一种改进的粒子群和 K 均值混合聚类算法

陶新民[①] 徐 晶[②] 杨立标[①] 刘 玉[①]
[①](哈尔滨工程大学信息与通信工程学院 哈尔滨 150001)
[②](黑龙江科技学院数力系 哈尔滨 150027)

摘 要：该文针对 K 均值聚类算法存在的缺点，提出一种改进的粒子群优化(PSO)和 K 均值混合聚类算法。该算法在运行过程中通过引入小概率随机变异操作增强种群的多样性，提高了混合聚类算法全局搜索能力，并根据群体适应度方差来确定 K 均值算法操作时机，增强算法局部精确搜索能力的同时缩短了收敛时间。将此算法与 K 均值聚类算法、基于 PSO 聚类算法和基于传统的粒子群 K 均值聚类算法进行比较，数据实验证明，该算法有较好的全局收敛性，不仅能有效地克服其他算法易陷入局部极小值的缺点，而且全局收敛能力和收敛速度都有显著提高。

关键词：K均值算法；粒子群优化算法；随机变异；适应度方差
中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2010)01-0092-06
DOI: 10.3724/SP.J.1146.2008.01698

Improved Cluster Algorithm Based on K-Means and Particle Swarm Optimization

Tao Xin-min[①] Xu Jing[②] Yang Li-biao[①] Liu Yu[①]
[①](*College of Information and Communication Engineering*, *Harbin Engineering University*, *Harbin* 150001,*China*)
[②](*Department of Mathematics and Mechanics*, *Heilongjiang Institute of Science and Technology*, *Harbin* 150027,*China*)

Abstract: To deal with the problem of premature convergence of the traditional K-means algorithm, a novel K-means cluster method based on the enhanced Particle Swarm Optimization(PSO) algorithm is presented. In this approach, the stochastic mutation operation is introduced into the PSO evolution, which reinforces the exploitation of global optimum of the PSO algorithm. In order to avoid the premature convergence and speed up the convergence, traditional K-means algorithm is used to explore the local search space more efficiently dynamically according to the variation of the particle swarm's fitness variance. Comparison of the performance of the proposed approach with the cluster method based on K-means, traditional PSO algorithm and other PSO-K-means algorithm is experimented. The experimental results show the proposed method can not only effectively solve the premature convergence problem, but also significantly speed up the convergence.

Key words: K-means algorithm; Particle Swarm Optimization(PSO) algorithm; Stochastic mutation; Fitness variance

1 引言

由 MacQueen 提出的 K 均值算法是解决聚类分析问题的一种经典算法[1,2]，广泛应用于数据挖掘和知识发现领域。由于 K 均值算法的聚类结果依赖于初始值的选取，并且基于梯度下降进行搜索常常使算法陷入局部最优，这两大缺陷极大地限制了它的应用范围。

为了解决这一问题，许多学者进行了大量的研究。例如引入遗传算法等优化技术来解决陷入局部极值的问题[3–5]。但是，实验结果表明当聚类问题的规模较大时，这些算法会出现早熟收敛而陷入局部极值的现象。而且由于进化算法在进化过程中可能产生的退化现象，导致 K 均值算法迭代次数过多，聚类准确率不高，并且可能产生进化后期的波动现象。同遗传算法相比较，PSO[6–9] 不但具有遗传算法的全局寻优能力，通过调整参数，PSO 还可同时具有较强的局部寻优能力，并且 PSO 的参数调整简单易行，适合计算机编程处理，多数情况下比遗传算法更快地收敛于最优解，而且可以避免完全随机寻优的退化现象[10–12]。为此，部分学者结合 PSO 算法对 K 均值算法进行改进。例如，文献[12]提出利用 PSO 算法进行聚类分析，充分利用了 PSO 算法的全局解搜索能力，缺点是没能克服 PSO 算法陷

2008-12-15 收到，2009-06-15 改回
哈尔滨工程大学校科研基金(002080260735)和黑龙江省博士后基金(LBH-Z08227)资助课题
通信作者：陶新民 taoxinmin@hrbeu.edu.cn

入局部最优的缺陷且没能利用 K 均值算法良好的局部搜索能力。进而，文献[13,14]利用在每次 PSO 迭代时对所有粒子进行一次 K 均值计算来寻求最优解，虽然聚类性能有所提高，但该算法仍没能摆脱 PSO 算法陷入局部最优解的困境反而大大增加算法的计算量，使得收敛速度很慢。因此，如何利用 PSO 算法的优点来弥补 K 均值算法的缺点，使两种方法更好的结合，值得我们进一步研究。

鉴于此，本文提出一种改进的粒子群和K均值混合聚类算法，该算法首先利用群体适应度方差来确定K均值算法操作时机，实现PSO算法与K均值算法的有机结合，增强算法局部搜索能力的同时加快算法的收敛速度。为了提高混合算法在初期的收敛速度，在进化过程中增加基于外推方向的粒子位置更新机制，解决了文献[13,14]中聚类算法收敛速度过慢的问题。在PSO算法进化过程中引入随机变异操作，且只对每次参与变异的粒子进行K均值搜索，这样在增强种群多样性的同时不影响算法的收敛速度，进而弥补了文献[12-14]中聚类算法陷入局部最优的缺陷。利用试验对不同实际数据进行测试均显示出新算法的优良性能。

2　改进的粒子群和 K 均值混合聚类算法

2.1 粒子群优化(PSO)算法

粒子群优化算法是一种基于群体智能的进化方法。优化问题的每一个潜在解都是搜索空间中的粒子，每一个粒子有其相应的速度、位置和一个由目标函数决定的适应度，算法通过适应度来评价粒子的优劣。算法首先初始化一群随机粒子，然后通过迭代找到最优解。在每一次迭代中，粒子通过跟踪两个“极值”来更新自己：一个是粒子本身所找到的最优解，即个体极值 pBest；另一个是整个粒子群目前找到的最优解，称之为全局极值 gBest。粒子在找到上述两个极值后，就根据下面两个公式来更新自己的速度与位置:

$$v_i(n+1) = wv_i(n) + c_1 \cdot \mathrm{rand}_1(\,) \cdot (\mathrm{pBest} - p_i(n)) + c_2 \cdot \mathrm{rand}_2(\,) \cdot (\mathrm{gBest} - p_i(n)) \quad (1)$$

$$p_i(n+1) = p_i(n) + v_i(n+1) \quad (2)$$

其中，$v_i(n)$ 是当前粒子的速度，$p_i(n)$ 是粒子的当前位置。$i = 1,2,\cdots,N$，N 是当前空间的维度。$\mathrm{rand}_1(\,)$, $\mathrm{rand}_2(\,)$ 是[0 ,1]之间的随机数，c_1 和 c_2 为学习因子，通常取 c_1=c_2 =1。w 是加权系数，一般在 0.1 到 0.9 之间取值。文献[7]通过大量实验证明，如果 w 随算法迭代的进行而线性减小，将显著改善算法的收敛性能。设 $w_{\max}$ 为最大加权系数，$w_{\min}$ 为最小加权系数，run 为当前迭代次数，runMax 为算法迭代总次数，则有

$$w = w_{\max} - \mathrm{run}\frac{(w_{\max} - w_{\min})}{\mathrm{runMax}} \quad (3)$$

2.2 粒子群和 K 均值混合聚类算法

K 均值聚类算法因其算法简单且收敛速度快等优点，在数据挖掘、图像分割、模式识别、特征提取等诸多领域得到了广泛应用，然而传统的 K 均值算法存在两个固有的缺点：对初始值敏感，易陷入局部最优，PSO 优化算法的出现为解决这一问题提供了新的思路。然而通过对以往文献分析发现，如何能充分利用 PSO 的全局搜索能力以及 K 均值算法精确的局部解搜索能力，且在提高解精度的同时加快算法的收敛速度是提高 PSO+K 均值混合聚类算法的关键所在。

2.3 确定 K 均值算法操作时机

为了将 PSO 算法和 K 均值聚类算法有机的结合，首先要确定 K 均值算法的操作时机。事实上当 PSO 算法在进行全局随机搜索阶段时，不需进行 K 均值算法，这样可以最大程度地利用 PSO 算法向着全局解子空间逼近，加快算法的收敛速度。当 PSO 算法进入收敛状态，此时引入 K 均值算法可以提高解的局部搜索能力，加快算法的收敛速度。因此为了真正实现 PSO 算法与 K 均值算法的有机结合，只需确定粒子群算法何时收敛。

由于 PSO 算法无论是早熟收敛还是全局收敛，粒子群中的粒子都会出现“聚集”现象，此时各粒子的位置一致即各粒子适应度相同。因此，研究 PSO 中所有粒子适应度的整体变化就可以跟踪粒子群的状态，判断算法是否收敛。

设粒子群的粒子数目为 n，f_i 为第 i 个粒子的适应度，f_{avg} 为粒子群目前的平均适应度，粒子群的群体适应度方差 σ^2 的定义如下：

$$\sigma^2 = -\sum_{i=1}^{n}\left(\frac{f_i - f_{\mathrm{avg}}}{f}\right)^2 \quad (4)$$

群体适应度方差反映的是粒子群中所有粒子的“收敛”程度。σ^2 越小，则粒子群趋于收敛；若 σ^2 为零，则群体适应度几乎相同，粒子群优化算法陷入早熟收敛或者达到全局收敛；反之，适应度不同粒子群则处于随机搜索阶段。 因此，利用群体适应度方差的大小来判断 K 均值算法操作时机，当群体适应度方差小于某一个阈值时，说明 PSO 算法进入收敛阶段这时开始执行 K 均值进行局部解精确搜索，这样既提高了混合聚类算法全局解搜索性能，加快了收敛速度，同时解的精度也随之增加。

2.4 提高混合聚类算法初期阶段的收敛速度

为了在迭代初始阶段加快 PSO 算法的收敛速

度，进而加快混合聚类算法的收敛速度，引入外推方向来更新粒子的位置。在粒子尚未达到最优点时，如果迭代后粒子位置的适应度大于粒子当前的适应度，则在它附近产生另一个外推最优粒子，新的外推粒子为

$$p(n+1)=p(n+1)+k(p(n+1)-p(n)) \tag{5}$$

其中，k 为调节系数。由于在初始阶段距最优解较远，调节幅度较大有利于加速进化；后期距最优解较近时，调节幅度较小，以逐步逼近最优解，因此调节系数如下：

$$k=\exp(-20\times(\text{run}/\text{runMax})^{10}) \tag{6}$$

对多变量优化问题，由于每个粒子位置的分量较多，很容易出现某些分量非常接近甚至相同的两个粒子，对此式(5)将不起作用。此时，可在式(5)后加上一个微小的随机数，它只在进化后期起作用以加强微调幅度，位置公式为

$$p(n+1)=p(n+1)+k(p(n+1)-p(n))+10^{-6}\cdot\text{rand} \tag{7}$$

2.5 提高混合聚类算法的全局解勘探能力

由于粒子群和 K 均值混合聚类算法的全局解搜索能力完全依赖于粒子群算法前期的全局解空间的勘探结果。因此，在粒子群中引入随机变异操作来避免 PSO 算法陷入局部极值而早期收敛。由于适应度好的粒子不需变异，因此本文只对适应度差的部分粒子进行随机变异操作，其它优选粒子继续保持原有种群结构进行局部搜索，这样就实现了提高收敛速度和保持种群多样性之间的平衡。具体公式如下：

$$\text{If } (r_i<C_v) \text{ then } v_{id}=r_2\times r_3\times V_{\max}/C_m \tag{8}$$

其中 r_i ($i\in M$)是均匀分布在[0,1]的随机变量，M 是经过适应度排序后较差的部分微粒，r_2 是均匀分布在[0,1]的随机变量，r_3 为随机变量，当随机数小于 0.5 时为 1，大于 0.5 时为-1，控制粒子飞行方向。

通过在粒子群中引入随机变异操作避免了 PSO 算法的早期收敛，进而提高了粒子群和 K 均值混合聚类算法的全局解勘探能力。为了实现聚类算法中全局解能力和收敛速度的平衡，新混合聚类算法只对每次参与变异的粒子进行一次精确的 K 均值搜索，增强了参与随机变异操作的粒子新空间的开采能力，并且由于参与随机变异的粒子数目及概率的限制，对算法的收敛速度影响不大。

3 粒子群和 K 均值混合聚类算法流程

粒子群和 K 均值混合聚类算法采用基于聚类中心的编码方式，也就是每个粒子的位置由 m 个聚类中心组成，粒子除了位置之外，还有速度和适应度。设样本向量维数为 d，因此粒子的位置和速度是 $m\times d$ 维变量，另外每个粒子还有一个适应度 f_i。这样，粒子就可以采用以下的编码结构：$c_1^1c_1^2\cdots c_1^d\cdots c_m^1c_m^2\cdots c_m^dv_1^1v_1^2\cdots v_1^d\cdots v_m^1v_m^2\cdots v_m^df_i$。当聚类中心确定时，聚类的划分由下面的最近邻法则决定，若 x_i,c_j 满足：

$$\|x_i-c_j\|=\min\|x_i-c_k\|,\ k=1,2,\cdots,m \tag{9}$$

则 x_i 属于第 j 类。对于某粒子，按照以下方法计算其适应度：

$$\text{fitness}_i=\sum_{i=1}^{L}\sum_{j=1}^{m}\|x_i-c_{ij}\|^2 \tag{10}$$

其中 L 为样本数，x_i 为输入样本。

具体算法流程如下：

(1)种群的初始化　在初始化粒子时，先将每个样本随机指派为某一类，作为最初的聚类划分，并计算各类的聚类中心，作为初始粒子的位置编码，计算粒子的适应度，同时作为粒子的个体最优位置，并随机初始化粒子的速度。反复进行 N 次，共生成 N 个初始粒子群；

(2)对每个粒子，比较它的适应度和它经历过的最好位置的适应度，如果更好，则更新该粒子的最好位置；

(3)对每个粒子，比较它的适应度和群体所经历的最好位置的适应度，如果更好，则更新全局最好位置；

(4)根据 PSO 算法调整粒子的速度和位置；

(5)利用式(7)生成最优外推粒子；

(6)利用式(8)增加小概率随机变异操作并对参与变异的粒子进行 K 均值操作；

(7)根据式(4)判断当前粒子群是否到达收敛状态，若群体适应度方差 σ^2 小于阈值 threσ，选择 P_m 个最优粒子进行 K 均值局部搜索算法，以此跳出早熟收敛陷阱；

对于选中的粒子，按照以下的 K 均值算法进行优化：

(1)根据粒子的聚类中心编码作为初始值，按照最近邻法则，来确定对应该粒子的聚类划分；

(2)按照聚类划分计算新的聚类中心，取代原来的编码值；由于 K 均值具有较强的局部搜索能力，因此引入 K 均值优化后的粒子群算法的收敛速度可以大大提高；

(3)如果达到结束条件(足够好的位置或最大迭代次数)，判断新聚类中心的适应度，如果比先前的粒子更优，则更新粒子并结束，否则转步骤(2)。

4　试验分析

4.1 对4维IRIS数据聚类性能评价试验

为了测试本文提出的新算法对低维数据的分类性能，采用著名的 IRIS 实际数据作为测试样本集[15]。试验分别采用本文改进的混合聚类算法，K 均值算法，基于 PSO 聚类算法[13]以及传统 PSO-K 均值算法[14]，分析几种算法在迭代过程中的最优解变化情况，以比较它们的聚类性能(K-means 聚类算法不需进行收敛性能的比较)。

试验种群为 10，粒子的维度为 12(4×3)，迭代次数为 1000，w 选择在[0.0962,0.06]之间随迭代次数线性递减，c_1,c_2 均为 1，为了排除算法内部随机操作对分类性能的影响，以 1000 次试验的统计结果进行分析。本文改进的算法中 $C_v=0.8$，$C_m=1$，参与变异操作的粒子群个数 $M=5$，参与 K 均值操作的个数 $P_m=4$，种群适应度方差阈值 $\text{thre}\sigma=0.1$，聚类结果如图 1 所示。

从图 1 中可知，基于 PSO 聚类算法在迭代的开始阶段由于 w 的取值较大，惯性运动占主导作用，因此全局空间的搜索能力较强，最优适应度变化较快，然后随着迭代次数的增加，w 的取值线性减少，算法逐渐进入局部空间搜索阶段，由于丧失了新空间的勘探能力，因此算法陷入局部最优解。PSO-K 均值聚类算法由于在种群中每一个粒子上增加了 K 均值算法使得其在每一个粒子附件又进行了一次精确的局部搜索，增加每一个粒子附近解空间的开采能力，使得算法的全局最优解搜索能力有所提高，进而优于基于 PSO 聚类算法。然而，随着迭代次数的增加，粒子群逐渐趋于收敛，进入了局部搜索阶段，这时 K 均值的作用逐渐减少，起主导作用的是其精确的局部空间搜索能力，全局解空间的勘探能力随着粒子群的聚集而减弱，进而陷入局部最优。与上述两种方法不同的是，本文改进的粒子群-K 均值聚类算法由于在迭代过程中引入了随机变异操作，增强算法全局解的搜索能力，由于变异粒子的出现，使得算法随着迭代次数的增加，不会丧失新空间的勘探能力，避免陷入局部最优的陷阱。同时由于在每一个参与变异操作的粒子中引入了 K 均值算法，使得全局解空间的开采能力进一步加强。因此，改进的算法在最优解搜索能力上都优于其它两种算法。

为了说明改进粒子群 K 均值混合算法的收敛速度以及算法的稳定性能，同样以迭代 1000 次作为停止条件，计算每一次迭代所需要的时间并求其统计值，比较每一次迭代的最优解统计值。试验结果如表 1 所示，K 均值算法的收敛速度最快，但其解的稳定性最差且易陷入局部最优解。由于 PSO 聚类算法每次迭代不需要进行 K 均值操作，其收敛速度较快。然而从最优解适应度的均值及中值可以看出，正是由于没有增强全局解空间的勘探能力使得算法得到的全局最优解较差，即易陷入局部最优解，并且从方差值可以看出，由于算法受到初始解的影响较大，其稳定性能也较差。PSO-K 均值聚类算法由于每一次迭代都引入了 K 均值操作，使得在迭代初始阶段具有较好的全局解搜索能力，遗憾的是算法所消耗的时间较长。同时，由于迭代过程中较强的全局解空间的开采能力使得算法本身降低了对初始解分布的依赖，算法的稳定性较好。与上述两种方法不同的是，由于本文改进的粒子群-K 均值聚类算法增加了随机变异操作，增强了算法的全局解搜索能力，由于每次操作仅仅针对参与变异的部分粒子，使得算法的迭代时间大大减少，同时由于全局解空间勘探性能的提高，使得算法对初始解的依赖减少，算法稳定性能大大提高。

4.2 对高维数据分类性能评价试验

为了测试本文提出的新算法对高维特征数据的聚类性能，利用 4 种算法对来自 UCI 的实际数据进行分类。试验中使用的数据来自 James Cook 大学数学与统计中心[15]。本次试验的种群个数为 30，粒子的维度为 39，迭代次数为 2000，每一次试验独立运行 1000 次，计算统计结果消除随机操作带来的影响。w 选择在[0.0962,0.06]之间随迭代次数线性递减，c_1,c_2 均为 1，本文改进的算法中 $C_v=0.8$，$C_m=1$，参与变异操作的粒子群个数 $M=15$，参与 K 均值操作的个数 $P_m=8$，聚类结果如图 2 所示。从图 2 可知，对于高维特征数据的聚类问题，PSO 聚类算法在算法初期阶段，由于惯性权重在开始时较大，使得粒子具有全局解空间的勘探性能，适应度变化较快；在搜索的中期阶段，由于惯性权重的减少，算法局限于局部空间的开采能力，因此陷入了局部最优解。而 PSO-K 均值聚类算法因引入了 K 均值运算，在算法初始阶段，全局能力较 PSO 算法有所增强，算法得到的最优解优于 PSO 聚类算法，同样到了搜索的中后期，在有限的迭代次数内无法摆脱局部最优解的困扰。而本文的混合算法在开始阶段由于参与随机变异的粒子较少，相对于 PSO-Kmeans 聚类算法来讲最优解下降缓慢，然而在搜索的中后期，由于随机变异操作的引入，使得种群的全局搜索能力伴随着搜索的全部过程，并不会因为迭代次数的增加而消失，因此本文算法在中期阶

段寻到搜索方向后开始下降，逃出局部最优解。

表 2 显示了 4 种方法对高维特征数据进行聚类时所用时间与稳定性的对比结果，从各种算法的均值，中值可以看出，本文算法具有较好的稳定性。由于 PSO 聚类算法在种群初始化和微粒飞行过程中均具有随机性，同时由于 PSO 聚类算法迅速陷入局部最优解使得算法的执行结果也具有随机性。因此，如果能在有限的迭代次数内，改善算法的全局解搜索能力，则算法的稳定性就会提高。这也正是本文算法的稳定性优于其它两种方法的根本原因。从算法所用的时间可以看出，本文算法在提高了全局搜索能力的同时，较好保持了算法的收敛速度。

4.3 对多样本高维数据聚类性能评价试验

为了测试本文提出的新算法对多样本多特征数据的分类性能，利用 4 种方法对乳腺癌实际数据进行分类。试验中使用的数据来自 Wisconsin 大学临床医学中心[15]。

为了排除随机影响，本文取1000次试验的统计结果进行分析，本次试验的种群个数为40，粒子的维度为60，迭代次数为1000，参与变异操作的粒子群个数 $M=20$，参与K均值操作的个数 $P_m=10$，其它参数同上，聚类结果如图3所示。

从试验结果可以看出，PSO聚类算法同样在经历初期的快速下降后陷入了局部最优解。而PSO-K均值聚类算法在初始阶段增加了全局解搜索能力外，在算法的中后期同样难逃陷入局部最优解的困扰。而本文改进的算法由于空间的复杂性，在搜索的中期经历了一段时间的平稳阶段后，跳出局部最优。同样，由于本文的聚类算法在初始阶段参与随机变异的粒子较少，相对于PSO-K均值聚类算法来讲最优解下降较缓慢。对于多样本高维度聚类问题，本文算法的全局搜索能力优于PSO聚类算法及PSO-K均值聚类算法，能够有效地逃出局部最优找到全局最优解，因此算法的稳定性不受算法内部随机操作的影响。如表3所示，本文算法的中值，均值以及标准方差等各项数据均显著优于其它算法，显示出了新算法的稳定性和鲁棒性。

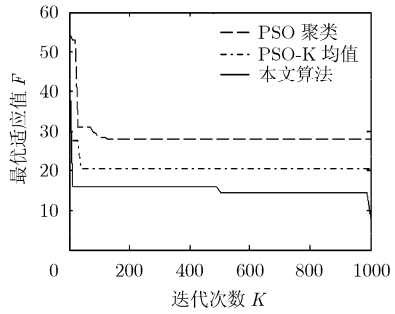

图1 3种算法对IRIS数据聚类性能比较

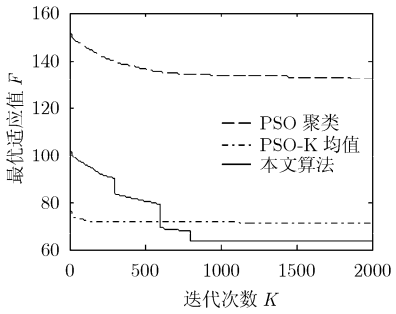

图2 3种算法对UCI高维实际数据聚类性能比较

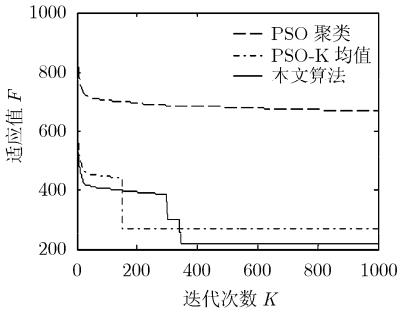

图3 3种算法对多样本高维数据聚类性能比较

表 1 4 种算法对 IRIS 数据聚类结果稳定性与收敛时间比较表

算 法	最大值	最小值	均值	中间值	方差	时间(s)
Kmeans 算法	79.3312	8.07	33.1022	28.1796	9.5668	1.0117
PSO 聚类[13]	61.0973	22.4782	39.8515	40.4224	11.2703	5.8472
PSO-KMeans 聚类[14]	24.4372	7.5353	12.9673	12.9520	4.0006	51.8312
本文算法	20.3020	7.4320	12.5369	12.6139	2.8341	14.3331

表 2 4 种算法对高维数据聚类结果稳定性与收敛时间比较表

算法	最大值	最小值	均值	中间值	方差	时间(s)
Kmeans 算法	279.3150	61.07	140.1022	146.0776	38.5623	13.3126
PSO 聚类[13]	188.2626	98.3892	132.5611	132.0250	19.7191	135.2365
PSO-KMeans 聚类[14]	80.8687	59.7045	71.1389	71.6183	5.1363	1037.4
本文算法	68.43	46.07	64.6272	65.7201	2.0236	189.4255

表 3 4 种算法对多样本高维数据聚类结果稳定性与收敛时间比较表

算法	最大值	最小值	均值	中间值	方差	时间(s)
K-means 聚类	1291.4706	216.7801	701.1236	638.2176	331.1211	20.1367
PSO 聚类[13]	991.4706	434.5537	664.1342	674.4488	117.5258	213.8927
PSO-KMeans 聚类[14]	377.8327	234.6559	271.8983	266.0389	44.9673	3076
本文算法	302.3893	201.0963	215.3282	218.6401	19.6236	434.4048

以上试验数据表明，本文改进的粒子群-K均值聚类算法具有较强的全局解搜索能力，同时保持了算法的收敛速度，实现了两者的平衡。

5　结论

针对 K 均值聚类算法存在的问题，本文提出一种改进的 PSO 和 K 均值混合聚类算法。结合试验得到如下结论：(1)通过利用群体适应度方差判别粒子的收敛程度，确定 K 均值算法的操作时机，实现 PSO 算法与 K 均值算法的有机结合。(2)引入粒子外推位置更新公式，既充分利用了原有 PSO 算法全局解空间的搜索能力，又发挥了 K 均值方法精确的局部空间搜索性能，同时收敛速度较快，所用时间较少。(3)改进算法通过增加随机变异操作，并只对参与变异的粒子进行 K 均值操作，使其在进化过程中始终保持最优解空间勘探的能力，增强了混合聚类算法全局解搜索能力的同时不影响聚类算法的收敛速度。(4)最后，利用不同特征的数据验证了算法的全局解搜索性能以及稳定性，同时与其它 K 均值聚类算法、基于 PSO 聚类算法和基于 PSO-K 均值聚类算法进行比较，试验结果表明本文算法有效实现了全局解搜索能力及收敛速度的平衡。需要指出的是，本文算法中对于参与随机变异操作的粒子个数对聚类算法全局解空间搜索能力和收敛性能的关系尚没有讨论，这也是本课题下一步研究的重点。

参 考 文 献

[1] 陈金山等. 遗传+模糊C-均值混合聚类算法[J]. 电子与信息学报, 2002, 24(2): 210–215.
Chen Jin-shan, *et al.*. A hybrid clustering algorithm incorporating fuzzy c-means into canonical genetic algorithm[J]. *Journal of Electronics & Information Technology*, 2002, 24(2): 210–215.

[2] Li M J and Ng M K, *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.

[3] Krishma K and Murty M N. Genetic Kmeans algorithm[J] . *IEEE Transactions on System, Man and Cybernetics, Part B*, 1999, 29(3): 433–439.

[4] Maulik U and Bandyopadhay S. Genetic algorithm-based clustering technique[J]. *Pattern Recognition*, 2000, 33(9): 1455–1465.

[5] 孟伟, 韩学东. 蜜蜂进化型遗传算法[J]. 电子学报, 2006, 34(7): 1294–1300.
Meng Wei and Han Xue-dong. Bee evolutionary genetic algorithm[J]. *Acta Electronica Sinica*, 2006, 34(7): 1294–1300.

[6] Kennedy J and Eberhart R. Particle swarm optimization[C]. Proceedings of IEEE international conference on neural networks, Perth, Australia, 1995: 1942–1948.

[7] 吕振肃, 侯志荣. 自适应变异的粒子群优化算法[J]. 电子学报, 2004, 32(3): 416–420.
Lu Zheng-su and Hou Zhi-rong. Particle swarm optimization with adaptive mutation[J]. *Acta Electronica Sinica*, 2004, 32(3): 416–420.

[8] Del V Y and Venayagamoorthy G K. Particle Swarm Optimization: Basic concepts, variants and applications in power systems[J]. *IEEE Transactions on Evolutionary Computation*, 2008, 12(2): 171–195.

[9] Van den Bergh F and Engelbrecht A P. A Cooperative approach to particle swarm optimization[J]. *IEEE Transactions on Evolutionary Computation*, 2004, 8(3): 225–239.

[10] 曾建潮, 崔志华. 一种保证全局收敛的PSO算法[J]. 计算机研究与发展, 2004, 41(8): 1333–1338.
Zeng Jian-chao and Cui Zhi-hua. A guaranteed global convergence particle swarm optimizer[J]. *Journal of Computer Research and Development*, 2004, 41(8): 1333–1338.

[11] 赫然, 王永吉. 一种改进的自适应逃逸微粒群算法及实验分析[J]. 软件学报, 2005, 16(12): 2036–2044.
He Ran and Wang Yong-ji. An improved particle swarm optimization based on self adaptive escape [J]. *Journal of Software*, 2005, 16(12): 2036–2044.

[12] 李晓晴, 焦素敏. 基于粒子群优化的带障碍约束空间聚类分析[J]. 计算机工程与设计, 2007, 28(24): 5924–5926.
Li Xiao-qing and Jiao Su-min. PSO spatical clustering with obstacles constraints [J]. *Computer Engineering and Design*, 2007, 28(24): 5924–5926.

[13] Pei Zhenkui and Hua Xia. The clustering algorithm based on particle swarm optimization algorithm[C]. International Conference on Intelligent Computation Technology and Automation (ICICTA), Hunan, Oct. 20–22, 2008, 1: 148–151.

[14] 刘靖明, 韩丽川. 基于粒子群的K均值聚类算法[J]. 系统工程理论与实践, 2005, 6: 54–58.
Liu Jing-ming and Han Li-chuan. Cluster analysis based on particle swarm optimization algorithm[J]. *Systems Engineering-Theory & Practice*, 2005, 6: 54–58.

[15] http://archive.ics.uci.edu/ml/datasets/

陶新民：　男，1973 年生，博士，副教授，研究方向为软计算、随机信号处理、统计检测、免疫算法与网络、信息安全等.

徐　晶：　女，1974 年生，副教授，研究方向为小波及多小波、统计学等.

杨立标：　男，1980 年生，硕士生，研究方向为信号处理、软计算等.

刘　玉：　男，1986 年生，硕士生，研究方向为信号处理、故障诊断.