Sons Ltd, 2002