Journal of System Simulation

Volume 27 | Issue 9 Article 19

8-7-2020

Affinity Propagation Based Improved Group Search Optimizer Clustering Algorithm

Zhang Kang
Key Laboratory of Advanced Control and Optimization for Chemical Process, Ministry of Education East China University of Science and Technology, Shanghai 200237, China;

Xingsheng Gu
Key Laboratory of Advanced Control and Optimization for Chemical Process, Ministry of Education East China University of Science and Technology, Shanghai 200237, China;

Affinity Propagation Based Improved Group Search Optimizer Clustering Algorithm

Abstract

Abstract: The essence of clustering is an optimization problem. It can be solved by swarm intelligent algorithms which are the popular research area in recent years. *A novel Group Search Optimizer (GSO) algorithm named Fast Global Group Search Optimizer (FGGSO) was proposed. FGGSO improved the individuals' updating strategies of GSO, adopting the campaign strategy, destruction-construction strategy and accelerating-jumping strategy. By this means, the proposed algorithm improved the global and local search capability of the original GSO. Furthermore, based on this FGGSO algorithm, a novel improved AP algorithm was proposed. On account of deficiency of AP clustering unable to deal with a user given cluster number, FGGSO and AP were combined. Firstly, AP algorithm was used to obtain any candidate exemplars, and then the clustering result was optimized using FGGSO algorithm, so that a certain cluster number can be obtained*. Experimental results show the effectiveness of the proposed algorithm.

Keywords

clustering analysis, affinity propagation, group search optimizer, optimization

第 27 卷第 9 期 系统仿真学报© Vol. 27 No. 9
2015 年 9 月 Journal of System Simulation Sep., 2015

基于近邻传播的改进组搜索优化聚类算法

张康，顾幸生
（华东理工大学教育部化工过程控制与优化重点实验室，上海 200237）

摘要：聚类问题本质上作为一个最优化问题，理论上是可以使用近年来流行的群智能优化算法来求解的。针对组搜索优化(GSO)算法在全局和局部搜索能力上的不足，*提出了一种新型的——快速全局组搜索优化算法(FGGSO)，采用了竞选策略、破坏——重建策略、加速与跳跃策略。基于该改进的组搜索算法，提出了一种基于近邻传播(AP)算法的改进组搜索优化聚类算法。针对 AP 算法不能设定输出类数的不足，通过将其与 FGGSO 算法结合，先使用 AP 算法得到候选类中心点，再利用 FGGSO 优化聚类结果，得到确定类数的聚类。*实验结果表明所提算法能够获得预期的聚类效果。

关键词：聚类分析；近邻传播；组搜索算法；优化
中图分类号：TP18 文献标识码：A 文章编号：1004-731X (2015) 09-2066-09

Affinity Propagation Based Improved Group Search Optimizer Clustering Algorithm

Zhang Kang, Gu Xingsheng

(Key Laboratory of Advanced Control and Optimization for Chemical Process,
Ministry of Education East China University of Science and Technology, Shanghai 200237, China)

Abstract: The essence of clustering is an optimization problem. It can be solved by swarm intelligent algorithms which are the popular research area in recent years. *A novel Group Search Optimizer (GSO) algorithm named Fast Global Group Search Optimizer (FGGSO) was proposed. FGGSO improved the individuals' updating strategies of GSO, adopting the campaign strategy, destruction-construction strategy and accelerating-jumping strategy. By this means, the proposed algorithm improved the global and local search capability of the original GSO. Furthermore, based on this FGGSO algorithm, a novel improved AP algorithm was proposed. On account of deficiency of AP clustering unable to deal with a user given cluster number, FGGSO and AP were combined. Firstly, AP algorithm was used to obtain any candidate exemplars, and then the clustering result was optimized using FGGSO algorithm, so that a certain cluster number can be obtained.* Experimental results show the effectiveness of the proposed algorithm.

Keywords: clustering analysis; affinity propagation; group search optimizer; optimization

引言

聚类分析的目标是使类内样本相似度较大，类间样本相似度较小，在对样本进行相似性度量后，算法还需要一定的准则函数优化聚类结果使其达到更理想的状态。因此聚类分析本质上是一种优化问题，基于距离的聚类算法则是一种求解该问题的优化算法。近年来，科学家们发明了越来越多的基于群智能的优化方法，典型的有 J. Kennedy 与 R. Eberthart 于 1995 年提出的粒子群优化算法(Particle Swarm Optimization，PSO)[1]以及 M. Dorigo 于 1997

收稿日期：2015-05-15 修回日期：2015-08-03；
基金项目：国家自然科学基金(61174040, 61104178, 61205017)；
作者简介：张康(1984-)，男，山东，博士生，研究方向为人工智能与模式识别；顾幸生，男，教授。

年提出的蚁群算法(Ant Colony Optimizer，ACO)[2]。两种算法分别基于鸟群与蚁群觅食的行为，都采用启发式随机搜索算法，在全局寻优的过程中，不依赖于梯度信息，对目标函数没有特殊要求，在没有集中控制的情况下，依靠群体之间的信息共享来求解复杂问题。群智能算法具有协作性、简单性、分布性、适应性和鲁棒性等特点，使得这些方法为模式识别[3]、函数优化[4]、组合优化[5]、数据聚类[6]、图像处理[7]等复杂问题的求解提供了新的思路。

组搜索优化算法是近年来提出的一种新型的群智能优化算法[8]，它具有较好的全局搜索能力以及高维度问题寻优能力。它基于动物捕食过程中的发现-搜寻(Producer-Scrounger，PS)模型[9]，将整个种群中的个体分为发现猎物的“发现者”和参与捕猎的“参与者”。其中“参与者”中的大部分个体直接向“发现者”周围移动，这一部分个体被称为“跟随者”; 另外一部分个体在自身周围搜寻猎物，它们被称为“游荡者”。每个个体在每次迭代过程中，都以其中一种身份来更新自身位置。近年来，组搜索优化算法被应用于解决各种实际优化问题，例如化工过程建模[10]、建筑结构优化[11]、生产调度[12]等等。为了采用组搜索优化算法进行聚类分析，只需将数据集的聚类准则函数作为组搜索优化算法的适应度函数，将聚类的某一种潜在的划分看作种群中的一个个体，就可以通过组搜索优化算法进行最优聚类划分的搜索。

2007 年，Frey 和 Dueck 在《Science》上提出了近邻传播算法(Affinity Propagation Clustering，APC)，该算法能够处理大规模数据的聚类问题，并且运行时间较短[13]。它是一种基于实例的聚类方法，与经典的 K 均值算法具有相同的目标函数，但在算法原理上该算法与 K 均值算法存在很大的不同。近邻传播算法是一种基于点与点之间信息传递的聚类算法，它将数据集中的每个样本点都视为网络中的一个节点，带有各节点吸引力的消息沿着节点连线递归传播，直到算法找到最优的类代表点集合，使得所有数据点与其最近的类代表点之间的相似度和最大。其中，吸引力信息是某个样本点适合被选为其他样本点的类代表点的程度。相较于传统的聚类算法，AP 算法以数据集的相似度矩阵作为输入，将每个数据点都视为潜在的类代表点，避免了对初始类中心点的选择困难。同时，算法在初始化时无需指定聚类类数，在消息传递的迭代过程中，自动生成聚类的类簇数。另外，AP 算法以具体的数据对象作为类代表点，免去了因为采用平均位置作为类中心的相似度重复计算，在处理大规模多类数据时运算速度快。

本文针对组搜索优化算法运算速度慢，局部搜索能力不强的缺点，采用一种改进的快速全局组搜索优化算法[14](FGGSO)。FGGSO 算法摒弃了原 GSO 算法中的角搜索策略，取而代之的是一种基于竞争与合作机制的竞选策略。并针对“跟随者”和“游荡者”各自的功能和特性，为它们设计了新的搜索与淘汰机制。基于该改进的组搜索算法，提出了一种基于近邻传播的组搜索优化聚类算法。针对 AP 算法不能设定输出类数的不足，通过将其与 FGGSO 算法结合，先使用 AP 算法得到候选类中心点，再利用 FGGSO 优化聚类结果，得到固定类数的聚类。最后，通过多组人工数据集和实际数据集的实验，并与其他聚类算法相比较，证明了本文提出的算法有效性。实验表明，APFGGSO 聚类算法能够聚类出指定数量的类簇，并取得较好的聚类效果。

1 近邻传播算法

近邻传播算法是基于中心点的聚类算法，样本中的每个数据点都被视为可能的类中心点。假设给定数据集中共有 N 个数据点，AP 算法将 N 个数据点两两之间的相似度组成的 $N\times N$ 相似度矩阵作为输入，两点之间的相似度(采用欧式距离为测度时)表示为 $s(i,j)=-\left\|x_i-x_j\right\|^2, i\neq j$，AP 算法用 $s(i,j)$ 表示点 x_i 在多大程度上适合作为数据点 x_i 的类代表点。AP 算法为选出合适的类代表点不断地从数据中搜集相关证据：“可信度”信息和“可

第 27 卷第 9 期 系统仿真学报 Vol. 27 No. 9
2015 年 9 月 Journal of System Simulation Sep., 2015

用度”信息，两种信息流在数据之间传递，代表了不同的竞争目的：一是从点 i 出发到候选代表点 k 的消息流，它为候选代表点 k 从每个数据点 i 处搜集证据 $R(i,k)$，被称作点 k 对点 i 的可信度，用来描述点 k 适合作为点 i 的类代表点的程度；二是从候选类代表点 k 到点 i 的消息流，它为数据点 i 从候选类代表点 k 搜集证据 $A(i,k)$，被称作点 i 对点 k 的可用度，用来描述点 i 选择点 k 作为类代表点的适合程度。$R(i,k)$ 与 $A(i,k)$ 越大，积累的证据越强，则点 k 作为最终聚类中心的可能性就越大。算法通过反复迭代，各样本点之间进行竞争，从而得到最终的聚类中心。算法初始阶段，$A(i,k)$ 和 $R(i,k)$ 都设为 0，两类消息的更新过程如下：

$$R(i,k)\leftarrow s(i.k)-\max_{k's.t.k'\neq k}\{A(i,k')+S(i,k')\} \quad (1)$$

$$A(i.k)\leftarrow \min\left\{0,R(k,k)+\sum_{i's.t.i'\notin\{i,k\}}\max\{0,R(i',k)\}\right\} \quad (2)$$

$$A(k,k)\leftarrow \sum_{i's.t.i'\neq k}\max\{0,R(i',k)\} \quad (3)$$

由公式(1)可以看出，当 $S(k,k)$ 越大使得 $R(k,k)$ 越大时，$A(k,k)$ 也越大，从而点 k 作为聚类中心的可能性也越大。文献[13]设定所有 $S(k,k)$ 为相同的值 p，称作偏向参数，增大或减小 p 的值可以增加或减少 AP 输出的聚类数目。

为防止迭代过程中出现震荡，在 AP 算法每一步迭代中，当前代的可信度 R_t 和可用度 A_t 都要与上一步的 R_{t-1} 和 A_{t-1} 进行加权更新：

$$R(i,k)=\lambda\cdot R_{t-1}(i,k)+(1-\lambda)\cdot R_t(i,k) \quad (4)$$

$$A(i,k)=\lambda\cdot A_{t-1}(i,k)+(1-\lambda)\cdot A_t(i,k) \quad (5)$$

其中 λ 被称作阻尼因子，当 AP 算法发生震荡而不能收敛时，增大 λ 可消除震荡。

2 快速全局组搜索优化算法

2.1 基本 GSO 算法

在 GSO 算法中，个体通过采用三个策略来更新自身位置来搜索新解。在 n 维搜索空间中，$X_i^k\in R^n$ 表示算法运行到第 k 代时，第 i 个个体的位置，将该个体的搜索角记为 $\varphi_i^k=(\varphi_{i1}^k,...,\varphi_{i(n-1)}^k)\in R^{n-1}$，其搜索方向记为 $D_i^k(\varphi_i^k)=(d_{i1}^k,d_{i2}^k,...,d_{in}^k)\in R^n$，其更新公式如式(6)所示：

$$\begin{aligned} d_{i1}^k&=\prod_{q=1}^{n-1}\cos\left(\varphi_{iq}^k\right)\\ d_{ij}^k&=\sin(\varphi_{i(j-1)}^k)\cdot\prod_{q=j}^{n-1}\cos(\varphi_{iq}^k)\quad (j=2,...,n-1)\\ d_{in}^k&=\sin(\varphi_{i(n-1)}^k). \end{aligned} \quad (6)$$

每轮迭代过程中，“发现者”为种群当前适应度最好的个体，其他个体则随机的被选为“跟随者”或“游荡者”。“发现者”在最优位置通过随机旋转搜索角，选取自身位置附近的 3 个方向上的点与当前点比较。若在 3 个方向上能够发现比当前位置更好的点，“发现者”则更新自己的位置；若没有找到适应值更高的点，则随机旋转一个角度，并回到此次迭代的初始位置。

“发现者”的位置确定以后，“跟随者”以随机行走的方式向“发现者”移动，如式(7)所示：

$$X_i^{k+1}=X_i^k+r_3\circ(X_p^k-X_i^k) \quad (7)$$

式中 r_3 为分布在(0,1)区间内的 n 维均匀分布随机序列。

若个体被选为“游荡者”，首先它会选择一个随机的方向，然后，生成一个随机步长，并移动到该点。因为“游荡者”的存在，算法具有了跳出局部最优的能力。为保证算法收敛，种群中的“跟随者”的数量要多于“游荡者”。

2.2 FGGSO 算法

本文采用我们提出的一种快速全局组搜索(Fast Global Group Search Optimizer，FGGSO)优化算法[14]。首先针对基本 GSO 算法角搜索策略搜索速度慢、计算复杂度高的问题，提出了一种新的“发现者”搜索策略——竞选策略。

(1) 竞选策略

竞选是人类社会中存在的一种组织制度，它通过一种自下而上的竞争——淘汰的机制，将社会中

的优秀个体选拔出来作为种群的领导者。若干优秀的候选个体通过竞选的方式选举出更优秀的个体，作为竞选的胜利者来领导种群进步，是一种高效的社会组织形式。在竞选过程中，为了获得种群领导地位，候选个体间不仅存在着相互竞争的关系，同时也存在候选者间相互合作的关系。因为种群中所有个体的目的都是一致的——为了使整个种群社会共同进步。

基于以上理论，提出了采用竞选策略来更新个体位置的方法。在竞选策略中，算法从当前代中选取适应值最好的 K 个个体作为候选个体，本节将竞选行为抽象后，提出了两种搜索机制。

1) 竞争机制

每个优秀个体都渴望成为群体领导者，它们之间存在着竞争关系。本算法中，每个候选个体任意选择另外一个候选个体作为自己的竞争对手，并以随机步长向该个体移动。如式(8)所示：

$$X_{i'}^k = X_i^k + r \cdot (X_j^k - X_i^k), (j \neq i) \tag{8}$$

其中：r 是(0, 1)区间中的均匀随机数，迭代第 k 代时，$X_{i'}^k$ 是第 i 个候选个体移动后的位置，X_i^k 是第 i 个候选个体移动前的位置，X_j^k 是第 j 个候选个体当前位置。计算移动后的个体适应度值，若优于先前位置的值，则该候选个体移动到此位置，并替换之前的适应度值；反之则仍然保持原位置。

2) 合作机制

有竞争就有合作，竞争使个体努力向更好的方向前进，合作则帮助整个种群进步。由 GSO 算法原理可知，“发现者”必是当前 K 个候选个体中的一员，并且是其中的最优个体。合作机制帮助“发现者”搜索其附近的区域，希望能够发现更优的位置。合作机制如式(9)所示：

$$X_{p'}^k = X_p^k + r \cdot (X_{\text{average}}^k - X_p^k) \tag{9}$$

其中：X_p^k 表示第 k 代时“发现者”的位置；X_{average}^k 代表剩余 K-1 个候选个体的平均位置，$X_{p'}^k$ 是移动后的“发现者”的位置。计算 X_{average}^k 和 $X_{p'}^k$ 适应度值，并与 $f(X_p^k)$ 比较，保留更好的并替换“发现者”的位置。从“发现者”的竞选策略中我们可以看出，为了简化了算法结构，FGGSO 算法抛弃了原算法中繁琐的角度——方向模型，“发现者”不再单独的对 3 个方向进行随机搜索，转而在与邻近较优个体的交互中，采用竞争——合作模型发现更优位置。

(2) 破坏——重建策略

其次，为了进一步提高算法的全局搜索能力，并且为了在去除个体搜索角之后，为“游荡者”提供新的更新机制，我们在 FGGSO 中采用了一种应用于“游荡者”的破坏——重建策略。

在基本 GSO 算法中，“游荡者”从始至终都是固定的个数(占种群个体总数的 20%)，这并不能真实地反映动物的觅食行为。当一个动物群体觅食时，若群体在长时间内仍然无法搜索到更丰富的食物源时，更多的个体将会分散开去更大的区域寻找新的食物源。进一步地，若在相当长的时间内，该区域仍然无法搜索到更好的食物源，动物群体会自觉地迁徙到另一个区域，重新开始搜寻食物源。我们将这种行为抽象为算法中的破坏——重建策略，并设计了两种实现机制：

1) 分散机制

算法迭代过程中，若种群的最优适应度值(即“发现者”的适应值)在设定的时间窗口内没有发生显著的改变，则可以认为算法陷入了停滞状态。如式(10)所示：

$$rate = \frac{f(X_p^k)}{f(X_p^{k+N})} < \delta \tag{10}$$

其中：N 表示时间窗口大小；X_p^k 和 X_p^{k+N} 分别是第 k 次迭代和第 k+N 次迭代后“发现者”的位置，适应度的变化率 $rate$ 即为前后适应度的改变比值。若该变化率小于给定的阈值 δ，“游荡者”的个数将会随代数递增，以扩大邻域的搜索范围，有助于找到新的最优位置。为了防止过多的随机搜索导致算法退化，我们将“游荡者”递增的个数上限设定为种群数量的 50%。

2) 重组机制

算法在经历分散机制，扩大“游荡者”比例后，

第 27 卷第 9 期 系统仿真学报 Vol. 27 No. 9
2015 年 9 月 Journal of System Simulation Sep., 2015

若式(10)中的适应度变化率在一定迭代次数内仍未超过给定阈值δ，则可认为算法无法通过分散机制跳出局部最优。此时算法将启动重组机制，以防止过多的“游荡者”导致算法对当前最优位置周围的深度开发能力下降。

为了保留当前种群已经获得的较优位置的信息，算法将在解空间内重新初始化除 K 个当前候选个体外所有的个体。通过重组机制，算法不仅保留了上一代种群中的最优个体，并且扩大了算法的搜索范围。为了保证重组后的种群能适应当前环境，算法将在一定代数内不再激活分散操作，以使得算法能够避免陷入分散——重组——分散频繁的循环中，造成算法性能下降。

(3) 改进个体更新策略

本文前述的两个更新策略都是基于种群层面的改进，如竞选策略是针对种群中最好的 K 个个体的，破坏——重建策略则是针对“游荡者”整体的改进。对于个体层面的位置更新策略，本小节提出两个加快寻优的机制，分别针对“跟随者”与“游荡者”个体。

1) “跟随者”加速机制

根据式(7)所示，“跟随者”以随机步长 $r_3 \circ (X_p^k - X_i^k)$ 移动到“发现者”附近。为了使“跟随者”更快的靠近“发现者”，我们在“跟随者”的更新公式(7)中引入了加速系数 $\omega(\omega>1)$，如式(11)所示：

$$X_i^{k+1} = X_i^k + \omega r_3 \circ \left(X_p^k - X_i^k\right) \tag{11}$$

基于加速搜索机制的“跟随者”将更快的移动到“发现者”周围开始进行深度搜索。

2) “游荡者”跳跃机制

因为 FGGSO 算法抛弃了 GSO 算法中的“角度——方向”模型，“游荡者”不能继续按照旋转一个角度，移动一定距离的方式来进行位置更新，因此我们设计了新的“游荡者”搜索机制，利用均值为 μ 方差为 σ^2 的高斯分布生成“游荡者”步长，如式(12)所示：

$$l_i = a \cdot norm(\mu,\sigma) l_{max} \tag{12}$$

3 基于近邻传播的改进的组搜索优化聚类

3.1 算法原理

最优化问题就是在一定约束条件下，寻找一组参数值，使得最优性度量得到满足。目前最优化方法已经被广泛研究，并应用到生产和科研各个领域。最优化问题可以根据其目标函数、约束条件的性质以及优化变量的选取等分成许多类型，一般地，最优化问题可定义为：

$$\min y = f(x)\, s.t.\, X \in S = \{X \mid g_i(X) \leqslant 0, i = 1,\ldots,m\}$$

其中：$y = f(x)$ 为目标函数；$g_i(X)$ 为约束函数；X 为 n 维优化变量。

由于聚类分析的目标是使类内样本相似度较大，类间样本相似度较小，在对样本进行相似性度量后，还需要一定的准则函数优化聚类结果使其达到更理想的状态，因此聚类分析也是一种最优化问题。本文将数据集的聚类误差函数，视作组搜索算法的适应度函数，将一种潜在的划分看作一个个体，便可以通过前文改进的组搜索优化进行最优聚类的划分。

AP 算法的优点是用户无需指定聚类簇数，算法自动的为输入数据集聚类。但是，通常情况下，算法所得到的聚类类别数往往与数据集固有的类别数不符。前文提到，AP 算法中控制类代表点个数(即聚类簇数)的参数是偏向参数，它的物理意义是表示数据点成为类代表点的概率。一般地，对所有对象的 p 值都设为相同的值，意味着所有对象都有相同的可能性被当做类代表点。因此，p 值越大，结果中的类数越多；p 值越小，结果得到的类数就越少。当用户需要输出指定类数聚类时，只能通过试错法不断调整偏好系数 p 的大小来观察输出结果，这不仅需要用户具有对待聚类数据的先验知识，而且要耗费一定时间。另一方面，AP 算法作为一种基于距离的划分式聚类算法，其本质是优化问题。在给定偏向参数及相似度矩阵的前提下，即

使最终分类数与目标分类数有所不同，AP 算法也能够给出一种满足当前参数条件的最优划分。

因此，为了解决确定分类数条件下的 AP 聚类问题，本节提出了一种 APFGGSO 混合聚类算法：首先利用 AP 算法为数据集做出初始划分，若此时得到的聚类类数 K'与目标类数 K 不同，则将 AP 算法得到的类中心点作为 FGGSO 算法的初始输入，利用 FGGSO 算法继续对数据集进行优化聚类，从而得到满足分类目标数的聚类结果。

3.2 数据编码

本章采用基于聚类中心的编码方式为种群设置解空间。给定一个待聚类数据集 D，目标分类数为 N_c，则每个个体 x_i 的位置是由 N_c 个聚类中心点的向量组成，如式(13)所示：

$$x_i = (m_{i1}, m_{i2}, \ldots, m_{ij}, \ldots, m_{iN_c}) \tag{13}$$

其中：m_{ij} 代表第 i 个个体中的第 j 个聚类的中心位置向量，该中心位置所代表的类簇为 C_{ij}。若数据集 D 中包含 n 个属性，则个体的维数为 $n \times N_c$。因此，个体位置向量的前 n 个值即为第一个聚类中心的 n 个属性值，其他值则以此类推。可以看出，GSO 算法只对 N_c 个类中心点位置进行优化，它所得到的聚类中心点也必然是 N_c 个。

但是，由于经过 AP 算法聚类后的输出为 $K'(K' \neq N_c)$ 个聚类中心点，无法直接作为 FGGSO 算法的输入。因此，本节设计了一种从 K' 个候选类中心点中选取 N_c 个点作为初始类中心点的方法：将从 K' 个点中选取 N_c 个点的所有可能性组合得到的聚类准则函数值作为适应度值，为适应度值由大到小排序。前 N 大(N 为 FGGSO 的种群大小)的适应值代表的数据点组合作为 FGGSO 的初始聚类中心点。这样便可以选择最有可能成为类中心的 N_c 个点作为 FGGSO 算法的初始聚类中心。

每个个体的适应值由式(14)给出：

$$J_k = \sum_{j=1}^{N_c} \left(\sum_{\forall Z_p \in C_{ij}} \frac{s(z_p, m_j)}{|C_{ij}|} \right) \Big/ N_c \tag{14}$$

其中：$s(Z_p, m_j)$ 表示数据点 Z_p 与聚类中心 m_j 之间的相似度；$|C_{ij}|$ 为数据集中属于类簇 C_{ij} 的对象的个数。该式表示了每个数据点到其最近的聚类中心的距离和的平均值。

3.3 算法流程

根据本文提出的基于 AP 的 FGGSO 聚类算法，具体步骤如下：

输入：样本数据集 D，样本总数 n，FGGSO 种群大小 N，聚类数目 K，最大迭代次数 MaxIter；

输出：数据集的 K 个划分，最优适应度值；

初始化：计算数据集 D 的相似度矩阵 d 以及偏好系数 p，p 取所有相似度值的中值；

STEP1：以相似度矩阵 d 和偏向参数 p 为输入，运行 AP 算法，得到预聚类的 K' 个聚类中心点；

STEP2：列出从 K' 个中心点中取出 K 个点的 $C_{K'}^{K}$ 种组合；

STEP3：由式(14)计算所有数据点到每种组合中 K 个点的最大相似度和的平均值，得到该种组合的适应值；

STEP4：对适应值进行由大到小排序，取前 N 个适应值对应的类中心点组合作为 FGGSO 个体的初始位置；

STEP5：以给定参数运行 FGGSO 聚类算法，直到达到最大迭代次数；

STEP6：算法结束。

4 实验与讨论

4.1 实验用数据集

为检验 APFGGSO 算法在实际应用中的有效性，本文选取了 5 个人工数据集和 4 个真实数据集。实验采用的人工数据集由程序自动生成，5 个人工数据集的对象个数、数据维数以及分类目标簇数如表 1 所示。

第 27 卷第 9 期 系统仿真学报 Vol. 27 No. 9
2015 年 9 月 Journal of System Simulation Sep., 2015

表 1 5 个人工数据集

数据集	对象个数	维数	聚类目标数
DS1	800	2	5
DS2	800	2	8
DS3	1200	2	5
DS4	1200	2	8
DS5	1500	2	10

人工数据集考察了不同对象个数、不同聚类数条件下的聚类算法效果。

实验采用的真实数据集来源于 UCI 标准数据库[15]，本文从 UCI 数据库上选取了 4 个真实数据集，如表 2 所示，分别列出了 4 个数据集的对象个数、属性个数、以及聚类目标数。

表 2 4 个 UCI 数据集

数据集	对象个数	属性个数	聚类目标数
Iris	214	4	3
Wine	178	13	3
Breast cancer	699	9	2
Ecoli	336	7	8

4.2 实验结果

本文采用聚类精度 AC(Accuracy)和 ARI 指数(adjusted Rand Index)作为聚类结果的评价指标。聚类精度 AC 表示正确分类的对象个数占总对象数的百分比，$AC \in [0,1]$，AC 越大，表示聚类准确度越高。ARI 指数是评价聚类算法的常用指标，它通过数据点两两之间的关系来评价聚类结果，它能反映聚类的质量以及聚类算法的优劣。$ARI \in [-1,1]$，ARI 指数越大，表示聚类效果越好。为检验 APFGGSO 算法的有效性，分别采用 AP 算法、SAP 算法[16]、本章的 FGGSO 算法以及 APFGGSO 算法进行对比实验。FGGSO 算法为未采用 AP 算法预聚类的本章所提的改进 GSO 算法，初始聚类中心点由解空间内的随机数产生。FGGSO 算法与 APFGGSO 算法均运行 30 次，取最好的一次作为实验结果。FGGSO 算法采用的参数设置如下：候选点个数 K=5，破坏重建的时间窗口设为 50 代，适应值的变化率阈值 δ=1.5，种群大小为 50，迭代次数设为 1 000；APFGGSO 算法与 FGGSO 算法的参数设置相同，AP 算法采用默认的偏好系数设定方法，即取所有点相似度值的中位数。以上算法都是以 C 语言实现，编译器为 GCC 4.7，运行环境为 64 位 Linux 3.11.10.fc18 操作系统，计算机平台为 Intel Core i5-2400 3.10GHz 处理器，2G 内存。4 种算法在 5 个人工数据集上的聚类结果如表 3 所示。

表 3 4 种算法在 5 个人工数据集上的聚类结果比较

数据集	算法			
	AP	SAP	FGGSO	APFGGSO
*DS*1				
AC	/	0.996 2	0.996 2	0.996 2
ARI	0.648 8	0.995 4	0.995 4	0.995 4
聚类数	8	5	5	5
*DS*2				
AC	/	0.9737	0.9700	0.9750
ARI	0.4608	0.9393	0.9298	0.9399
聚类数	16	8	8	8
*DS*3				
AC	/	0.9991	0.9991	0.9991
ARI	0.3206	0.9979	0.9979	0.9979
聚类数	12	5	5	5
*DS*4				
AC	/	0.9816	0.9825	0.9825
ARI	0.7390	0.9633	0.9649	0.9649
聚类数	11	8	8	8
*DS*5				
AC	0.9920	0.9920	0.9927	0.9927
ARI	0.9847	0.9847	0.9862	0.9862
聚类数	10	10	10	10

从表 3 中可以看出，本文所提的 APFGGSO 算法能在所有 5 个人工数据集上取得比其他 3 中算法更好的聚类效果，因为人工数据集的类簇区分度较高，SAP 和 FGGSO 聚类算法在某个数据集上也能达到和 APFGGSO 算法相同的效果。另外，经过 FGGSO 重新优化过的 AP 算法比原始 AP 算法的聚类效果更好，并且能够帮助 AP 算法获得指定类数的结果。4 种算法在 4 个 UCI 数据集上的聚类结果如表 4 所示。

从表 4 中可以看出，默认参数的 AP 算法虽然能够对 4 个数据集成功聚类，但获得的聚类结果却与目标类数相差很大，这是因为 AP 算法的聚类结

果受偏向参数控制，而用户很难将偏向参数与聚类类数一一对应，无法通过快速调节偏向参数来实现目标聚类数目。SAP 算法能够自适应的调节偏向参数，使得 AP 算法能够获得目标类数的结果。FGGSO 算法在 4 个数据集上与 SAP 算法的效果基本一致，这说明单独的 FGGSO 聚类算法已经具有较好的性能。混合 AP 与 FGGSO 的 APFGGSO 算法在 FGGSO 的基础上，采用 AP 算法进行预聚类，将聚类结果作为 FGGSO 个体的初始位置，从 4 个数据集的聚类结果上看，APFGGSO 算法不仅使得 AP 算法能够自动获得目标聚类数，而且聚类效果也好于 SAP 和 FGGSO 算法。特别是对 Ecoli 数据集而言，Ecoli 具有 8 个目标类数和 7 个特征属性，FGGSO 算法需要在 56 维的解空间内搜索最优值，仅能达到 62.2%的准确率。将 AP 算法的聚类结果作为初值的 APFGGSO 算法能够在已有聚类中心的基础上进行再聚类，达到 83.9%的聚类准确率，高于 SAP 和 FGGSO 效果的 30%，这说明本章提出的 APFGGSO 算法能够根据指定类数来指导 AP 算法进行聚类，并且比 AP 算法得到更好的效果。

表 4 APFGGSO 算法与 AP 算法在 4 个 UCI 数据集上的聚类结果比较

数据集	算法			
	AP	SAP	FGGSO	APFGGSO
Iris				
AC	/	0.9133	0.8866	0.9333
ARI	0.3769	0.7720	0.7163	0.8178
聚类数	9	3	3	3
Wine				
AC	/	0.8539	0.9213	0.9494
ARI	0.2689	0.6159	0.7713	0.8537
聚类数	14	3	3	3
Breast cancer				
AC	/	0.9585	0.9585	0.9628
ARI	0.1918	0.8390	0.8390	0.8551
聚类数	31	2	2	2
Ecoli				
AC	/	0.6160	0.6220	0.8393
ARI	0.2251	0.4486	0.5035	0.7463
聚类数	21	8	8	8

5 结论

由于聚类分析的目标是使类内样本相似度较大，类间样本相似度较小，在对样本进行相似性度量后，算法还需要一定的准则函数优化聚类结果使其达到更理想的状态。因此聚类分析本质上是一种优化问题，基于距离的聚类算法则是一种求解该问题的优化算法。组搜索优化算法是近年来提出的一种新型的群智能优化算法，它具有较好的全局搜索能力以及高维度问题寻优能力。将数据集的聚类准则函数作为组搜索优化算法的适应度函数，将聚类的某一种潜在的划分看作种群中的一个个体，就可以通过组搜索优化算法进行最优聚类划分的搜索。

本文首先概述了组搜索优化算法以及近邻传播算法的原理，针对组搜索优化算法运算速度慢，局部搜索能力不强的缺点，采用了我们提出的一种改进的快速全局组搜索优化算法(FGGSO)。FGGSO 算法摒弃了原 GSO 算法中的角搜索策略，取而代之的是一种基于竞争与合作机制的竞选策略。并针对“跟随者”和“游荡者”各自的功能和特性，为它们设计了新的搜索与淘汰机制。然后，本文基于该改进的组搜索算法，提出了一种基于近邻传播的组搜索优化聚类算法。针对 AP 算法不能设定输出类数的不足，通过将其与 FGGSO 算法结合，先使用 AP 算法得到候选类中心点，再利用 FGGSO 优化聚类结果，得到固定类数的聚类。最后，通过多组人工数据集和实际数据集的实验，并与其他聚类算法相比较，证明了本文提出的算法有效性。实验表明，APFGGSO 聚类算法能够聚类出指定数量的类簇，并取得较好的聚类效果。

参考文献:

[1] Eberhart R C, Kennedy J. A new optimizer using particle swarm theory [C]// Proceedings of the Sixth International Symposium on Micro Machine and Human Science. 1995, 1: 39-43.

[2] Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents [J]. Systems, Man, and Cybernetics, Part B, IEEE

Transactions on (S1083-4419), 1996, 26(1): 1-13.

[3] Li R, Shi S, Nian Q, *et al*. Research on Pattern Recognition of Gas Explosion Disaster Risk in Coal Mines Based on PSO-SVM [J]. China Safety Science Journal (S1003-3033), 2013, 23(5): 38-43.

[4] Das S, Suganthan P N. Differential evolution: a survey of the state-of-the-art [J]. Evolutionary Computation, IEEE Transactions on(S1089-778X), 2011, 15(1): 4-31.

[5] Banks A, Vincent J, Anyakoha C. A review of particle swarm optimization. Part II: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications [J]. Natural Computing (S1567-7818), 2008, 7(1): 109-124.

[6] Van der Merwe D W, Engelbrecht A P. Data clustering using particle swarm optimization [C]// The 2003 Congress on Evolutionary Computation, IEEE. USA: IEEE: 2003, 1: 215-220.

[7] Horng M H. Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation [J]. Expert Systems with Applications (S0957-4174), 2011, 38(11): 13785-13791.

[8] He S, Wu Q H, Saunders J R. Group search optimizer: an optimization algorithm inspired by animal searching behavior [J]. Evolutionary Computation, IEEE Transactions on (S1089-778X), 2009, 13(5): 973-990.

[9] Barnard C J, Sibly R M. Producers and scroungers: a general model and its application to captive flocks of house sparrows [J]. Animal Behaviour (S0003-3472), 1981, 29(2): 543-550.

[10] Yan X, Yang W, Ma H, *et al*. Soft Sensor for Ammonia Concentration at the Ammonia Converter Outlet Based on an Improved Group Search Optimization and BP Neural Network [J]. Chinese Journal of Chemical Engineering (S1004-9541), 2012, 20(6): 1184-1190.

[11] Li L, Xu X T, Liu F, *et al*. The group search optimizer and its application to truss structure design [J]. Advances in Structural Engineering (S1369-4332), 2010, 13(1): 43-52.

[12] Cui Z, Gu X. A discrete group search optimizer for hybrid flowshop scheduling problem with random breakdown [J]. Mathematical Problems in Engineering (S1024-123X), 2014.

[13] Frey B J, Dueck D. Clustering by passing messages between data points [J]. Science (S0036-8075), 2007, 315(5814): 972-976.

[14] Zhang K, Gu X S. A Fast Global Group Search Optimizer algorithm [C]// Proceeding of the IEEE International Conference on Information and Automation. USA: IEEE, 2014: 59-64.

[15] Lichman, M. UCI Machine Learning Repository. [DB/OL]. (2015-07-29) [2015-5-15] http://archive.ics.uci.edu/ml.

[16] Zhang J, Tuo X, Yuan Z, *et al*. Analysis of FMRI data using an integrated principal component analysis and supervised affinity propagation clustering approach [J]. Biomedical Engineering, IEEE Transactions on (S0018-9294), 2011, 58(11): 3184-3196.

(上接第 2065 页)

参考文献:

[1] 吴朝全. 一体化联合作战指挥研究[M]. 北京: 解放军出版社, 2005: 79-83.

[2] 王精业. 系统论与作战复杂性[M]. 北京: 黄河出版社, 2006: 233-251.

[3] 韩江洪, 刘小平, 杜兆芳, 等. 基于案例推理的纺纱质量预测模型研究[J]. 系统仿真学报, 2009, 21: 1347-1351.

[4] 袁晓芳. 基于案例推理的煤矿瓦斯预警支持系统研究[D]. 西安: 西安科技大学, 2009.

[5] 徐宁. 基于案例推理的上市公司失败预警[D]. 合肥: 合肥工业大学, 2006.

[6] 周辅疆, 朱小冬, 程永伦, 等. 云模型在训练弹药消耗预测中的应用研究[J]. 舰船电子工程, 2009, 29: 57-59.

[7] 陈震威, 宋志强, 曹明宝, 等, 灰色理论在机场油库油料消耗预测中的应用[J]. 徐州空军学院学报, 2009, 20: 72-74.