广义加权 Minkowski 距离及量子遗传聚类算法

钱国红 黄德才 陆亿红
(浙江工业大学计算机科学与技术系 杭州 310023)

摘 要 相异度和相似度度量是聚类算法中非常重要的一种因素,往往会影响到聚类分析的结果。很多聚类算法采用欧式距离作为计算数据相似度的度量。而欧式距离不能反映属性值的全局特性,且不顾及各属性之间的量纲差异,因此当不同属性间具有明显量纲或值域差异时,不能取得很好的效果。对此,提出了一种广义加权 Minkowski 距离,即由各属性的量纲和值域信息来确定各属性的广义权值,既考虑了整个数据集的特性,又消除了各属性之间的不和谐,同时分位数的引进在一定程度上减弱了噪声属性值对距离度量的影响。将提出的新的距离度量用于经典的 k-means 算法和量子遗传聚类算法,实验结果表明,采用新的距离度量和引进量子遗传算法的聚类是更加有效的。

关键词 数据聚类,Minkowski 距离,分位数,全局信息,量子遗传算法
中图法分类号 TP18 **文献标识码** A

General Weighted Minkowski Distance and Quantum Genetic Clustering Algorithm

QIAN Guo-hong HUANG De-cai LU Yi-hong
(Department of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Difference and similarity are very important factor in clustering algorithms, and always affect the results of clustering analysis. A lot of clustering algorithms use Euclidean distance as it's similarity measure. Euclidean distance can't reflect the global information of attributes, and don't consider the unit differences between each attribute, so it can't make a good result when there is obvious unit and domain differences. So, this paper put forward a generally weighted Minkowski distance which is determined by the unit and domain information of each attributes value. Not only characteristics of whole data are considered, but also dicord between attributes is removed, at the same time, using of fractional bits weakens the noise data influence. We used new distance measure in classic k-means. And quantum genetic k-means and the experimental result show that the new algorithm is effective.

Keywords Data clustering, Minkowski distance, Fractional bits, Global information, QGA

1 引言

聚类分析是模式识别、人工智能、数据挖掘等领域重要的研究方向之一,数据聚类在勘探数据内部关系和结构方面有着非常重要的作用,同时也能作为其他工作的预处理过程。迄今为止,学者们对数据聚类进行了大量的研究,并提出了很多优秀的聚类算法[1-5]。聚类算法的基本思想是根据待分样本的相似度将相似的归为一类、不相似的归为另一类。因此,对于具有不同结构特征的数据,如何衡量数据点之间的相异或相似程度对聚类算法起着至关重要的作用,会直接影响到聚类结果的优劣程度。

于是出现了很多关于距离函数的研究和很多新的距离形式,比较常见的形式如下:

对于数据集中任意的两个数据个体 $X_i=(x_{i1},x_{i2},\cdots,x_{in})$ 和 $X_j=(x_{j1},x_{j2},\cdots,x_{jn})$,有:

1. 基于向量 p 范数的距离(Minkowski 距离)

$$d(X_i,X_j)=[\sum_{k=1}^{n}|x_{ik}-x_{jk}|^p]^{\frac{1}{p}} \quad (1)$$

已有研究表明,p 的不同取值会对聚类分析结果产生重要的影响[6,7]。当 $p=2$ 时,Minkowski 距离转化为最常见的欧式距离,但是欧式距离只能发现低维空间球状簇,且对噪声数据敏感;当 $p=1$ 时,演变成 Mahanttan 距离;当 $p=\infty$ 时,转化成 Sup 距离。

分析可知,该类距离并没有考虑各个属性之间的相关性,同样没有考虑属性值量纲的影响,这显然对具有该类特性的数据集的聚类分析具有不利的影响。

2. 马氏距离(Mahalanobis 距离)

$$d(X_i,X_j)=[(X_i-X_j)S^{-1}(X_i-X_j)^T]^{\frac{1}{2}} \quad (2)$$

式中,S 为数据总体的协方差矩阵。

马氏距离是原空间的数据在线性投影空间的欧式距离[8],能够发现椭球形状的簇,也就是说,它能够考虑各个属性之间的相关特性,同时对一切非奇异变换是不变的,不受属

到稿日期:2012-07-11 返修日期:2012-11-23 本文受国家水体污染控制与治理科技重大专项(2009ZX07318-003-01-02),水利部公益性行业科研专项(201001031)资助。

钱国红(1987—),男,硕士,主要研究方向为量子计算、人工智能、数据挖掘;黄德才(1958—),男,博士,教授,博士生导师,主要研究方向为网格调度、人工智能、量子计算、数据挖掘等,E-mail:hdc@zjut.edu.cn(通信作者);陆亿红(1968—),女,硕士,副教授,主要研究方向为程序设计方法、数据库与数据挖掘等。

性值量纲的影响，但是协方差矩阵有可能不可逆以及该距离计算具有较大的时间复杂度成为了致命的弱点，对大规模数据聚类存在不足且难以得到推广。

3. Camberra 距离[9]

$$d(X_i,X_j)=\sum_{k=1}^{n}\frac{|x_{ik}-x_{jk}|}{x_{ik}+x_{jk}} \tag{3}$$

式中，$x_{ik},x_{jk}\geqslant 0$，$x_{ik}+x_{jk}\neq 0$。

Camberra 距离消除了量纲影响，但是只能适用于正实数域，且不考虑属性之间的相关性。因此很多情况下不适用。

4. Alternative 距离

$$d(X_i-X_j)=1-\exp(-\beta\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2) \tag{4}$$

Alternative 距离对噪声数据不敏感[10]，能够在一定程度上消除噪声对聚类效果的影响，使聚类算法具有很好的鲁棒性，但该距离同样没有考虑量纲和属性间相关性。

综上所述，我们发现量纲、属性间相关性、计算时间复杂度以及噪声数据对于聚类分析的结果具有很大的影响，因此选择对这几个方面友好的距离将是改进聚类结果的一种有效的方法。本文针对上述距离的存在缺陷，结合这几个重要特性，提出了一种新的基于极差的广义加权 Minkowski 距离，其将建立在分位数极差基础上的广义权值作用于 Minkowski 距离上，同时考虑数据集的全局特性，并将新的距离度量用于改进 k-means 算法，使之获得更好的性能。另外，针对 k-means 算法采用梯度下降的方法优化代价函数而存在的对初始值敏感、容易陷入局部最优等缺点[11]，本文借鉴 3D 角度编码量子遗传算法[12]充分的量子并行特性、优秀的全局寻优能力，来实现聚类原型的全局寻优，进一步改进算法性能。仿真实验分别对 k-means 算法、引进新距离的 k-means 算法以及进一步引进量子遗传算法的 k-means 进行了比较，验证了新的距离度量和引入量子遗传算法的有效性。

2 基于极差的广义加权 Minkowski 距离

引言中已经讨论了由于量纲、属性相关性、计算时间复杂度、噪声数据因素的存在，使得在数据聚类分析中选取合适的距离度量变得至关重要，而 k-means 算法中采用的欧几里得距离没有充分考虑属性之间的相关性、数据集的全局特性以及各属性量纲的影响，因此欧氏距离对于包含全局空间信息、属性间信息相关的数据集存在巨大的缺陷。本文在此基础上，充分利用全局数据集特性和属性相关性，引进属性值的分位数概念，提出了一种全新的简单有效的距离度量，这里将其命名为基于极差的广义加权 Minkowski 距离。

2.1 相关概念的定义

设 $X_1,X_2,\cdots,X_n$ 为 n 个数据个体，其中 $X_i=(x_{i1},x_{i2},\cdots,x_{im})$，$i=1,2,\cdots,n$。$m$ 为 X_i 的属性个数。则数据总体可以表示为 $G=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，即：

$$G=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1m}\\ x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nm}\end{bmatrix}$$

定义 1(属性值的 q 分位数) 对于给定的数据集总体 G 和实数 $q\in(0,0.5)$，将所有个体的第 j 个属性值 $x_{1j},x_{2j},\cdots,x_{nj}$ 进行排序，得到 $x'_{1j}\leqslant x'_{2j}\leqslant\cdots\leqslant x'_{nj}$，则定义第 j 属性值的 q 分位数 $R_q^j=x'_{[n*q]}$。其中 $[n*q]$ 是对 $n*q$ 取整。

定义 2(属性值的 q 极差) 对于给定的数据集总体 G，给出第 j 个属性值的 q 分数极差(dif_q^j)的定义如下：

$$dif_q^j=|R_q^j-R_{1-q}^j|$$

式中，R_q^j 和 R_{1-q}^j 分别是第 j 属性值的 q 分位数和第 j 属性值的 $(1-q)$ 分位数。显然 dif_q^j 满足对称性，即：

$$dif_q^j=dif_{1-q}^j$$

定义 3(广义权) 如果给定一组权值 $w_1,w_2,\cdots,w_m\in R$，其对于任意 $j\in\{1,2,\cdots,m\}$，满足 $w_j>0$，不必满足 $w_1+w_2+\cdots+w_m=1$ 并且 w_j 不必是无量纲的数值，可以包含特殊的信息，这样的权值在本文中被定义为广义权。

2.2 基于分位数极差的广义加权 Minkowski 距离

定义 4(广义加权 Minkowski 距离) 对于给定的向量集合 G，设 x 与 y 是集合中的任意两个数据，dif_q^j 是数据全体 G 中第 j 个属性值的 q 分位极差，那么：

$$d_G(x,y)=[\sum_j(\frac{1}{dif_q^j}|x_j-y_j|)^p]^{\frac{1}{p}} \tag{5}$$

式中，$j\in\{1,2,\cdots,m\}$ 且 $j\notin\{j\,|\,dif_q^j=0\}$ 是一种距离度量，称之为基于分位数极差的广义加权 Minkowski 距离。

通常而言，数据集上的距离度量函数应该满足对称性和非负性，另外它还应该满足三角不等式，如此该函数才可以定义为一个距离度量，数据集上的相似性函数也有类似的性质。下面给出新距离函数 3 个性质的证明。

证明：

引理 1 对于任意向量 $x=[x_1,x_2,\cdots,x_m]^{\mathrm{T}}$ 和 $y=[y_1,y_2,\cdots,y_m]^{\mathrm{T}}$ 满足下面不等式：

$$(\sum_{i=1}^{m}|x_i+y_i|^p)^{1/p}\leqslant(\sum_{i=1}^{m}|x_i|^p)^{1/p}+(\sum_{i=1}^{m}|y_i|^p)^{1/p}$$

本引理也称为闵可夫斯基不等式，其证明由文献[13]给出。

下面分别证明距离度量需要满足的 3 个必要条件：

(1)对称性。显然可以从式(5)看出，由于

$$|x_j-y_j|=|y_j-x_j|,\ \forall x,y\in G$$

因此 $d_G(x,y)=d_G(y,x)$ 成立。

(2)非负性。从式(5)可以看出每一项均非负，所以非负性成立。而且当且仅当 $x=y$ 时，距离取得最小值 0。

(3)三角不等式。

$$\begin{aligned}d_G(x,z)&=[\sum_j(\frac{1}{dif_q^j}|x_j-z_j|)^p]^{\frac{1}{p}}\\&=[\sum_j|\frac{x_j-y_j+y_j-z_j}{dif_q^j}|^p]^{\frac{1}{p}}\\&=[\sum_j|\frac{(x_j-y_j)}{dif_q^j}+\frac{(y_j-z_j)}{dif_q^j}|^p]^{\frac{1}{p}}\end{aligned}$$

此处正好是 Minkowski 不等式的左边，由引理 1 可得：

$$\begin{aligned}&\leqslant[\sum_j|\frac{(x_j-y_j)}{dif_q^j}|^p]^{\frac{1}{p}}+[\sum_j|\frac{(y_j-z_j)}{dif_q^j}|^p]^{\frac{1}{p}}\\&=d_G(x,y)+d_G(y,z)\end{aligned}$$

因此三角不等式成立。证毕。

由于该距离的本质是一种加权 Minkowski 距离，但其权值之和不必等于 1，且由于权值定义中引进了基于属性值的 q 分位极差，使得距离度量包含了数据集 G 的全局信息，因此也可以称为广义加权 Minkowski 距离。从式(5)可以看出，我们将其广义权值用属性值 q 分位极差的倒数表示，同时对 j 取值进行限制使得其倒数值不可能为 0，显然，倒数权值的引进带来的除法运算消除了量纲的影响；另外，两个 q 分位值

是选自全体数据集，因此能够体现整个数据集的全局特性和消除各个属性之间的不和谐差异性，即如果数据集某个属性值跨度很大，我们的权值正好能够解决距离度量中单个属性影响过大的问题；最后选择一个合适的属性值 q 分位数而不是最值，可以在一定程度上有效避免离群点或者噪声数据的影响，提高算法的鲁棒性。

3 基于广义加权 Minkowski 距离的聚类

3.1 数据聚类问题描述

聚类是空间中包含相对高密度点的连续区域，由相对低密度点区域将其与其他相对高密度点区域分开。总之，聚类 C_i 中包含的数据对象彼此更相近，与其他类中数据对象彼此更不相似。数学定义如下[14]：

假定被研究的数据集为 G，即

$G=\{x_1,x_2,\cdots,x_N\}$

定义 G 的 k 聚类 R，将 G 划分为 k 个集合(聚类)$C_1,C_2,\cdots,C_k$，使其满足下面 3 个条件：

$C_i\neq\phi,i=1,\cdots,k$

$\bigcup_{i=1}^{k}C_i=G$

$C_i\cap C_j=\phi,i\neq j$

3.2 基于广义加权 Minkowski 距离的聚类算法

传统的 k-means 算法是最简洁实用的硬聚类算法之一，它是典型的基于中心点原型的聚类方法的代表，而基于中心点原型的聚类问题的求解往往可以转换成基于目标代价函数的优化问题，它采用所有数据点到相应原型的距离和作为优化问题的目标函数，定义如下：

$$fitness=\sum_{i=1}^{k}\sum_{X\in C_i}dist_G(X,cent^j(i)) \tag{6}$$

式中，$fitness$ 是聚类问题的目标函数，$dist_G(X,cent^j(i))$ 是距离度量公式，$cent^j(i)$ 是第 j 代聚类 C_i 的聚类中心。

采用迭代运算的方式可不断优化聚类问题对应的目标代价函数，以达到算法给定的最高迭代次数或者满足算法的收敛条件，即目标代价函数的极值不再具有明显改变。在本文算法中，采用式(6)定义的适应度函数，并应用文中定义的基于极差的广义加权 Minkowski 距离作为适应度函数中的距离度量公式，求解得到算法收敛时的最优聚类中心 $cent^j(i)$，使得目标函数 $fitness$ 取得最小值。

最后，根据优化得出的聚类中心和本文的距离度量将数据集中的数据分配到相应的簇中。

在本文选取的广义加权 Minkowski 距离中，由于文献[7]中说明了 $p=1$ 时的 Minkowski 距离对噪声数据和离群点具有较好的鲁棒性，故本文取 $p=1$。另外分位数 q 的选取不能太大，取太大极差跨度就很小，不能反映属性值域特性；也不能太小，太小就容易使奇异数据对距离度量产生影响，因此本文给出 q 的一个经验范围 $q\in[0.05,0.2]$，在这个范围内，既包含了大部分数据，又基本去除了一般数据集的奇异数据的影响。

3.3 基于量子遗传改进的聚类算法

经典算法中聚类中心的寻优往往采用梯度下降的方式[11]，该类方式存在对初始值敏感、容易陷入局部最优等缺点，因此本文采用 3D 角度编码的量子遗传算法来改进聚类原型的优化，构造一种量子遗传数据聚类算法，3D 角度编码量子遗传算法充分地利用了量子特性，相比传统的量子遗传算法具有更好的全局寻优能力和更快的寻优效率[12]。以此替换梯度下降的方式能够很好地解决对初始值敏感、容易陷入局部最优等缺点，从而进一步改进聚类效果。下面简单介绍 3D 角度编码量子遗传算法量子并构造相应的量子遗传聚类算法流程。

3.3.1 3D 角度编码量子遗传算法编码方式

由于在 3D 球坐标空间中，量子比特可以由一对相位角 (θ,φ) 唯一确定，因此，我们构造以下形式的编码方式：

$$q_i^t=\left|\begin{matrix}\theta_{i1}^t\\ \varphi_{i1}^t\end{matrix}\right|\cdots\cdots\left|\begin{matrix}\theta_{ij}^t\\ \varphi_{ij}^t\end{matrix}\right|\cdots\cdots\left|\begin{matrix}\theta_{in}^t\\ \varphi_{in}^t\end{matrix}\right| \tag{7}$$

式(7)表示第 t 代种群、第 i 条染色体。$i=1,2,\cdots,m,j=1,2,\cdots,n$。其中 m 是种群规模，n 是量子位数，θ_{ij}^t 和 φ_{ij}^t 是 i 染色体的第 j 量子位的相位角大小。两条基因链 θ 链和 φ 链分别代表一个优化解 θ 解和 φ 解，表示为：

$$q_{i\theta}=(\theta_{i1},\theta_{i2},\cdots,\theta_{in}) \tag{8}$$

$$q_{i\varphi}=(\varphi_{i1},\varphi_{i2},\cdots,\varphi_{in}) \tag{9}$$

3.3.2 解空间映射

记优化问题的第 j 个变量 $X_j\in[a_j,b_j]$，染色体 P_i 第 j 个量子位为 $\begin{pmatrix}\theta_{ij}\\ \varphi_{ij}\end{pmatrix}$，则量子位到 X_j 的解空间变换式为：

$$X_{ij}^{\theta}=a_j+\frac{(b_j-a_j)\theta_{ij}}{\pi} \tag{10}$$

$$X_{ij}^{\varphi}=a_j+\frac{(b_j-a_j)\varphi_{ij}}{2\pi} \tag{11}$$

对于连续优化问题，每条染色体对应优化问题的两个解，分别在 $[0,\pi]^n$ 和 $[0,2\pi]^n$ 尺度空间得到两个最优解 θ 解和 φ 解，然后利用式(10)、式(11)通过线性变换得到相应优化问题的解。可以看出，该方法把任何尺度空间的优化问题都映射到统一的空间来解决，具有很好的适应性。

3.3.3 染色体的更新机制

对于染色体的更新机制，本文充分利用 3D 角度编码的优势，考虑当代染色体和最优染色体的相关度，采用一种自适应的动态调整策略。

具体转角调整方案由式(12)－式(15)确定：

$$\Delta\theta=\frac{w}{|w|}\theta_{\min}+w*(\theta_{\max}-\theta_{\min}) \tag{12}$$

$$w=\frac{f_{\max}^{\theta}-f_x^{\theta}}{f_{\max}^{\theta}} \tag{13}$$

$$\Delta\varphi=\frac{v}{|v|}\varphi_{\min}+v*(\varphi_{\max}-\varphi_{\min}) \tag{14}$$

$$v=\frac{f_{\max}^{\varphi}-f_x^{\varphi}}{f_{\max}^{\varphi}} \tag{15}$$

式中，$\theta_{\min}$ 和 $\varphi_{\min}$ 是旋转角的经验最小值[15] 0.005π，$\theta_{\max}$ 和 $\pi_{\max}$ 是相应的经验最大值 0.1π。其中 $f_{\max}^{\theta}$ 是搜索到的最优个体的适应度，f_x^{θ} 是当前个体的适应度，w、v 分别是式(13)和式(15)给定的一个无量纲的跨度值，描述了当代染色体和全局最优解的差异。差异大，则旋转角度大，能加速地向全局最优解靠拢；差异小，则旋转角小，防止跨过全局最优解。

因此更新操作如式(16)、式(17)所示：

$$\theta_i'=\theta_i+\Delta\theta_i \tag{16}$$

$$\varphi_i'=\varphi_i+\Delta\varphi_i \tag{17}$$

3.3.4 量子位变异机制

对于变异操作，以变异概率 P_m 随机选择量子位，对选中

的量子位进行量子非门变异，对应的角度 θ 和 φ 变为 $\pi/2-\theta$ 和 $\pi/2-\varphi$。

3.3.5 算法流程如下

输入：给定一个数据集 G，聚类数目 K；

输出：最终的聚类结果和聚类中心；

步骤 1(种群初始化) 对聚类中心点进行编码。按照式(7)的角度编码方式以及初始化方式对 k 个中心点进行初始化，每个中心点有 m 个属性值，取量子位数 k * m，随机产生 k 个量子聚类中心。

步骤 2(解空间变换) 按照式(10)、式(11)进行解空间变换，将量子基因链从单位空间转化成优化问题的解空间。

步骤 3(适应度计算) 根据式(6)计算适应度值，评估每条基因链，将最佳适应度对应的解 GX 作为全局最优解。

步骤 4(染色体更新) 按照式(16)、式(17)的方案对每条染色体进行更新，产生更优的聚类中心。

步骤 5(染色体变异) 用 P_m 的变异概率选择量子位进行染色体的变异。

步骤 6(染色体评估) 进行新染色体的解空间变换并计算适应度，记当代最优解为 BX，若 fit(BX)>fit(GX)，则 GX=BX。

a. 由新的染色体产生新的聚类中心。

b. 根据新的聚类中心对数据集重新聚类。

c. 由适应度函数计算出新的适应度。

d. 与保留的精英进行对比，产生新的精英，来指导进一步的寻优过程。

步骤 7(停机条件)

由于聚类问题目标函数最优解未知，因此本文采用连续 10 次适应度函数值没有明显改进作为停机条件，当算法满足收敛条件，转步骤 8；否则，转步骤 4，继续迭代。

步骤 8 按照最优的聚类中心，将数据分配到距离最近的聚类中。

4 对比实验及结果分析

4.1 实验数据与实验环境

为了直观地考察本文所提出的距离度量和算法的性能，实验以 UCI 的 3 个实际数据集作为聚类对象进行聚类分析，3 个 UCI 数据集分别是 Iris、wine 与 glass，具体如表 1 所列。本文在 Windows XP 操作系统下以 matlab7.0.1 作为开发环境，SPSS 作为辅助工具来进行仿真。

表 1 实验数据集描述表

数据集名称	数据个数	数据维度	标准类数
iris	150	4	3
wine	178	13	3
glass	214	9	6

4.2 实验结果和实验分析

仿真实验分别比较了 k-means 算法、采用本文距离度量改进的 k-means 算法以及进一步采用量子遗传算法改进的 k-means 算法。下面分别从分类的正确率(accuracy)、类精度(precision)、召回率(recall) 3 个方面来分析算法的聚类质量。正确率(AC)、类精度(PE)、召回率(RE)分别定义[16]如下：

$$AC=\frac{\sum_{i=1}^{k}a_i}{n},PE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+b_i}}{k},RE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+c_i}}{k}$$

式中，n 表示数据集对象数，a_i 表示正确分到 i 类的对象数，b_i 表示误分到第 i 类的对象数，c_i 表示应该分到第 i 类却没有分到 i 类的对象数，k 表示聚类个数。

基于 3D 角度编码量子遗传聚类算法的参数设置参照文献[12]，对于广义加权 Minkowski 距离，本文取范数 $p=1$，属性值分位数 $q=12.5\%$。分别对 3 个数据集采用传统的 k-means 算法、新距离度量的 k-means 算法(NDK-means)以及进一步由 3D 角度编码量子遗传算法改进的聚类算法(3DQGA_C)运算 50 次。计算其正确率、类精度和召回率的平均值，如表 2—表 4 所列。

表 2 各算法正确率(AC)评估表

数据集名称	K-means	NDK-means	3DQGA_C
iris	81.63%	82.15%	94.71%
wine	87.95%	92.13%	95.83%
glass	51.42%	58.36%	60.57%

表 3 各算法类精度(PE)评估表

数据集名称	K-means	NDK-means	3DQGA_C
iris	82.49%	83.09%	94.17%
wine	88.18%	92.47%	94.82%
glass	53.10%	60.32%	62.47%

表 4 各算法召回率(RE)评估表

数据集名称	K-means	NDK-means	3DQGA_C
iris	80.65%	82.17%	93.72%
wine	85.07%	91.35%	95.06%
glass	49.93%	56.98%	60.29%

从上述表中可以看出：对于 3 个数据集，无论是正确率、类精度，还是召回率，采用本文引进的广义加权 Minkowski 距离的 k-means 算法要比相应的采用欧式距离的 k-means 算法效果要好，而进一步引进量子遗传算法改进的聚类具有最好的聚类效果；同时发现，对于 wine 数据集和 glass 数据集，正确率的改进比较明显，相对来说，因为这两个数据集属性多且属性值域比较多样化，这就使得数据集变得相对复杂，针对这类数据集，采用新的距离能够取到明显的优势。而对于相对简单、全局特性不复杂的 Iris 数据集，新的距离度量也起到了的一定的作用，虽然不是特别明显，但在进一步引入量子遗传算法优化之后能取得较好的聚类效果。从而说明本文提出的新的距离度量和 3D 量子遗传算法在聚类分析中是有效的。

结束语 在分析研究现有距离度量的基础上，针对 k-means 聚类算法采用欧式距离存在的问题和不足，本文提出了一种基于极差的广义加权 Minkowski 距离，新的距离度量不仅能充分利用数据集的全局相关特性、不受属性值量纲的影响，而且新的距离度量计算简洁，没有像马氏距离一样产生过高的计算复杂度，因此新距离度量使得算法具有更好的性能。最后将其分别用于传统的 k-means 算法和用 3D 角度编码量子遗传算法改进的 k-means 算法，仿真结果表明，通过新的距离度量改进的算法比对应的未改进算法具有更好的聚类精度。同时也可以看出，量子遗传算法也具有了经典算法不具备的优势。因此研究新的距离度量及算法融合在数据聚类中具有较大的意义，另外，本文提出的新的距离度量在 q 值的选取上采用统计经验值，如何更加合理地选取 q 将是值得研究的方向。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

[2] Nguyen C D，Cios K J. GAKREM：A novel hybrid clustering al-

gorithm[J]. Information Sciences,2008,178(22):4205-4227
[3] Liu Jing-wei, Xu Mei-zhi. Kernelized fuzzy attribute C-means clustering algorithm [J]. Fuzzy Sets and Systems, 2008, 159 (18):2428-2445
[4] Graves D, Pedrycz W. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study[J]. Fuzzy sets and systems,2010,161(4):522-543
[5] Li M J, Ng M K, Cheung Y-M, et al. Agglomerative fuzzy K-Means clustering algorithm with selection of number of clusters [J]. IEEE Transactions on Knowledge and Data Engineering, 2008,20(11):1519-1534
[6] Bobrowski L, Bezdek J C. c-means clustering with the L_1 and L_∞ norms[J]. IEEE Trans on System, Man and Cybernetics, 1991, 21(3):545-554
[7] Hathaway R J, Bezdek J C, Hu Y. Generalized fuzzy c-means clustering strategies using Lp norm distances[J]. IEEE Trans on Fuzzy Systems,2000,8(5):576-582
[8] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. J of Machine Learning Research,2009,10(2):207-244
[9] 叶斌,胡修林,张蕴玉,等. 基于 3D Zernike 矩的三维地形匹配算法及性能分析[J]. 宇航学报,2007,28(5):1241-1245
[10] Zhang D, ChenS. A comment on alternative c-means clustering algorithms[J]. Pattern Recognition,2004,37(2):173-174
[11] Bezdek J C. A convergence theorem for the fuzzy ISODATA clustering algorithms [J]. IEEE Transactions on Pattern Anal Machine Intel,1980,2(1):1-8
[12] 钱国红,黄德才. 基于 3D 角度编码的量子遗传算法[J]. 计算机科学,2012(8):242-245
[13] 张愿章. Young 不等式的证明及应用[J]. 河南科学,2004,22 (1):23-29
[14] 西奥多里德斯,等. 模式识别[M]. 北京:电子工业出版社,2006
[15] 张葛祥,李娜,金炜东,等. 一种新量子遗传算法及其应用[J]. 电子学报,2004,32(3):476-479
[16] Yang Yi-ming. An evaluation of statistical apporachs to text categorization[J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88

(上接第 200 页)

都高于 Holistic Twig 算法,这主要因为 ProTwigList 算法不需要归并中间结果,且在查询前使用 Tag+Level 流剪枝能提前减少需要处理的节点数。

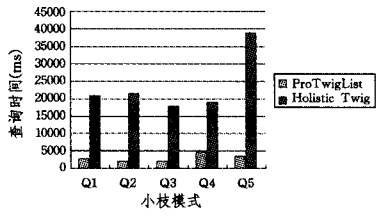

图 4 查询比较(1)

实验 2 文档大小变化,小枝模式固定(以 Q1 为例),对比两算法的查询效率,实验结果如图 5 所示。实验结果表明随着文档大小的逐渐增加,两算法的运行时间都会增加,但 ProTwigList 算法仍优于 Holistic Twig 算法,同样证明了 ProTwigList 算法的查询效率。

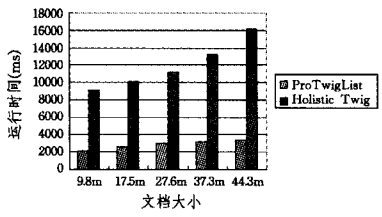

图 5 查询比较(2)

结束语 本文提出了一种非归并不确定 XML 小枝模式查询算法:ProTwigList 算法,该算法提前减少待处理节点的数目,小枝模式查询时使用队列存储可能参与结果的元素节点,输出查询结果并根据最低公共祖先节点计算各结果的概率值。理论分析和实验结果证明了 ProTwigList 算法的性能。下一步将针对包含 NOT、OR 等谓词的不确定 XML 复杂小枝模式查询算法展开研究。

参 考 文 献

[1] 周傲英,金澈清,王国仁,等. 不确定性数据管理技术研究综述[J]. 计算机学报,2009,32(1):1-16
[2] 李文凤,彭智勇,李德毅. 不确定性 Top-k 查询处理[J]. 软件学报,2012,26(3):1-19
[3] Li Jian, Deshpande A. Ranking Continuous Probabilistic Datasets[C]// Proceeding of the 36th International Conference on Very Large Data Bases. Singapore: VLDB Endowment,2010:13-17
[4] Nierman A, Jagadish H V. ProTDB: Probabilistic data in xml [C]// Proceeding of the 28th Very Large Data Bases. Hong Kong: VLDB Endowment,2002:29-41
[5] Kimelfeld B, Sagiv Y. Matching Twigs in Probabilistic XML[C]// Proceeding of the 33th International Conference on Very Large Data Bases. Vienna: VLDB Endowment,2007:23-28
[6] Li Ya-wen, Wang Guo-ren, Xin Juan-chang. Holistically Twig Matching in Probabilistic XML[C]// Proceedings of the 25th International Conference on Data Engineering. Shanghai: IEEE, 2009:1649-1656
[7] Liu Si-qi, Wang Guo-ren. Boosting Twig Joins in Probabilistic XML[C]// Proceeding of the 22ed International Conference on Database and Expert Systems Applications. France: Springer-Verlag,2011:51-58
[8] Ning Bo, Liu Cheng-fei, Yu J X, et al. Matching Top-k Answers of Twig Patterns in Probabilistic XML[C]// Proceeding of the 15th International Conference on Database Systems for Advanced Applications. Tsukuba: Springer-Verlag,2010:125-139
[9] Nierman A, Jagadish H V. Probabilistic data in XML[C]// Proceeding of the 28th international conference on Very Large Data Bases. Hong Kong: Springer-Verlag,2002:646-657
[10] Lu Jia-heng, Meng Xiao-feng, Ling T W. Indexing and querying XML using extended Dewey labeling scheme [J]. Data & Knowledge Engineering,2011,70(1):35-39