DOI:10.13295/j.cnki.jlut.2009.05.034
第35卷 第5期 兰州理工大学学报 Vol.35 No.5
2009年10月 Journal of Lanzhou University of Technology Oct.2009

文章编号:1673-5196(2009)05-0089-06

一种基于量子机制的分类属性数据层次聚类算法

赵正天[1],赵小强[1],李 炜[1],段晓燕[2],卢 勇[3]

(1.兰州理工大学 电气工程与信息工程学院,甘肃 兰州 730050;2.兰州石化职业技术学院 电子电气工程系,甘肃 兰州 730060;3.中国石油兰州石化电仪事业部,甘肃 兰州 730060)

摘要:受物理学中量子机制特性的启发,结合层次凝聚思想,通过引入新的相异性度量测度以及聚类度量尺度步长β_{step}概念,重新定义以紧致性指标AIAD和离散性指标AIED为基础的聚类有效性函数CVF,提出一种针对分类属性数据的基于量子机制层次聚类算法CQHC.该算法首先在不同粒度水平上划分数据样本产生初始类(簇),然后以聚类有效性函数CVF为评价标准,动态地合并初始类(簇)完成聚类.仿真实验采用2个真实数据集,即:线性可分的大豆疾病样本数据集和线性不可分的动物园数据集.实验结果表明,该算法与已有的其他几个算法相比,不仅具有更高的聚类准确率,而且能够准确地检测出最佳类别数,是有效且可行的.

关键词:分类属性;量子机制;层次凝聚;聚类度量尺度步长;聚类有效性函数

中图分类号:TP301.6 **文献标识码**:A

A hierarchical clustering algorithm of categorical attributive data using quantum mechanism

ZHAO Zheng-tian[1], ZHAO Xiao-qiang[1], LI Wei[1], DUAN Xiao-yan[2], LU Yong[3]

(1. College of Electrical and Information Engineering, Lanzhou Univ. of Tech., Lanzhou 730050, China; 2. Department of Electronic and Electrical Engineering, Lanzhou Petrochemical College of Vocational Technology, Lanzhou 730060, China; 3. Petro China Lanzhou Petrochemical Company ELEC. & INTR. Headquarter, Lanzhou 730060, China)

Abstract: Enlightened by quantum mechanics in physics and incorporated with agglomerative hierarchical clustering, a quantum mechanism-based hierarchical clustering algorithm of categorical attributive data CQHC was proposed by introducing a new dissimilarity measure and a concept of clustering measure scale step β_{step}, and redefining the cluster validity function CVF based on compactness index AIAD and discreteness index AIED. In this algorithm of CQHC, the data sample was partitioned first according to different granularities levels to generate initial clusters. Then the initial clusters were dynamically merged by taking the cluster validity function CVF as evaluation standard and the clustering was completed. Two real data sets, including linear separable soybean disease data sets and linear inseparable zoo data sets, were used for simulation experiment. Experimental result demonstrated that the proposed algorithm was effective and feasible, which not only had higher clustering accuracy, but also accurately detected the best cluster number when compared to other algorithms available.

Key words: categorical attribute; quantum mechanism; hierarchical clustering; clustering measure scale step; cluster validity function

俗话说"物以类聚,人以群分",聚类是人类认识活动的一个重要组成部分.所谓聚类,就是将数据对象分组成为多个类(簇),使同一个类(簇)中的对象具有较高的相似性,而不同类(簇)中的对象具有较大的相异性.从机器学习的角度看,聚类属于无指导学习,与分类不同,不依赖预先定义的类和带类标号的训练对象.良好的聚类方法产生的聚类结果具有类(簇)内对象高度相似,类(簇)间对象很少相似的特性[1].

分类属性数据是待聚类的数据类型中常见的一类,其属性是有限无序的,且不可比较大小,如{红,

收稿日期:2009-04-08
基金项目:甘肃省自然科学基金(0809RJZA005)
作者简介:赵正天(1980-),男,甘肃民勤人,硕士,讲师.

黄,蓝,绿},{正方形,圆形,梯形}等.由于分类属性数据分布固有的无序性,使得只有诸如 k-mode 算法[2-3]、k-prototype 算法[4]和模糊 k-prototype 算法[5]、CQC(categorical quantum clustering)算法[6]等少数几种算法能实现对其聚类.然而,这些算法或多或少的存在不稳定、随机性差等缺点.其中,k-mode 算法和 k-prototype 算法在聚类模式的初始点选择上采用了随机选择初始点的方法,导致聚类结果对初始点过于敏感,甚至使聚类结果偏离实际情况;模糊 k-prototypes 算法是软聚类(soft clustering)格式,具有较好的性质,但遇到特殊分类属性时会出现属性值丢失现象;CQC 算法对聚类度量尺度 β 较敏感,而 β 往往凭经验确定,没有通用的原则,可操作性差,且该算法对线性可分数据聚类效果显著,而对线性不可分数据不能奏效,其有效性很大程度上取决于样本的分布情况.因此,积极探索新的、更有效的针对分类属性数据的聚类算法依然是聚类研究的一个开放问题.

对于分类属性数据,本文作者曾采用 Ahmad 相异性度量测度替换 Hamming 相异性度量测度,针对 CQC 算法从计算分类属性数据间相异度的角度提出一种改进算法[7],但改进算法由于受到固定不变聚类度量尺度 β 的限制,在对线性不可分数据的聚类上也仍显不足.本文在量子聚类的基础上,通过重新计算分类属性数据样本间的的相异性度量测度,引入聚类度量尺度步长(clustering measure scale step)β_{step},定义聚类有效性函数,提出一种基于量子机制的层次聚类算法(categorical quantum hierarchical clustering,CQHC).

1 量子势能

波函数是粒子量子态的描述,薛定锷方程的目的就是求解有势场约束的波函数.不"显含"时间的薛定锷方程为

$$H\varphi=\left(-\frac{\delta^2}{2}\tilde{\nabla}^2+V(X)\right)\varphi=E\varphi \tag{1}$$

若 φ 已知,用薛定锷方程式(1)求解粒子分布的势能函数,如下式所示:

$$V(X)=E+\frac{\left(\delta^2/2\right)\tilde{\nabla}^2\varphi}{\varphi} \tag{2}$$

式中:φ 为波函数;H 为 Hamilton 算子;V 为势能函数;E 为 H 算子的能量特征值;$\tilde{\nabla}$ 为劈形算子;δ 为波函数的宽度调节参数.薛定锷方程的物理含义是:在一定的势场中,求解粒子的分布,即波函数 φ.从式(1)不难看出,势场相同,那么粒子的分布状态相同,当粒子的空间分布缩变到一维无限深势阱时,粒子聚集在势能为 0 的一定宽度的势阱中.因此,势能函数相当于一个抽象的源,对粒子具有吸引作用,随着势能趋近于 0 或比较小时,势阱中往往分布有比较多的粒子.这一过程的逆就是量子聚类的物理思想的依据.

假设 V 是非负且确定的,即 V 的最小值为 0,那么通过式(2)可以求得 E,如下式所示:

$$E=-\min\frac{\left(\delta^2/2\right)\tilde{\nabla}^2\varphi}{\varphi} \tag{3}$$

已知高斯函数是薛定锷方程的解之一,高斯波包如下式所示,它代表粒子的分布状态:

$$\varphi(X)=\sum_{i=1}^{n}e^{-(X-X_i)^2/2\delta^2} \tag{4}$$

假定它对应于尺度空间(scale-space)中的一个观测样本集 $X=\{X_1,X_2,\cdots,X_i,\cdots,X_n\}\subset R^m$,$X_i=(X_{i1},X_{i2},\cdots,X_{im})^{\mathrm{T}}\in R^m$,问题变成使用宽度为 δ(scale parameter)的高斯波包来描述样本点的分布.根据 Mercer 理论,高斯函数 $\varphi(X)$ 在此相当于一个核函数,可以认为它定义了一个到 Hilbert 空间的非线性映射,作用是把非线性的输入空间转换成 Hilbert 空间,所以同样可以认为 δ 相当于一个核宽度调节参数.

把式(4)代入式(2),得样本服从高斯分布时的量子势能计算公式:

$$V(X)=E-\frac{d}{2}+\frac{1}{2\delta^2\varphi}\sum_i(X-X_i)^2\exp\left[-\frac{(X-X_i)^2}{2\delta^2}\right] \tag{5}$$

其中 d 为算子 H 最小的可能特征值,可以用样本的维数 m 表示.

由于样本的势能是可以确定计算的,根据量子理论可知,低势能的粒子振动小,相对比较稳定.对于聚类,就相当于势能为 0 或最小的样本周围分布着较多的样本,因此可以用于确定聚类中心.

2 基于量子机制的分类属性数据层次聚类算法

2.1 分类属性数据样本的相异性度量

多数分类属性数据的相异性度量测度规定同一属性不同取值之间的距离相等.然而,按照人们一般的经验,有些分类属性取值之间的相似性高,有些则恰好相反,例如"颜色"属性的取值"红色"与"橙色"间的距离小于"白色"与"黑色"间的距离.因此,本文

采用文献[8]重新定义的同一属性不同取值间关于其他属性的相异性度量测度,基于此分类属性数据样本间的相异性度量可描述如下:

定义 1　设样本 $X_i, X_j \in R^m$,其中 $X_i=(X_{i1}, X_{i2}, \cdots, X_{im})$, $X_j=(X_{j1}, X_{j2}, \cdots, X_{jm})$,则 X_i 与 X_j 间的相异性度量测度为

$$d(X_i, X_j)=\sum_{k=1}^{m}\delta(X_{ik}, X_{jk})/m \qquad (6)$$

其中,$\delta(X_{ik}, X_{jk})$是样本的第 k 维属性 2 个不同取值 X_{ik}和 X_{jk}间关于其他属性的相异性度量测度[8], m 是样本的维数.显然,当 $X_i=X_j$ 时,$d(X_i, X_j)=0$, $d(X_i, X_j)=d(X_j, X_i)$, $0<d(X_i, X_j)<1$.

2.2　聚类度量尺度步长 β_{step}

聚类度量尺度是对样本间的相异性度量测度 $d(X_i, X_j)$的程度度量,是一个可变参数 β,显然取值范围是$(0,1)$.其实质是对相异性度量测度值的一个度量参数(即:划分粒度).在进行聚类分析时,用它作为聚类的度量标准.

从聚类过程可以直观地理解,β应小于聚类簇间的相异性测度,而又大于聚类簇内的相异性测度,由这样一个理想的 β或它的经验近似估计 $\hat{\beta}$ 指导样本的划分.然而,对于样本空间中分布混乱,归属于不同潜在类(簇)的样本相互混杂重叠,导致这样一个理想的 β并不存在,牵强地选取一个所谓的经验 $\hat{\beta}$ 值指导样本的划分势必导致聚类效果变差,甚至面目全非.鉴于此,引入聚类度量尺度步长 β_{step},使 β以步长 β_{step} 在其取值范围内适宜地区间遍历,以保证聚类过程能够动态地在不同粒度水平上划分样本.

2.3　聚类有效性函数

大多数聚类有效性函数兼顾了紧致性和分离性这 2 个主要影响因素[9].紧致性度量类(簇)内各样本之间的紧密程度或一致程度.分离性表示类(簇)之间的离散程度或相异程度.

2.3.1　紧致性指标 AIAD

定义 2　设 $C=\{X_1, X_2, \cdots X_i, \cdots, X_u\}$是聚类产生的一个包含 u 个样本的类(簇),则 C 的簇内平均距离(average intra-cluster distance, AIAD)为

$$\mathrm{AIAD}(C)=\left(\sum_{i=1}^{u}\sum_{j=1}^{u}d\left(X_i, X_j\right)\right)/u^2 \qquad (7)$$

其中,$d\left(X_i, X_j\right)$ 是 X_i 与 X_j 间的相异性度量测度.以 AIAD 作为描述类(簇)内各样本间的紧密程度的紧致性指标,显然,其值越小,类(簇)内的样本间越紧密,类(簇)的紧致性越高.在不同划分粒度下可通过比较 AIAD 值选择与之相应的最合理类(簇).

2.3.2　离散性指标 AIED

定义 3　设 $C_h=\{X_1, X_2, \cdots X_i, \cdots, X_u\}$和 $C_k=\{Y_1, Y_2, \cdots, Y_i, \cdots, Y_v\}$是聚类产生的 2 个类(簇),分别包含 u 个和 v 个样本,则 C_h 和 C_k 的簇间平均距离(average inter-cluster distance, AIED)为

$$\mathrm{AIED}(C_h, C_k)=\left(\sum_{i=1}^{u}\sum_{j=1}^{v}d\left(X_i, Y_j\right)\right)/uv \qquad (8)$$

其中,$d\left(X_i, Y_j\right)$ 是 X_i 与 Y_j 间的相异性度量测度.以 AIED 作为描述类(簇)间的离散程度的离散性指标,显然,其值越大,类(簇)间越分散.

2.3.3　聚类有效性函数 CVF

根据上述紧致性和分离性度量,提出如下聚类有效性函数(cluster validity function, CVF).

定义 4　聚类有效性函数 CVF 为

$$\mathrm{CVF}=\frac{\varphi\left(c\right)\cdot\max\limits_{i=1}^{c}[\mathrm{AIAD}(C_i)]}{\min\limits_{h\neq k}[\mathrm{AIED}(C_h, C_k)]} \qquad (9)$$

其中,c 是类别数,$\varphi\left(c\right)=(c+1)/\left(c-1\right)$,引入该项的目的是抵消类别数 c 的变化对离散性指标 AIED 的影响.在聚类过程中,对数据样本固有类(簇)的错误合并会首先反映到最不合理的类(簇)中,引起相应的紧致性和离散性指标较显著的变化,如果使用整体或平均的紧致性和离散性,会弱化指标的这种变化,使聚类有效性函数的敏感性降低[10].所以,在本文提出的聚类有效性函数 CVF 中分别使用了 AIAD 和 AIED 的最大值和最小值,它们分别是紧致性和离散性的最不合理取值,代表最不利的聚类状态.

2.4　算法描述

CQHC 算法可分为两部分,首先第 1～18 行利用步长 β_{step}使 β在其取值范围的适宜区间内遍历,使得能够在不同的划分粒度 β上划分数据样本,并以紧致性指标 AIAD 为选择依据产生初始的 $\hat{c}$ 个类(簇);然后,第 19～32 行依据凝聚层次聚类算法的思想对已划分的初始类(簇)进行合并,产生最终的 c 个类(簇).

层次聚类是常用的聚类方法之一,其缺陷在于:一旦一个步骤(合并或分裂)完成,就无法回到先前的聚类状态,如果在某一步所作的合并的决策不合适,那么将会导致低质量的聚类结果.因此,为克服上述缺点,在层次聚类过程中动态地进行类(簇)的合并与分裂,即按照一定的准则合并类(簇),同时评价合并前后的聚类质量,如果合并使得聚类质量下降,就取消合并,从而提高聚类效果.

合并准则:计算各类(簇)间的相异性测度,合并

具有最小相异性测度的 2 个类(簇).

取消合并准则:使用聚类有效性函数衡量 2 个类(簇)合并前后的聚类质量.若合并后聚类有效性函数比合并前小,表明类(簇)内样本分布比合并前紧密,聚类质量上升,保留合并;反之则不合并.

此外,由文献[11]可知,采用高斯核宽度参数估计方法,对量子势能中的参数 δ 作指导性估计,如下式所示:

$$\overset{\circ}{\delta}=\left\{\frac{4}{(m+2)}\right\}^{1/(m+4)} n^{-1/(m+4)} \tag{10}$$

其中,m 是样本的维数,n 是数据样本集的个数.它包含了样本的维数、大小等信息,能体现样本集的潜在结构对聚类性能的影响.

算法中通过用相异性度量测度式(6)代替式(5)中的欧氏距离部分来计算样本的势能.CQHC 算法描述如下:

输入:包含 n 个样本,且每个样本具有 m 维分类属性的数据集;聚类度量尺度步长 β_{step};

输出:c 个类(簇).

方法:

(1) 初始化 β_{step};

(2) 用式(6)计算样本间的相异性度量测度,得度量矩阵 $\boldsymbol{D}$,$\boldsymbol{D}$ 是有关 $d(X_i, X_j)$ $(1\leqslant i\leqslant n, 1\leqslant j\leqslant n, i\neq j)$的一个高维矩阵;

(3) 用式(10)估算参数 δ,根据式(5)计算样本的势能 V;

(4) for $\beta=0$ to 1 do

(5) $\beta=\beta+\beta_{\text{step}}$;

(6) $\hat{c}=0$, $\hat{X}=X$;

(7) do {

(8) $\hat{c}=\hat{c}+1$;

(9) 根据样本的势能 V,选择势能最小的样本点,即 $V_{\min}=\min_n(V_n)$,且令:$V(X_k)=V_{\min}$, $v_{\hat{c}}=X_k$ 为第 $\hat{c}$ 类聚类中心;

(10) 根据 $\boldsymbol{D}$,满足相异性度量测度 $d(X_i, v_{\hat{c}})\leqslant\beta$ $(1\leqslant i\leqslant n(\hat{c}), i\neq k)$的所有样本聚成第 $\hat{c}$ 类,并从样本集 $\hat{X}$ 中删除这些样本;

(11) } while($\hat{X}$ 不为空);

(12) 计算每个类(簇)的 AIAD;

(13) 保存 AIAD 值最小的类(簇)并从样本集 X 中删除该类(簇)的样本,放弃剩余的类(簇),它们的样本待下次划分;

(14) if 样本集 X 为空 then

(15) $c=\hat{c}$;

(16) break;

(17) end if

(18) end for

(19) 用式(8)计算每两个类(簇)间的离散性指标 AIED,构建类(簇)间差异度矩阵 $\boldsymbol{DM}$;

(20) 对 $\boldsymbol{DM}$ 中的值按从小到大排序,构建向量 $\boldsymbol{V}$;

(21) 从 $\boldsymbol{V}$ 中查找出最小的离散性指标 AIED (C_q, C_p)以及对应的 2 个类(簇)C_p 和 C_q;

(22) 分别计算合并前的 CVF_b 和合并后的 CVF_a;

(23) if $\text{CVF}_\text{a}\leqslant\text{CVF}_\text{b}$ then

(24)合并类(簇)C_p 和 C_q,选择合并后类(簇)中势能最小的样本点为其新的聚类中心;

(25) $c=c-1$;

(26) 更新 $\boldsymbol{DM}$,转至(20);

(27) else

(28) if 达到向量 $\boldsymbol{V}$ 的最后一项 then

(29) break;

(30) end if

(31) 从 $\boldsymbol{V}$ 查找出 AIED(C_q, C_p)后继的离散性指标 AIED(C_h, C_k)以及对应的 2 个类(簇)C_h 和 C_k,转至(22);

(32) end if

值得注意的是,上述算法的层次聚类部分理想情况下应随 CVF_a 收敛至全局最小值,产生最佳的类别数 c 和理想的聚类效果.然而,通常情况下这部分会止于某个局部最小的 CVF_a,导致所产生的类别数 c 偏离最佳类别数,且聚类效果不佳.针对这种情况,对 CQHC 算法从 19 行开始作如下调整:

(19) 用式(8)计算每两个类(簇)间的离散性指标 AIED,构建类(簇)间差异度矩阵 $\boldsymbol{DM}$;

(20) 对 $\boldsymbol{DM}$ 中的值按从小到大排序,构建向量 $\boldsymbol{V}$;

(21) 从 $\boldsymbol{V}$ 中查找出最小的离散性指标 AIED (C_q, C_p)以及对应的 2 个类(簇)C_p 和 C_q;

(22) 分别计算合并前的 CVF_b 和合并后的 CVF_a;

(23) if $\text{CVF}_\text{a}\leqslant\text{CVF}_\text{b}$ then

(24) 合并类(簇)C_p 和 C_q,选择合并后类(簇)中势能最小的样本点为其新的聚类中心;

(25) $c=c-1$;

(26) 保存最小的 CVF_a 和对应的聚类状态,清除 Store_CVF 中的内容;

(27) 更新 $\boldsymbol{DM}$,转至(20);

(28) else

(29) 保存 CVF_a 至向量 Store_CVF,并保存对应的类(簇)C_p 和 C_q

(30) if 达到向量 $\boldsymbol{V}$ 的最后一项 then

(31) 从 Store_CVF 中选择最小的 CVF_a,以及对应的类(簇)C_p 和 C_q,转至(24);

(32) end if

(33)从 $\boldsymbol{V}$ 查找出 AIED(C_q, C_p)后继的离散性指标 AIED(C_h, C_k)以及对应的 2 个类(簇)C_h 和 C_k,转至(22);

(34) end if

从上述调整可以看出:首先,在完成类(簇)合并与类别数 c 的修改后保存了全局最小的 CVF_a 和相应的聚类状态;其次,当 $\boldsymbol{V}$ 中所有的 AIED(C_q, C_p)以及对应的类(簇)C_p 和 C_q 均不能满足 $CVF_a \leqslant CVF_b$ 时,从 Store_CVF 中选择最小的 CVF_a,以及对应的类(簇)C_p 和 C_q 进行合并,保证聚类过程的继续,直至所有样本都归为一个类(簇).待聚类过程结束后,被保存下来的与全局最小 CVF_a 相应的聚类状态(包括类别数 c 和样本划分)即为最佳聚类结果.

3　实验结果

为了验证 CQHC 算法对分类属性数据样本聚类的有效性和可行性,仿真实验采用源于 UC Irvine Machine Learning Repository 的 2 个标准数据集 Soybean 和 Zoo 作为测试数据集,如表 1 所示.它们是常用的知名数据集,已知其聚类结果可靠、并取得一致意见,适合做聚类分析的基准数据集.

表 1　实验样本数据集的组成

Tab.1　Experimental data sets

数据集名称	样本个数	维数	类
soybean disease(大豆疾病数据集)	47	35	4
Zoo(动物园数据集)	101	16	7

其中,大豆疾病数据集是线性可分的,共有 47 个记录样本,每个样本由 35 维分类属性特征描述.每个样本都被标记为 4 种疾病中的一种:Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot 和 Phytophthora Rot.除了 Phytophthora Rot 有 17 个样本外,其他的每种疾病都有 10 个样本.

动物园数据集是线性不可分的,共有 101 个记录样本,分为 7 类,每个样本由 16 维属性特征描述,其中 15 维是布尔属性{0,1}和 1 维分类属性(腿的数量{0,2,4,5,6,8}).

3.1　大豆疾病数据实验

首先引入 2 个评价聚类效果的性能指标:错分率和聚类准确率 γ[12].其中,错分率是指某类中被错分的样本与该类中样本数的比值;聚类准确率 γ 是指各个聚类中所有被正确聚类的样本数总和与样本集样本总数的比值,如下式所示:

$$\gamma = \sum_{i=1}^{c} \text{number}_i / n \tag{11}$$

式中:number_i 为第 i 类中被正确聚类的样本个数;n 为样本总数.γ 值越大,说明聚类准确度越高,聚类效果越好;γ 值越小,则相反.

大豆疾病数据实验结果如表 2 所示,其中本文提出的 CQHC 算法选取 0.001 的聚类度量尺度步长 β_{step},所有样本均被正确分类,聚类准确率 γ 为 100%;CQC 算法的聚类度量参数 β 选择为 0.45 时,该算法产生最佳聚类结果,聚类准确率 γ 为 89.4%;k-modes 算法的参数 α=1.1,并选用文献[2]中的初始模式(initial modes),由于该算法在聚类过程中,每次迭代均需要重新计算新的聚类中心,造成算法对初始中心的敏感,每两次聚类间的结果不一定相同,因此表 2 中的错分率指标是该算法的最佳聚类结果[2],聚类准确率是该算法进行 100 次随机实验的统计平均值.显然,对于线性可分数据 CQHC 算法的聚类效果明显优于已有的 2 种算法.

对于大豆疾病数据集,图 1 显示了本文提出的 CQHC 算法中函数 CVF 随类别数 c 的变化曲线.由于大豆疾病样本数据的线性可分,CVF 函数在类

表 2　大豆疾病数据聚类结果

Tab.2　Clustering result of soybean disease data

算法	类 1	类 2	类 3	类 4	聚类准确率 γ/%
CQHC	0∶10	0∶10	0∶10	0∶17	100
CQC	2∶10	0∶10	1∶10	2∶17	89.4
k-modes	1∶10	0∶10	0∶10	3∶17	78.9

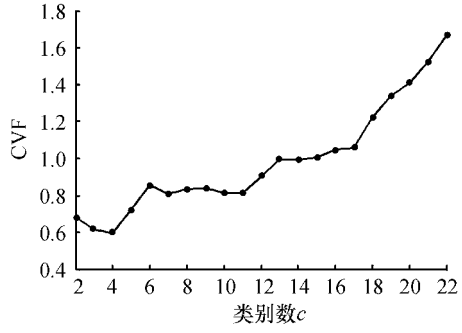

图 1　大豆疾病数据集的 CVF 与类别数 c 关系

Fig.1　Relationship between CVF value and number of clusters of soybean disease data

别数 $c=4$ 处出现较陡峭的谷点，即正确检测到最佳类别数.

3.2 动物园数据实验

本文提出 CQHC 算法以 0.001 的聚类度量尺度步长 β_{step} 对动物园数据进行聚类，结果如表 3 所示. 其中，数据样本被聚类为 7 类，爬行类和两栖类被错分成一类，软体类中的一个样本被单独分为一类，其他类基本正确，错分数目 8 个，表中带“*”的为错分样本数，聚类准确率 γ 为 92.08%. 对于 CQC 算法，由于动物园数据的线性不可分，使得适合于该数据集的聚类度量参数 β 不存在，导致 CQC 算法对动物园数据的聚类失败. 对于 k-modes 算法，用动物园数据进行 20 次随机实验，其平均聚类准确率为 85%. 显然，CQHC 算法对线性不可分数据的聚类效果明显优于已有的 2 种算法.

表 3 动物园数据聚类结果

Tab.3 Clustering result of zoo data

标准分类(样本数)	实际聚类(样本数)						
	类 1 (41)	类 2 (20)	类 3 (14)	类 4 (10)	类 5 (7)	类 6 (8)	类 7 (1)
哺乳类(41)	41						
鸟类(20)		20					
鱼类(13)			13				
昆虫类(8)				8			
软体类(10)				2*	7		1*
爬行类(5)			1*			4	
两栖类(4)						4	

图 2 显示了 CQHC 算法对动物园数据聚类过程中函数 CVF 随类别数 c 的变化曲线. 由于动物园数据的线性不可分，导致曲线随类别数 c 的减小变得比较平滑，CVF 函数最小点出现在最佳类别数 $c=7$ 处.

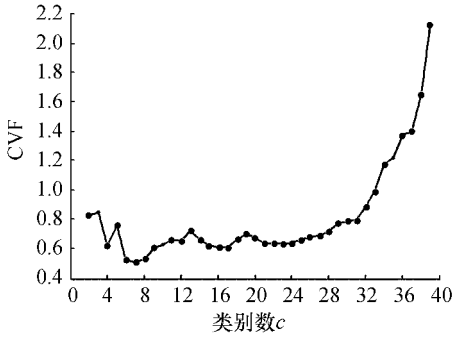

图 2 动物园数据集的 CVF 与类别数 c 关系

Fig. 2 Relationship between CVF value and number of clusters of zoo data

通过上述实验来看，CQHC 算法对线性可分和线性不可分数据均能实现有效聚类，其聚类效果较之先前已有的算法有较大改进，不仅具有更高的聚类准确率，而且能够准确检测出最佳类别数.

4 结论

本文在量子机制的基础上，通过重新计算分类属性数据样本间的相异性度量测度，引入聚类度量尺度步长 β_{step}、紧致性指标 AIAD、离散性指标 AIED 和聚类有效性函数 CVF，结合层次凝聚算法，提出 CQHC 算法，并分别在线性可分数据集和线性不可分数据集中进行了验证. 实验证明，CQHC 算法的聚类效果较之先前已有算法有较大的改进. 下一步将着重研究 CQHC 算法的可伸缩性，以进一步提高其对数据的处理能力.

参考文献：

[1] SANGUTHEVAR R. Efficient parallel hierarchical-clustering algorithms [J]. IEEE Transactions on Parallel and Distributed Systems, 2005, 16(6): 497-502.

[2] HUANG Zhexue, MICHAEL K N. A fuzzy k-modes algorithm for clustering categorical data [J]. IEEE Trans on Fuzzy Systems, 1999, 7(4): 446-452.

[3] HUANG Zhexue. A fast clustering algorithm to cluster very large categorical data sets in data mining [C]//Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. New York: ACM Press, 1997: 1-8.

[4] HUANG Zhexue. Extensions to the k-means algorithm for clustering large data sets with categorical values [J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.

[5] CHEN Ning, CHEN An, ZHOU Long-xiang. Fuzzy k-prototypes algorithm for clustering mixed numeric and categorical valued data [J]. Journal of Software, 2001, 12(8): 1107-1119.

[6] 李志华，王士同. 一种基于量子机制的分类属性数据模糊聚类算法 [J]. 系统仿真学报，2008，20(8)：2119-2122.

[7] 赵正天，赵小强，李 炜. 基于量子机制的改进的分类属性数据聚类算法 [J]. 兰州理工大学学报，2009，35(3)：98-102.

[8] AHMAD A, DEY L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set [J]. Pattern Recognition Letters, 2007, 28(1): 110-118.

[9] KIM D W, LEE K H, LEE D. On cluster validity index for estimation of the optimal number of fuzzy clusters [J]. Pattern Recognition, 2004, 37(10): 2009-2025.

[10] KIM M, RAMAKRISHNA R S. New indices for cluster validity assessment [J]. Pattern Recognition Letters, 2005, 26(15): 2353-2363.

[11] 李志华，王士同. 一种改进的量子聚类算法 [J]. 数据采集与处理，2008，23(2)：211-214.

[12] SUN Y, ZHU QM, CHEN Z X. An iterative initial-points refinement algorithm for categorical data clustering [J]. Pattern Recognition Letters, 2002, 23(7): 875-884.