基于量子谐振子模型的聚类中心选取算法

燕京京[1,3],王 鹏[2],范家兵[1,3],黄 焱[1,3]
(1. 中国科学院成都计算机应用研究所,四川成都 610041; 2. 成都信息工程学院并行计算实验室,四川成都 610225;
3. 中国科学院大学,北京 100049)

摘 要: 提出了一种基于量子谐振子模型的聚类中心选取算法. 该算法以量子谐振子波函数从高能态到基态过程中的概率变化过程为理论模型来描述聚类问题中数据对象向聚类中心点的聚集行为,能够快速查找到最优的聚类个数及较好的聚类中心点所在的网格;数据读入网格结构之后,算法的处理时间与数据集规模无关. 实验结果表明:CCSA-QHOM 算法较适合于处理每个子类局部区域的网格密度分布呈单峰特性的数据集的聚类中心选择问题.

关键词: 聚类中心; 量子谐振子; 聚类个数; 网格; 单峰特性
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112 (2016)02-0405-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.02.023

Clustering Center Selecting Algorithm Based on Quantum Harmonic Oscillator Model

YAN Jing-jing[1,3], WANG Peng[2], FAN Jia-bing[1,3], HUANG Yan[1,3]
(1. *Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, Sichuan* 610041, *China*;
2. *Parallel Computing Laboratory, Chengdu University of Information Technology, Chengdu, Sichuan* 610225, *China*;
3. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: This article puts forward a clustering center selecting algorithm based on quantum harmonic oscillator model (CCSA-QHOM). The algorithm describes the way of data objects finding center of the cluster in clustering problem by taking the change of wave function's probability in the process of high energy level to a lower energy level for theoretical model. It can quickly find the optimal number of clusters and cluster center, computing time has nothing to do with the size of the data set after the dataset being got in grid space. Experiments show that CCSA-QHOM is more suitable for processing the clustering center selection question of dataset in which grid density distribution of each subclass haves a single peak characteristic.

Key words: cluster center; quantum harmonic oscillator; number of clusters; grid; peak characteristic

1 引言

聚类分析也称为"无师可学"的分类[1],是根据物理或抽象对象之间的相似性将它们聚为若干类的过程. 属于同一类的对象之间的相似性高,不属于同一类的对象之间的相异性高[2,3]. 聚类分析发展至今,产生的聚类方法主要可以分为以下几类:基于划分的方法,基于密度的方法,基于网格的方法,基于层次的方法,基于模型的方法及基于约束的方法[2]. 其中,基于划分的方法以聚类中心点来表征数据集的位置,如 K-means 聚类算法. K-means 聚类算法以其简单,有效性广泛地应用于工程研究之中. 但是,K-means 聚类算法从产生之时就面临着一个重要问题:初始聚类中心点的选择不当会使算法过早的收敛于局部最优解[4]. 此后,很多学者对 K-means 聚类算法的初始中心点的选择方法进行了研究. 到目前为止,初始中心选择方法大概可以分为三种[5]:随机选择法,距离优化法,密度估计法. 这三种方法中的典型代表分别是传统 K-means 算法[6],简易中心搜索法[7],KR 方法[8]. 当前,还有一些学者将研究方向转向智能算法来解决初始中心点随机选择导致的局部极值问题,如将粒子群算法[9~11]等优化算法运用于该问题的研究. 这些算法虽然能够在很大程度上降低

收稿日期:2014-07-01;修回日期:2014-12-06;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 60702075);广东省科技厅高新技术产业化科技攻关项目(No. 2011B010200007);四川省青年科学基金(No. 09ZQ026-068);成都市科技局创新发展战略研究项目(No. 11RXYB016ZF)

结果陷入局部最优的概率,但是不能自动发现聚类的数目.同粒子群算法相比,量子谐振子算法[12]不仅具有全局寻优能力,通过参数的合理调整,其还有较强的局部寻优能力.

本文基于文献[12]提出了一种基于量子谐振子模型的聚类中心选取算法(Clustering Center Selecting Algorithm Based on Quantum Harmonic Oscillator Model,CCSA-QHOM),来解决初始中心点选择不当导致局部极值及在运算过程中不能自适应的确定聚类个数的问题.该算法以量子谐振子从高能态到低能态跃迁过程中的波函数概率收敛过程为理论模型来描述聚类问题中数据对象向聚类中心点的聚集行为.CCSA-QHOM 能够快速的定位到聚类中心的位置,并在合理的参数设置下自动得到实际的聚类个数.实验阶段通过七组实验来说明 CCSA-QHOM 算法的适用性和有效性.

2 量子谐振子模型

量子力学是现代物理学的理论基础,而量子谐振子是其中重要的模型系统之一.量子谐振子波函数可以采用高斯分布函数来解释从高能态向基态的收敛过程.在一维谐振子中,粒子的理想运动模型是指一个质量为 m 的粒子被置于势能无穷大的阱中,沿着某一方向(如 x 轴方向)运动,以平衡位置为坐标原点,则一维谐振子的势能表示为[13]:

$$V(x)=\frac{1}{2}m\omega^2x^2 \tag{1}$$

利用薛定谔方程来求一维谐振子的波函数,可得:

$$\psi_n(x)=\sqrt{\frac{1}{2^n n!}}\left(\frac{m\omega}{\pi\hbar}\right)^{\frac{1}{4}}\cdot\exp\left(-\frac{m\omega x^2}{2\hbar}\right)\cdot H_n\left(\sqrt{\frac{m\omega}{\hbar}}x\right) \tag{2}$$

波函数的概率密度 $|\psi_n(x)|^2$ 表示为:

$$|\psi_n(x)|^2=\frac{1}{2^n n!}\left(\frac{m\omega}{\pi\hbar}\right)^{\frac{1}{2}}\cdot\exp\left(-\frac{m\omega x^2}{\hbar}\right)\cdot\left|H_n\left(\sqrt{\frac{m\omega}{\hbar}}x\right)\right|^2 \tag{3}$$

量子谐振子模型中波函数的概率密度函数与能级之间的关系如图 1.

从图 1 中可以看出,随着量子谐振子波函数从高能态到基态,其概率密度函数从多个高斯函数叠加逐渐收敛到一个高斯函数.在基态时的概率密度函数为:

$$|\psi_0|^2=\left(\frac{m\omega}{\pi\hbar}\right)^{\frac{1}{2}}\cdot\exp\left(-\frac{m\omega x^2}{\hbar}\right) \tag{4}$$

量子谐振子物理模型是根据图 1 中的概率变化规律来建立的模型.其主要思想是:开始时量子谐振子处于运动活跃的高能态.此时的波函数概率密度函数是多个高斯函数的叠加.随着能级的降低,概率密度函数

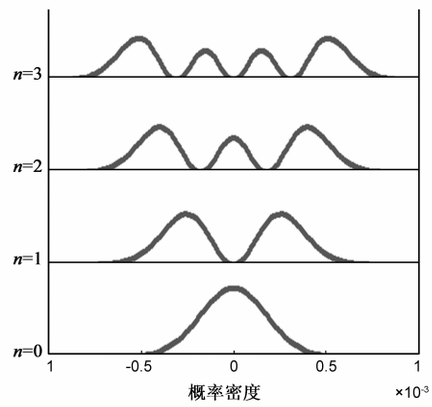


图1 波函数的概率密度函数与能级之间的关系

的个数逐渐减少,量子谐振子的运动逐渐趋于稳定.到达基态时,概率密度函数收敛到一个高斯函数,此时量子谐振子处于稳定状态.当前,对量子谐振子的物理模型应用的研究如下:文献[14]将量子谐振子模型用于解决粒子群算法的全局收敛问题,文献[15]提出了量子谐振子蚁群算法,来解决组合优化中旅行商寻优问题,文献[12]中的量子谐振子算法是根据量子谐振子物理模型设计的一种解决高维函数全局优化问题的算法.其主要包括两个收缩过程:一个是高斯采样函数的尺度从大到小的收敛过程,来获得优化函数的信息.另一个是在同一尺度下高斯采样函数的个数从多到少的收敛过程,该过程是在多维优化函数信息引导下的量子谐振子向能量基态的运动过程.CCSA-QHOM 算法主要是根据该算法中同一尺度下多个高斯采样函数的收缩聚焦过程是一种聚集区间的定位过程来设计的.

3 CCSA-QHOM 算法原理

量子谐振子向基态的物理变化过程与聚类算法中聚类中心的逐步确定过程非常相似.类比于量子谐振子的物理过程,CCSA-QHOM 算法用量子谐振子的物理模型解释如下:算法起始时,初始化 l 组初始中心点,由于此时整个聚类系统中的初始中心点分布是杂乱、无序的,此时中心点所在的网格可以看作是量子谐振子模型中谐振子运动比较频繁的高能态.在网格空间中,从多个初始中心点开始的高斯函数经过多次迭代收敛到网格密度局部最大处的过程,类比于图 1 中所展示的在目标函数指引下的量子谐振子向能量基态运动过程中的概率收敛过程.由于 CCSA-QHOM 算法最终要定位到的位置是所有的网格密度局部最大处,到达这些位置时,高斯函数的迭代过程结束,所以高斯函数收敛到密度局部最大网格时的状态类比于量子谐振子模型中的稳定态-基态.

为了下文描述的方便,本文给出了以下四个定义和一个定理.

定义 1　网格密度 S 指每个网格中包含的原始数据集中的点数.

定义 2　网格尺度 *scale* 指的是每一维上一段网格的长度.

定义 3　采样尺度 σ 指的是高斯函数的方差,描述了高斯函数采样区间的范围.

定义 4　网格密度阈值是指该网格区域成为稠密网格区域所包含的最少数据对象数目.

定理　若 S_i 代表第 i 次高斯采样迭代到的网格对应的密度,则下一步的移动方向评判函数 $\Delta S_i = S_i - S_{i-1}$.

CCSA-QHOM 算法的收敛过程

量子谐振子模型向基态的收敛过程在网格空间上的描述如下:首先,在数据集中随机选择 l 个点,这些点在每一维上的坐标值分别作为该维上高斯采样的初始均值位置,以该 l 个均值位置分别构造标准差为 σ 的高斯函数 $N(X_i,\sigma^2)(i=1,\cdots,l)$,形成 l 个采样区域,从而构造了高能态的量子谐振子. 然后,每个区域分别按 $N(X_i,\sigma^2)$ 在定义域上采样产生 l 个解;将所有维度上第 j 次迭代产生的解构成的向量对应到一个网格,若移动方向评判函数 $\Delta S_j \geq 0$,则以第 j 次迭代产生的解作为每一维上新的均值位置. 否则,分别重新进行第 j 次迭代采样. 这相当于量子谐振子模型中谐振子向低能态跃迁时的概率选择过程. 这样,l 个标准差为 σ 的高斯函数,随着高斯采样区域均值位置的改变而逐渐向聚合中心点所在的网格处聚集. 该过程对应于 l 个高斯函数叠加形成的波函数逐渐收敛于 $m(m \leqslant l)$ 个高斯分布 $N(\bar{X}_{im},\sigma^2)$ 的能量基态. 聚集在聚合中心处的高斯函数构造了基态的量子谐振子. 该收敛过程用 2 维数据集来描述如图 2 所示.

在图 2 中,以量子谐振子中一个高斯采样的运动过程为例来描述 CCSA-QHOM 算法收敛过程. 图中每个网格中的数据代表的是原始数据集被划分到该网格中的数据点数. 图 2 中的收敛过程描述如下:首先,随机选取初始中心点(x_0,y_0),点(x_0,y_0)所对应的网格密度 S_0 为 3,如图 2(a)所示. 然后,分别以 x_0,y_0 作为 x,y 轴上高斯函数的均值,在 x 轴上得到高斯函数 $x \sim N(x_0,\sigma_x^2)$,以此高斯函数在 x 轴上进行采样,选出 x_1. 在 y 轴上得到高斯函数 $y \sim N(y_0,\sigma_y^2)$,以此高斯函数在 y 轴上进行采样,选出 y_1. 点(x_1,y_1)对应的网格密度 S_1 为 63,该过程如图 2(b)所示. 比较 S_0,S_1,由于 $\Delta S_1 \geq 0$,则 x_1,y_1 分别作为 x,y 轴上新的均值. 然后,以高斯函数 $N(x_1,\sigma_x^2)$,$N(y_1,\sigma_y^2)$分别在 x,y 轴上进行采样,选出 x_2,y_2. 点(x_2,y_2)对应的网格密度 S_2 为 245,该过程如图 2(c)所示. 由于 $\Delta S_2 \geq 0$,则以 x_2,y_2 分别作为新的均值的取值,继续高斯采样,由于点(x_2,y_2)对应的网格密度是局部最大的,此后高斯采样选出的作为新的均值的点都将出现在点(x_2,y_2)对应的网格中,整个迭代过程结束.

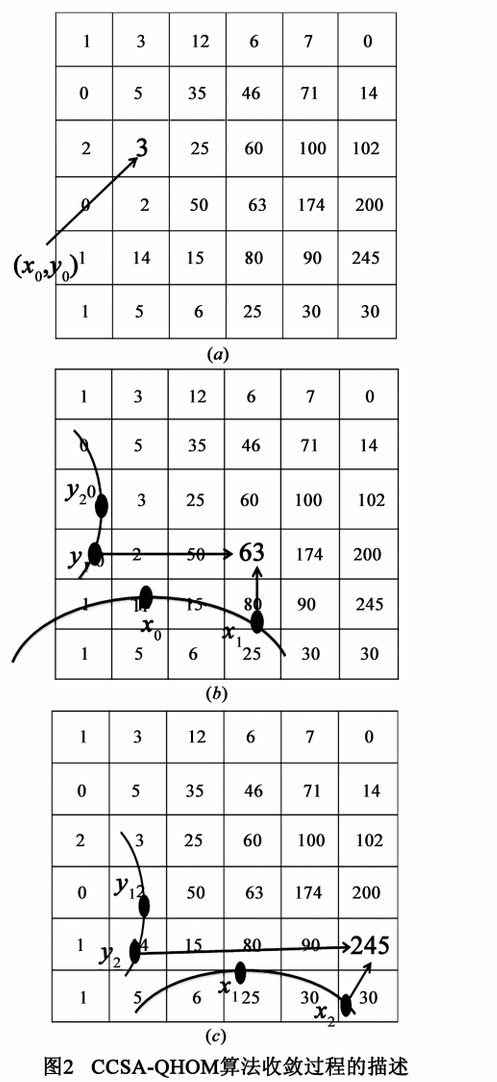


图2　CCSA-QHOM算法收敛过程的描述

4　CCSA-QHOM 算法过程

CCSA-QHOM 算法分为以下几个部分:(1)网格划分阶段,该阶段是整个算法的基础,网格划分的是否合理直接关系到算法最终定位的位置的合理性;(2)数据投影;(3)算法的量子谐振子收敛阶段,该阶段采用高斯函数来进行数据对象的采样与网格位置的定位. 主要过程是对当前解重复“产生新解→计算移动方向评判函数→接受或舍弃”的迭代.

4.1　网格划分阶段

已知数据集为 2 维数据集 D(数值型数据的数据集)划分网格空间 S,CCSA-QHOM 算法采用以下的网格划分方法:由用户设定每一维网格的尺度 scale,若每一维数据取值的最大值为 *imax*,最小值为 *imin*(i 代表第几维),则每一维的网格段数 m_i 为$\dfrac{\lceil imax \rceil - \lfloor imin \rfloor}{scale}$,

网格总数为:$\Pi_{i=1}^{2} m_i = m_1 \times m_2$.

4.2 数据投影阶段

已知在2-维数据集 D 中,x_{ij} 代表数据集中的第 i 个数据点的第 j 维坐标的取值,$jmin$ 代表第 j 维数据中的最小值 R,$\lfloor jmin \rfloor$ 表示小于 R 的最大整数值. 则 x_{ij} 的数据投影方法为 $\frac{x_{ij} - \lfloor jmin \rfloor}{scale}$. 经过投影后的数据集的聚类问题就类似于求解函数 $f(m,n)$(m,n 分别为网格的第一维、第二维坐标)的优化问题. 文献[12]中的量子谐振子算法是一种能够得到完善的全局最优解或局部最优解的函数优化算法,其具有高效的搜索区域收缩定位能力. 鉴于此,设计了下面的量子谐振子收敛阶段的算法.

4.3 算法的量子谐振子收敛阶段

该阶段用对2维数据的处理过程来描述,这一阶段对应于量子谐振子从高能态向基态的能量变化过程.

按照4.1、4.2所示方法划分网格并将数据集 D 中的点投影到网格空间中,记录每个网格的坐标及其密度.

Step1 从数据集 D 中随机选取 l 组点(x_{i0},y_{i0})(其中 i 代表第几个对象,i 的取值范围为0到 l-1)作为初始聚类中心. 并分别记录每组(x_{i0},y_{i0})所对应的网格密度 S_{i0}.

Step2 以 x_{i0},y_{i0} 分别作为 x,y 轴上高斯随机采样函数的均值,设定 x,y 轴的标准差 σ_x,σ_y,在 x 轴上构造以 x_{i0} 为均值的高斯函数 $N(x_{i0},\sigma_x^2)$,可以在 x 轴上形成 l 个采样区域. 同理,在 y 轴上构造以 y_{i0} 为均值的高斯函数 $N(y_{i0},\sigma_y^2)$,在 y 轴上形成 l 个采样区域. 类比于图1,该过程构造了高能态的量子谐振子.

Step3 在 x 轴上的 l 个采样区域内分别以相应的高斯函数 $N(x_{i0},\sigma_x^2)$ 进行高斯采样,选出 x_{i1}. 然后,在 y 轴上的 l 个采样区域内分别以高斯函数 $N(y_{i0},\sigma_y^2)$ 进行高斯采样,选出 y_{i1}. 记录点(x_{i1},y_{i1})所对应的网格密度 S_{i1}.

Step4 分别比较 S_{i0},S_{i1},若 $\Delta S_{i1} \geq 0$,则 x_{i1},y_{i1} 分别作为 x 轴,y 轴上新的高斯采样函数的均值,即新的 x_{i0},y_{i0},继续步骤 Step2. 否则,仍以原来的 x_{i0},y_{i0} 作为高斯采样函数的均值,继续 Step2.

Step5 重复上述步骤 Step2,Step3,Step4 的过程,直到连续迭代多次定位到的网格位置不变为止. 此时,多个高斯函数叠加到密度局部最大的网格处,类比于图1,构造了基态的量子谐振子.

Step6 对所得到的 l 组聚类结果进行比较,将网格位置相邻近的结果进行合并,以其中密度最大的网格位置为代表.

Step7 程序自动设定网格密度阈值,来屏蔽掉噪声数据.

经由以上几步,输出聚类最终个数及聚类中心所在的网格位置.

Step2 到 Step5 整个过程相当于图1中描述的波函数从高能态到基态过程中的概率收敛过程. Step6 类比于每一组初始中心点到达的基态位置的最终确定.

4.4 算法复杂度分析

在数据投影阶段,从数据对象投影到相应网格的计算方法分析知,该阶段的时间复杂度为 $O(n)$,其中 n 为数据集中的点数. 在核心算法阶段,初始化聚类数为 l,分别以每组初始聚类中心点的位置为起始位置进行高斯采样重定位,该过程主要是高斯采样的迭代过程及网格的选择与比较. 则该过程的时间复杂度为 $O(\sum_{i=1}^{l} m_i)$,其中 m_i 为以第 i 个初始值为起始位置所进行的所有高斯采样的迭代次数. 将高斯采样迭代重定位后所得到的 l 组聚类结果中位置相邻的归并为一个结果,这一过程主要是将结果进行两两比较,在最坏情况下所需要的计算次数是 $\frac{l(l-1)}{2}$,则时间复杂度为 $O(l^2)$. 若是比较过程选出 p 个网格位置,由于每一个被选出的网格附近的网格数为 $3q-1$(其中 q 的取值为1或2或3),则对以该 p 个网格位置为中心进行一次遍历的时间复杂度为 $O(pq)$. 最后,在所得的聚类结果中查找大于密度阈值的网格单元,并将其输出,时间复杂度为 $O(p)$. 由于 l,p,q 都为常数,所以该算法核心阶段的时间复杂度为 $O(\sum_{i=1}^{l} m_i)$. 由于每组初始聚类中心位置的随机选择导致不同初始聚类中心开始的高斯采样次数 m_i 是一个不确定的值,但是该值通常情况下不大于网格划分数. 在 l 较大的情况下,即 $\sum_{i=1}^{l} m_i > n$ 时,该算法的整体时间复杂度集中在核心算法阶段,为 $O(\sum_{i=1}^{l} m_i)$. 否则,该算法的整体时间复杂度集中在数据投影阶段,为 $O(n)$.

5 实验分析

5.1 数据集及实验说明

本文所采用的数据来源于标准数据集和仿真生成,数据集 Data1 来自于联合程序开发网,数据集 Data2 和 Data3 由仿真生成. Data1 中有10000组数据和3个类,用于说明噪声数据对算法结果的影响. Data2 中有8829组数据和四个大小不一样的类,用于说明类大小多样的数据对 CCSA-QHOM 算法结果的影响. Data3 中有6175组数据和3个分布密度不同的类,用来检验 CCSA-QHOM 算法对密度多样数据的聚类中心定位情况. 本文还采用 UCI 数据库中的 Iris 数据集,Haberman 数据集,Wilt 数据集(实验时对 Wilt 数据集进行了降维

处理)来验证 CCSA-QHOM 算法的有效性. 通过经典 DBSCAN 算法的数据集来验证 CCSA-QHOM 算法的适用性. 图 3 描述的是数据集 Data1 到 Data3 的二维网格分布图. 此外,本文中的所有实验都是在同一台机器上完成的,具体的软硬件参数如下:操作系统为 Windows XP,CPU 为 Intel Core i3,主频 2.26GHz.

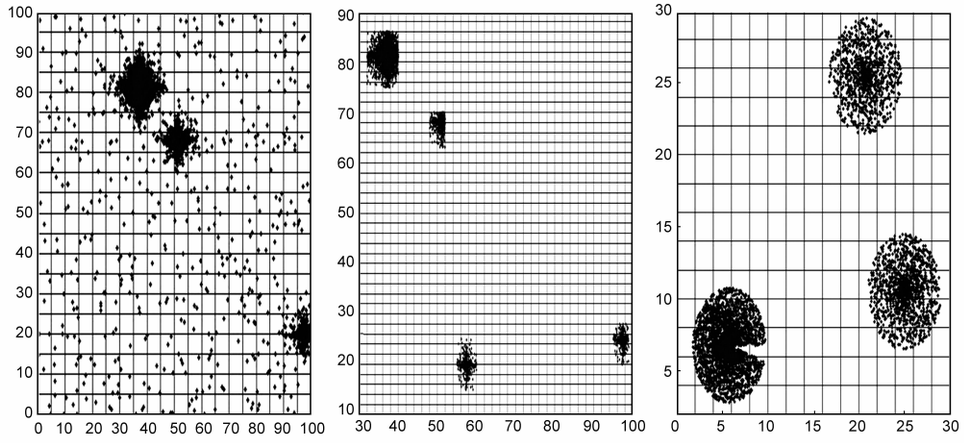


图3　仿真数据集的二维网格分布图

5.2　实验中参数设定对结果的影响

在 CCSA-QHOM 算法中,由于网格尺度 *scale*,初始化聚类数与采样尺度都是未知的,它们的实验数据随着数据集的不同而有所改变,需要根据经验给出大概取值. 其中网格尺度与采样尺度之间的关系是:网格尺度与采样尺度取值之间是正相关的,当网格尺度较大时,较大的采样尺度才更容易得到较好的结果. 由于两者之间的这种关系,所以在下面的分析中只分析了初始化聚类数和采样尺度对结果正确率的影响.

5.2.1　初始化聚类数对结果正确率的影响

从算法的原理分析知:初始化聚类数 l 最小要大于实际的聚类个数. 在实际的聚类个数未知的情况下,l 的取值应尽量大. 文献[16]通过对 k 近邻分类算法中参数 k 设置的研究,给出了如下规则:$k \propto n^{\frac{3}{8}}$($n$ 为样本中点的总数),在 CCSA-QHOM 算法中,由于网格是度量的最小单位,所以 l 的取值不仅与数据点的总数有关,还跟网格的划分尺度有关.

对 CCSA-QHOM 算法进行多次实验发现:当网格尺度 *scale* 划分合适时,上面的规则也是适用的. 当网格尺度足够小时,较大的 l 才能使聚类结果较优.

表 1 描述的是对于数据集 Data1,Data2,Data3,不同的初始化聚类数 l 对结果正确率的影响情况.

对表 1 中的数据进行分析发现:在 100 次算法调试中,随着 l 的增大,算法 CCSA-QHOM 对这三组数据调试的结果正确率都在逐渐增加. 当 l 增大到一定程度时,对结果正确数的影响将不再明显. 出现这种现象的原因是在 CCSA-QHOM 算法中,虽然初始中心点的位置是随机选取的,采样过程也是随机的,但是由于量子谐振子模型中基态的"引力"作用导致从不同的初始中心点开始的采样最后汇聚在同一位置处. 这种情况虽然是 CCSA-QHOM 算法的基本思想过程,但是当 l 较小时,l 与实际的聚类数相差不大,导致这种情况的发生对结果聚类数的影响非常显著. 当 l 较大时,l 远远大于实际聚类数,这种情况发生的次数越多,导致聚类结果质量越优.

表 1　初始化聚类数 l 与结果正确率之间的关系

	5	10	15	20	25	30	35	40
Data1	18%	42%	57%	70%	75%	87%	89%	95%
Data2	1%	15%	21%	43%	55%	74%	73%	76%
Data3	29%	46%	45%	49%	61%	60%	59%	64%

5.2.2　采样尺度 σ 的影响

采样尺度 σ 的选择关系到搜索最优解要进行的采样次数的多少,所以要慎重. 经过多次实验发现:当初始化聚类数,网格尺度一定时,随着 σ 的增大,高斯随机采样的计算次数会减小. 如图 4 所示.

图 4 描述了对于不同数据集 Data1,Data2,Data3,采样尺度 σ 对高斯采样函数迭代次数的影响. 从图中可以看出,对于数据集 Data1,Data2,Data3,随着高斯函数采样尺度的增加,高斯采样的迭代次数在逐渐减少.

这是由于当高斯函数的采样尺度 σ 较小时,采样跨度较小,历经整个网格区间需要行走很多步,即生成很多点. σ 较大时,采样跨度较大,采样频率较低,导致所需的计算次数相对较少. 从图中可以看出 $\sigma = 2$ 处是变化趋势快慢的分水岭,所以,CCSA-QHOM 算法对于 Data1,Data2,Data3 这三种数据集进行调试时,采样尺度选择的都是 2.

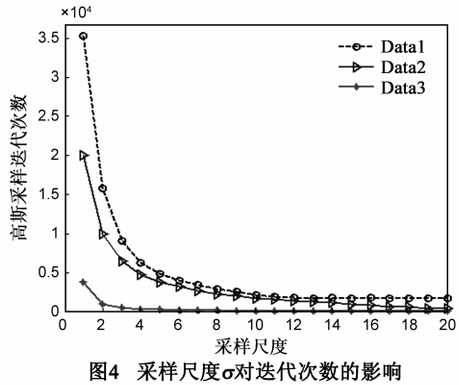


图4 采样尺度σ对迭代次数的影响

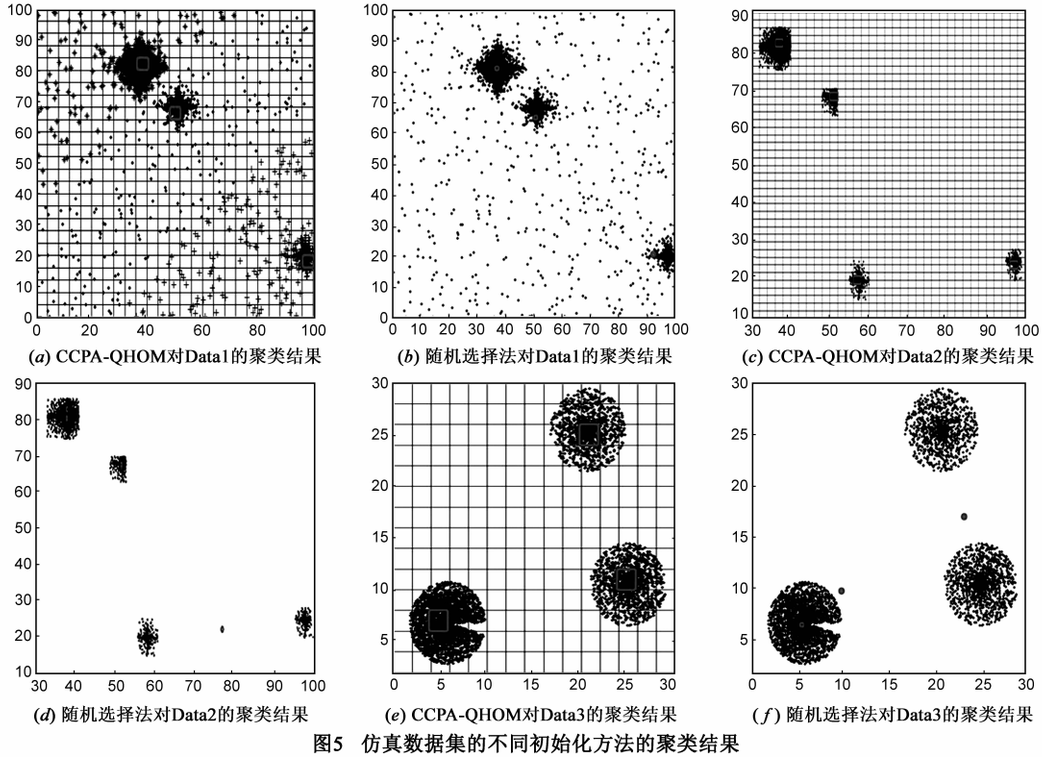
5.3 算法结果评估

5.3.1 仿真数据的结果研究

在下面的实验中,将 CCSA-QHOM 算法与随机选择法这两种 K-means 初始化方法对数据集 Data1,Data2,Data3 的聚类结果分别用图 5 中的图(*a*)~(*f*)表示. 其中的随机选择法展示的是 10 次随机选取出现次数最多的结果. 图 5 中,图(*a*),(*c*),(*e*)中标出的区域是 CCSA-QHOM 算法找到的中心点的位置. 图 5(*b*),(*d*),(*f*)中标出的位置是进行随机选择初始中心点的 K-means 运算最终得到的聚类中心的位置. 可以看出 CCSA-QHOM 算法对于噪声数据,类的大小相差很大的数据,类的密度相差较大的数据等随机选择法不善于处理的数据类型找到的聚类中心点位置更准确. 这是由于 CCSA-QHOM 算法中采用的量子谐振子模型以收敛到密度局部最大的网格处为终止条件,而密度局部最大的区域一般都在子类的聚合中心附近.

表 2,表 3 描述的是对于不同的数据集 Data1,Data2,Data3;CCSA-QHOM,Canopy 算法和传统的 K-means 算法的平均运行时间和最后得到的聚类中心结果与实际聚合中心的距离比较. 对于 CCSA-QHOM 算法中的聚合中心用最终定位到的网格的几何中心代表. Canopy 算法是根据文献[17]描述的第一阶段进行的设计. K-means 算法的数据取自 10 次运算中出现次数最多的聚类结果,对于数据集 Data1,Data2,Data3,K 的取值分别为 4,4,3. 从表 2 可以看出:Canopy 算法对数据集 Data1,Data2,Data3 的平均运行时间最短,但是表 3 表明 CCSA-QHOM 算法定位到的聚合中心效果最优. 其平均运行时间虽比 Canopy 算法长,但依然在一个数量级上,相差不大.

对于 Data1,Data2,Data3 这三种特殊的数据集,CCPA-QHOM 算法最终定位到的聚类中心最优的原因是:CCPA-QHOM 算法由于其基本思想是始终向网格密度局部最大处(量子谐振子中的基态)靠拢,只有定位到密度局部最大处时迭代过程才停止,而密度局部最大处都是在类的聚合中心附近,导致了初始中心点的随机选择只影响了其找到网格密度最大处时的迭代次数及采样的初始范围,对聚类结果质量影响不大. 但是,在 CCPA-QHOM 算法中网格的划分,采样尺度的设置



(*a*) CCPA-QHOM对Data1的聚类结果 (*b*) 随机选择法对Data1的聚类结果 (*c*) CCPA-QHOM对Data2的聚类结果

(*d*) 随机选择法对Data2的聚类结果 (*e*) CCPA-QHOM对Data3的聚类结果 (*f*) 随机选择法对Data3的聚类结果

图5 仿真数据集的不同初始化方法的聚类结果

是否合理直接关系着算法最终定位到的位置的合理性.

表 2　不同聚类算法对三种数据集的平均运行时间(s)的比较

算法 / DataSet	CCSA-QHOM	Canopy	K-means
Data1	0.5430	0.2470	3.0630
Data2	0.5760	0.2140	7.8750
Data3	0.4387	0.1260	1.6410

表 3　不同聚类算法得到的聚合中心与实际聚合中心的距离

算法 / DataSet	CCSA-QHOM	Canopy	K-means
Data1	3.7358	88.3772	6.2654
Data2	1.2578	17.0252	69.8911
Data3	1.2287	4.9640	17.6850

5.3.2　真实数据集的结果研究

表 4 描述了不同的初始聚类中心选择方式对真实数据集进行处理得到的聚类中心与实际聚合中心的距离. 表 4 说明了 CCSA-QHOM 算法始终以找到网格密度最大的区域为目标,由于网格划分尺度的影响,其找的中心位置与随机选择算法的某些结果相比并不是最好的,但是其定位到的中心位置始终比 Canopy 算法要好,且在网格尺度、采样尺度不变的情况下,其定位到中心位置始终不变,稳定性比较好. 从而说明 CCSA-QHOM 算法进行中心点的确定是可行的.

表 4　不同初始中心选择方式对真实数据集得到的聚类中心与实际聚合中心的比较

聚类中心选择方式 / DataSet	Iris	Haberman	Wilt
随机选取法 1	0.5196	16.9872	328.6005
随机选取法 2	0.5302	16.8821	732.0291
随机选取法 3	4.5079	16.8077	177.3188
Canopy 算法	2.7650	36.5144	534.3387
CCSA-QHOM	1.4037	18.2761	342.2879

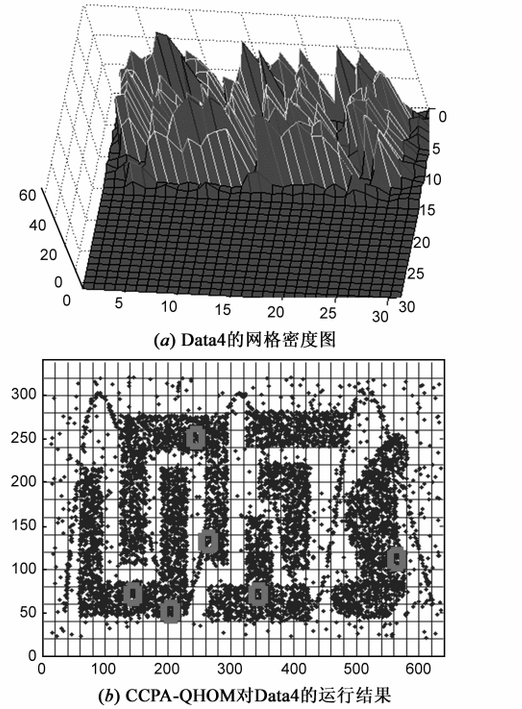
表 5 描述的是:对于不同的初始聚类中心选择方法产生的初始聚类中心,分别作为 K-means 算法的初始聚类中心进行 K-means 聚类运算得到的结果中数据点被正确划分所占比率. 从表中可以看出:CCSA-QHOM 算法的结果作为 K-means 算法的初始聚类中心来进行 K-means 算法有相对较优的数据点划分正确率.

表 5　不同初始中心选择方式的 k-means 算法性能比较

聚类中心选择方式 / DataSet	Iris	Haberman	Wilt
随机选取法 2	88.0%	81.0%	82.5%
随机选取法 3	54.0%	77.5%	67.9%
Canopy 算法	93.0%	80.0%	52.0%
CCSA-QHOM 结果为初始聚类中心	93.0%	82.0%	52.0%

5.3.3　算法的适用性说明

上文中的仿真数据集和真实数据集在网格空间上的密度分布都是每个子类中有唯一的密度最大处,转换成函数优化问题是一种单个子类的网格密度呈单峰分布的多极值函数优化问题. 量子谐振子算法[12]由于其高斯函数聚集收缩的特性导致其对于单峰的函数优化问题有较好的效率,对于子类局部区域中有多个峰值的函数优化问题效果不好. 从而暗示了根据量子谐振子模型设计的 CCSA-QHOM 算法对于在网格空间上的每个子类的密度分布呈单峰特性的数据集的聚类问题效果较好,不适合处理均匀分布的数据集的聚类中心定位问题,如图 6 所示. 对于均匀分布的数据集等某一子类的局部区域中的多极值现象需要进一步网格划分或者进行极值处理,即需要更进一步的研究.



(*a*) Data4的网格密度图

(*b*) CCPA-QHOM对Data4的运行结果

图6　Data4的网格密度分布图及CCPA-QHOM对其的运行结果

6　总结

CCSA-QHOM 算法通过以量子谐振子模型中波函数从高能态到基态过程中的概率变化为理论模型来描述数据对象向聚类中心移动的过程,最终定位到密度局部最大的网格位置. 该算法在合理的参数设置下,不需要先验知识给出实际的聚类个数,运算过程中会自动聚为若干类. 本文通过对三种具有显著特征的数据集的实验研究,说明了 CCSA-QHOM 算法对于噪声数据,类大小差别较大,密度多样等数据类型在参数设置合理时都不太敏感,聚类中心定位较准确. 通过对真实数据集的实验研究说明该方法能够定位到离实际的聚合中心比较近的网格位置,从而说明量子谐振子模型

收敛过程的有效性.通过对 CCSA-QHOM 算法所采用的数据集及运算结果分析发现:该算法较适合于网格空间上的每个子类的密度分布呈单峰特性的数据集的聚类问题,对于网格空间上密度分布相对均匀及发散分布的数据集不太适合.通过对该算法的原理及结果进行分析发现:该算法可以作为类似于 K-means 算法的需要选取初始聚类中心点的动态聚类算法的预处理过程,也可以用于对离群点的检测等.

参考文献

[1] 徐利治.数学辞海(第四卷)[M].山西太原:山西教育出版社,2002.444 -447.

[2] Han J W, Micheline K. Data Mining Concepts and Techniques[M]. San Francisco: Morgan Kaufmann, 2006. 383 -464.

[3] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48 -61.
Sun Ji-gui, Liu Jie, Zhao Lian-yu. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48 -61. (in Chinese)

[4] Rana S, Jasola S, Kumar R. A boundary restricted adapti-ve particle swarm optimization for data clustering[J]. International Journal of Machine Learning and Cybernetics, 2013, 4(4): 391 -400.

[5] Ji He, Man Lan et al. Initialization of cluster refinement algorithms: A review and comparative study[A]. IEEE International Joint Conference on Neural Networks[C]. Budapest: IEEE, 2004. 1 -6.

[6] Mac Q J. Some methods for classification and analysis of multivariate observations [A]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability[C]. Berkeley, CA: University of California Press, 1967. 281 -297.

[7] Toy J T, Gonzalez R C. Pattern Recognition Principles [M]. Massachusetts: Addison-Wesley, 1974.

[8] Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis [M]. Hoboken NJ: John Wiley & Sons, 2009.

[9] Omran M, Salman A, Engelbrecht A P. Image classification using particle swarm optimization[A]. Proc of the 4th Asia-Pacific Conference on Simulated Evolution and Learning[C]. Singapore: 2002. 18 -22.

[10] Merwe D W, Engelbrecht A P. Data clustering usingparticle swarm optimization[A]. Proc of IEEE Congresson Evolution Computation[C]. Piscataway, NJ: IEEE, 2003. 215 -220.

[11] Omran M, Salman A, Engelbrecht A P. Dynamic clustering using particle swarm optimization with application in unsupervised image classification[A]. Fifth World Enformatika Conference (ICCI 2005)[C]. Prague: Czech Republic, 2005. 199 -204.

[12] 王鹏,黄焱,任超,郭又铭.多尺度量子谐振子高维函数全局优化算法[J].电子学报,2013,41(12):2468 -247.
Wang Peng, Huang Yan et al. Multi-scale quantum harmonic oscillator for high dimensional function global optimization algorithm[J]. Acta Electronica Sinica, 2013, 41(12): 2468 -2473. (in Chinese)

[13] 曾谨言.量子力学教程[M].北京:科学出版社,2003. 47 -50.

[14] 冯斌,须文波.基于粒子群算法的量子谐振子模型[J].计算机工程,2006,32(20):18 -21.
Feng Bo, Xu Wen-bo. Quantum oscillator model of particle swarm system[J]. Computer Engineering, 2006, 32(20): 18 -21. (in Chinese).

[15] 秦永波,王鹏,肖黎彬,江炳坤,任超,孟玉.量子谐振子蚁群算[J].计算机应用,2011,31(z2):54 -69.
Qin Yong-bo, Wang Peng, Xiao Li-bin et al. Ant colony optimization of quantum harmonic oscillators[J]. Journal of Computer Applications, 2011, 31(z2): 54 -69. (in Chinese)

[16] Hand DJ, Vinciotti V. Choosing k for two-class nearest neighbor classifiers with unbalanced classes[J]. Pattern Recognition Letters, 2003, 24(9): 1555 -1562.

[17] Mc Callum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching[A]. Proceedings of theSixth ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2000. 169 -178.

作者简介

燕京京 女,1989 年生于河南安阳.硕士研究生,主要研究领域为数据挖掘.

王 鹏(通信作者) 男,1975 年生于四川乐山.教授,博士生导师,CCF 高级会员,主要研究领域为并行计算,信号处理,智能算法.
Email: wp002005@163.com