

基于全局信息的人工蜂群聚类算法

邓玉婷, 宋 威, 马 伟

(江南大学 物联网工程学院, 江苏 无锡 214122)

摘 要: 针对人工蜂群算法易陷入局部最优和收敛速度慢的不足, 提出了一种基于全局信息的人工蜂群聚类算法。基于全局信息的人工蜂群聚类算法通过加入食物源平均丰富度 (richness), 利用中间聚类效果, 更好地更新食物源; 并且通过引入全局最优信息, 提高跟随蜂的搜索效率, 以获取聚类问题的全局最优解。同时在 UCI 机器学习库的 4 个标准数据集上进行了大量的实验来评估算法的性能。并将该算法和基本人工蜂群算法、粒子群算法和 K-means 算法进行比较。实验结果证明提出的基于全局信息的人工蜂群聚类算法具有更好的性能。

关键词: 人工蜂群算法; 聚类; 群体智能; 搜索策略; 全局信息

中图分类号: TP3

文献标识码: A

文章编号: 1000-7180(2017)02-0020-05

Artificial Colony Clustering Algorithm based on Global Information

Deng Yu-ting, Song Wei, Ma Wei

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: An improved artificial bee colony algorithm was proposed for data clustering aiming at overcoming the shortcomings of being trapped in local optimum and slow convergence rate in this paper, called artificial colony clustering algorithm based on global information (GI-ABC). On the one hand, the modified algorithm added the average richness of food to the updating equation of food source, taking advantage of the intermediate clustering consequence to update the food source better. On the other hand, it utilized global information to improve the search efficiency of the onlooker bees, which contributed to getting the global optimum. Meanwhile, abundant experiments were conducted to evaluate it. In the experiments, four of topical real data sets selected from the UCI Machine Learning Repository were used to test the performance of the strategies compared with the other clustering algorithms, such as ABC algorithm, Particle Swarm Optimization (PSO) algorithm and K-means algorithm. The results indicate that the modified algorithm can generate better results than other algorithms and has better general performance.

Key words: artificial bee colony algorithm; clustering analysis; swarm intelligence; search strategy; global information

1 引言

聚类算法在信息检索、图像分析^[1]、数据挖掘^[2]等领域得到广泛应用。聚类主要是将数据对象调整到不同的类中, 使得同一类中的数据对象彼此尽可能的相似, 不同类中的数据对象尽可能的不同。聚类

分析方法主要有划分方法和层次方法^[3]等。划分方法被认为是比较适合聚类分析的^[4]。K-means 算法是被广泛应用的划分方法之一。

近几年, 一些研究者开始关注生物种群行为。他们根据生物种群行为提出了几种仿生算法, 分别为遗传算法、鱼群算法、蚁群算法、粒子群算法、人工蜂

群算法等等。人工蜂群算法在算法中引入了侦察蜂,并让侦察蜂寻找新的食物源,利用 Deb 规则直接研究蜂群聚类算法的搜索方向的潜在解^[5]。Gao 和 Liu 提出了一种改进的人工蜂群算法来解决全局数值优化问题^[6]。虽然改进的蜂群算法在收敛速率方面优于其他算法,但是找到全局最优解的准确度不高,而且收敛速度比较慢。X Yan 等提出了一种混合人工蜂群算法^[7],为了加强蜜蜂之间的信息交流,引用了遗传算法的交叉算子,但是交叉算子的参数选择比较难定义。

针对基本人工蜂群算法易陷入局部最优和收敛速度慢的不足,本文提出了一种基于全局信息的人工蜂群算法,并将其应用数据聚类分析。

2 聚类性能评价

假定一个数据集 $D(d_1, d_2, \dots, d_n)$ 有 n 个数据对象,每个数据对象有 c 维,每个维度对应一个实数值。因此数据集 D 可以由矩阵 $\mathbf{X}_{n,c}$ 来表示,其中 n 表示矩阵 $\mathbf{X}_{n,c}$ 的行数; c 表示矩阵 $\mathbf{X}_{n,c}$ 的列数;矩阵 $\mathbf{X}_{n,c}$ 中的元素 $x_{i,j}$ 对应数据集 D 中的第 i ($i = 1, 2, \dots, n$) 个数据对象的第 j ($j = 1, 2, \dots, n$) 维特征值。聚类算法主要目标是调整聚类中心 $C = \{C_1, C_2, \dots, C_k\}$ ($C_i \neq \Phi, C_j \cap C_i \neq \Phi, \bigcup_{j=1}^k C_i = D; i, j = 1, 2, \dots, k; i \neq j$) 到不同的类中,使得在相同的类中的数据对象尽可能相似,不同类中的数据对象尽可能不一样。为了度量聚类效果,利用了均方差误差,其计算公式如式(1)所示^[8]:

$$\text{Perf}(D, C) = \sum_{i=1}^n \text{Min}\{ \|d_i - C_j\|^2 (j = 1, 2, \dots, k) \} \quad (1)$$

其中, $\|d_i - C_j\|$ 表示数据对象 d_i 和聚类中心 C_j 的相似度; n 表示数据对象的总数目; k 表示聚类中心总个数。在聚类算法中,最广泛使用的相似度距离是来源于 Minkowski 距离的欧几里得距离,表达式如式(2)所示:

$$\begin{aligned} \text{dist}(d_i, C_j) &= \left(\sum_{n=1}^{\text{Dim}} (x_{in} - C_{jn})^r \right)^{\frac{1}{r}} (\text{令 } r = 2) \\ \Rightarrow \text{dis}(d_i, C_j) &= \sqrt{\sum_{n=1}^{\text{Dim}} (x_{in} - C_{jn})^2} \quad (2) \end{aligned}$$

其中, d_i 表示第 i 个数据对象; C_j 表示聚类中心 C 的第 j 个聚类中心; Dim 表示每个数据对象维数。

针对聚类问题,聚类效果的度量函数 f (或指聚类问题的目标函数) 如式(3)所示:

$$f = \sum_{i=1}^n \text{Min}\{ \text{dist}(d_i, C_j) (j = 1, 2, \dots, k) \} \quad (3)$$

其中, n 表示数据对象的数目, k 表示聚类中心的个数。当函数 f 的值越小,表示相似的数据对象会被分到同一个类中,也说明聚类的效果越好。

3 人工蜂群算法

人工蜂群算法主要由两部分组成:食物源和蜂群,其中蜂群被分为雇佣蜂、侦察蜂和跟随蜂三种个体。雇佣蜂和跟随蜂在蜂群总数中各占一半。雇佣蜂的数目或者跟随蜂的数目与食物源的数目相等。食物源处花蜜的含量(即数量函数)表示食物源的优劣程度(即适应值),计算如式(4)所示:

$$\text{fitness} = \frac{1}{1 + f} \quad (4)$$

其中, f 表示目标函数,它由式(3)计算得到。

假定初始蜂群中雇佣蜂或跟随蜂的数量为 SN(即含有 SN 个解)。对于每个雇佣蜂或者跟随蜂,ABC 算法使用公式(5)来更新旧的食物源,产生新的潜在食物源。

$$v_{ij} = x_{ij} + r_{ij} * (x_{ij} - x_{kj}) \quad (5)$$

其中, $k \in \{1, 2, \dots, \text{SN}\}$ 且 $k \neq i, j \in \{1, 2, \dots, \text{Dim}\}$, (Dim 表示问题维数), k 和 j 都是随机整数。 r_{ij} 是服从均匀分布 $U(-1, 1)$ 的随机函数。新的潜在食物源 v_{ij} 和其对应的旧食物源 x_{ij} , 使用贪婪选择策略确定是否更新食物源,做出最优选择。即选择适应度值较大的食物源,放弃适应度值较小的食物源。

根据食物源的富集程度(或适应值),跟随蜂的选择概率公式如(6)所示:

$$P_i = \frac{\text{fitness}_i}{\sum_{n=1}^{\text{SN}} \text{fitness}_n} \quad (6)$$

其中, P_i 表示第 i 个食物源的选择概率, fitness_i 表示第 i 个食物源的丰富程度,它是通过公式(4)来计算的。

跟随蜂选择食物源后,根据相邻蜜蜂的共享信息,使用公式(5)进行搜索,利用公式(4)为每个食物源计算适应值 fitness,再使用贪婪选择策略淘汰掉较差的食物源。

为了记录某个食物源位置(解)未被更新的次数,引入参数“limit”。若某个食物源位置(解)在循环 limit 次后未被更新,则食物源未被改善,该位置被放弃,对应的雇佣蜂转变成侦察蜂,随机产生一个新的食物源代替旧食物源,公式如(7)所示:

$$x_{ij} = x_{\min,j} + \text{rand}(0, 1)(x_{\max,j} - x_{\min,j}) \quad (7)$$

4 基于全局信息的人工蜂群聚类算法

4.1 算法原理

聚类的主要目标是通过调整聚类中心来适应所属聚类的数据对象,使在聚类中的每一个数据对象被视为问题的一个潜在解.即每个实际要被聚类的数据对象被看作是一个食物源,并且这些数据对象经过特征提取后,数据对象在解空间中的位置不变,即食物源的位置是固定的.因此,可以将聚类问题变成寻找若干个虚拟食物源(即若干个聚类中心),使得每个虚拟食物源到各自所代表的数据对象(即被聚集到一类的食物源)的距离最近.为每个虚拟食物源(即聚类中心)计算一下虚拟食物源周围食物源的平均水平(丰富度),计算公式如(8)所示:

$$\text{richness} = \frac{\sum_{i=1}^N d_i}{N} \quad (8)$$

其中, N 表示属于某个虚拟食物源(聚类中心)的数据对象(食物源)的个数; d_i 表示第 i 个数据对象(食物源)向量.因为随着聚类的迭代次数的增多,聚类的效果会逐渐变好,所以对于每一个雇佣蜂,使用参数 richness,把中间的结果反馈给算法用于寻找潜在解.在雇佣蜂搜索的过程中,如果当代的目标评价函数值优于上一代的目标函数评价价值,则在解更新公式中加入 richness 信息,否则按照公式(5)更新解.一旦目前最好的食物源得到了改善,根据 richness 信息的引导,可以在食物源周围寻找到更好的食物源.这种引导能够被应用于食物源向量的所有维度上,产生新的食物源.新的潜在的食物源的计算公式如(9)所示:

$$v_{ij} = v_{ij} + r_{ij} * (v_{ij} - \text{richness}) \quad (9)$$

其中, r_{ij} 是服从均匀分布 $U(-1, 1)$ 的随机函数.式(9)是在式(5)的基础上操作的,若本次的聚类结果评价得到改善,则先使用式(5),再使用式(9)进行虚拟食物源的更新操作,否则仅使用式(5)来更新虚拟食物源的信息.

跟随蜂根据雇佣蜂的舞蹈选择要跟随的食物源,具有一定的随机性,为了保证跟随蜂能更快地找到食物源周围的更优解,使用全局信息对候选解在某个维度上进行更新,替换原始 ABC 算法中随机抽取一个食物源在某个维度上的更新操作.对于聚类问题,实际的数据对象被视为一个食物源(即一个解),所有聚集在同一个类中的食物源都是聚类中心的邻居,因此可以选择实际的数据对象来更新解,这

个操作主要是维护虚拟食物源的多样性.对跟随蜂中解的更新公式如(10)、(11)所示:

$$v_{ij} = x_{ij} + r_{1ij} * (gp_j - x_{ij}) \quad (10)$$

$$v_{ij} = v_{ij} + r_{2ij} * (d_{kj} - x_{ij}) \quad (11)$$

其中, gp_j 表示全局最好的食物源; r_{1ij} 和 r_{2ij} 是两个随机函数,服从均匀分布 $U[-1, 1]$; d_{kj} 表示第 i 个聚类中第 k 个数据对象的第 j 维分量,而且 k 是一个不同于 i 的随机整数, $k \in \{1, 2, \dots, N\}$ (N 是雇佣蜂或者跟随蜂的数目).

式(5)中,雇佣蜂或跟随蜂进行邻域搜索时,仅仅朝着 $r_{ij}(x_{ij} - x_{kj})$ 的矢量方向迭代更新,并没有对迭代前与迭代后的位置进行优劣比较.跟随蜂在搜索的过程中没有考虑整个蜂群的全局最优位置,只得到了自己的历史最优位置以及当前的位置信息.式(10)引入了全局信息,对数据对象的某一个维度进行更新,而且能够更快搜索到全局最优解.而式(11)对数据对象的所有维度进行更新.这样能够保持式(10)与蜂群算法的等式更新策略一致,维护虚拟食物源的多样性,同时使用式(11)在所有维度上进行更新是加强食物源的特征.

4.2 算法流程

与 ABC 的算法相比,改变了更新公式. GI-ABC 算法的伪代码如下所示:

```

1: 初始化食物源,并且使用式(4)和式(3)为每个食物源计算蜜蜂富集程度;除此之外,与食物源对应的雇佣蜂被分配到相应的食物源上
2: flag = 1;
3: while 不符合结束条件 do
4:   for 每个雇佣蜂 do
5:     基于相邻蜜蜂的共享信息使用式(5)搜索新的潜在食物源
6:     if flag then
7:       使用式(8)计算 richness
8:       使用式(9)产生候选食物源
9:       flag = 0;
10:    end if
11:    使用式(4)和式(3)为每个食物源计算适应值 fitness
12:    对新产生的食物源和其对应的原始食物源使用贪婪选择策略确定是否更新食物源;如果食物源有改善,flag被重置为 1
13:  end for
14:  使用式(6)为每个食物源计算相应的概率  $P$ 
15:  for 每个跟随蜂 do
16:    if rand() <  $P$  then
17:      根据全局最好食物源共享信息使用式(10)、(11)搜索新的食物源并产生潜在的候选解
18:      使用式(4)和式(3)为每个食物源计算适应值 fitness
19:      对新产生的食物源和其对应的原始食物源使用

```

```
贪婪选择策略确定是否更新食物源;如果食物源有改善,flag  
被重置为 1  
20:     end if  
21: end for  
22: if 任何雇佣蜂变成侦察蜂 then  
23:     侦察蜂重新随机产生新的食物源  
24: end if  
25: 记录到目前为止找到的最好食物源  
26: end while
```

5 实验与分析

5.1 实验数据集介绍

为了更好的评估 GI-ABC 数据聚类算法的性能,利用 ABC、PSO 以及 K-means 算法在 UCI^[10]的 4 个标准的数据集上做了测试. 每个数据集的名称、数据大小 N 、聚类中心数目 K 以及维数(或者特征数目) D 总结在表 1 中.

表 1 用于实验的数据集

数据集 名称	数据集 大小 N	聚类中心 个数 K	特征 数目 D
Wine	178	3	13
Glass	214	6	9
Dermatology	366	6	34
Thyroid	215	3	5

Wine 数据集是基于 3 个不同品种产生于意大利同一个地区的葡萄酒化学成分分析的结果,在这 3 个不同品种的葡萄酒里面分析确定了 13 种成分(即特征值)^[9].

Glass 数据集是我们所面临的问题中两外一个比较大的类别(包含 6 个类). 该数据集经常被用在分类问题上,这 6 个类包括:浮动建筑物窗户、非浮动建筑物窗户、车窗、容器、餐具以及前灯,与这六个类分别对应包含 70、76、17、13、9 和 29 条数据记录^[9].

皮肤病 Dermatology 数据集是包含最多类别的数据集之一,共包含 6 个类别,分别是:牛皮癣、脂溢性皮炎、扁平苔癣、玫瑰糠疹、慢性皮炎以及毛发红糠疹^[9].

甲状腺 Thyroid 数据集是用来诊断甲状腺功能是否过高或者过低,数据集每个样本都由 5 个特征值用来区分成为 3 个类别,这三个类别是:甲状腺功能过高、甲状腺功能正常、甲状腺功能过低,数据集一共包含了 215 例样本^[9].

5.2 算法和相关参数设置

所提出来的算法 GI-ABC 与 K-means 算法、群体智能算法中的 PSO、以及 ABC 算法进行对比实

验. 对于 PSO 算法,粒子种群的大小设置为 60,学习因子 c_1 和 c_2 设置为 2.0^[10];惯性权重系数 w 是服从均匀分布 $U(0,1)$ 的随机函数. 对于基本的 ABC 算法与 GI-ABC 算法,参数“limit”被设置为 100. ABC、与 GI-ABC 算法的种群大小被设置为小于或者等于 PSO 算法的种群大小. 在实验中,ABC、GI-ABC 算法的种群大小被设置为 40. 对上述算法,设置最大的迭代次数是 2 500 代. 因为动态确定聚类中心的数目不是本文研究的重点,聚类中心的数目按照数据集的类别预先指定好.

5.3 实验结果与分析

GI-ABC、ABC、PSO 以及 K-means 算法数据聚类的结果如表 2 所示. 这 4 种算法用于聚类分析,在介绍的 4 个数据集上做了对比实验. 平均值 Mean 是在某一数据集上独立运行 30 次所得到类内评价指标 f (见式(3))的平均值. 标准差 Std 反映了 30 次独立运行的实际结果与对应平均值的波动性. 等级评分 Rank 表示每种算法在每个数据库上的性能表现,Rank 的值越小表示算法的性能越好. Rank 评分定义如下:(1)平均值 Mean 的值越大,Rank 的值就越大;(2)如果平均值 Mean 是相等的,那么 Rank 的值将随着标准差的值越大而变大.

对于 Wine 数据集,各个算法性能的顺序是 GI-ABC>ABC>PSO>K-means,在这个数据集上 GI-ABC 算法的优势相对比较的明显. 对于 Glass 数据集而言,4 种算法的性能排序为 GI-ABC>ABC>PSO>K-means,同样在这个数据集上 GI-ABC 算法的优势很明显,要好于其他算法. 对于 Dermatology 数据集而言,算法性能顺序是 GI-ABC>ABC>K-means>PSO. 在这个数据集上 GI-ABC 算法取得了最好的效果,但是在这个数据集上 K-means 算法也取得了不错的效果,而且 K-means 的效果要好于 PSO 的效果.

对于 Thyroid 数据集来说,算法性能顺序是 GI-ABC>ABC>PSO>K-means. 同样 GI-ABC 算法取得了最好效果.

下面讨论 GI-ABC 算法与其它 3 种算法的收敛性问题. 在独立运行 30 次中取得最好性能所对应的收敛性曲线图. 表明 GI-ABC、ABC 以及 PSO 算法的收敛速度基本上差不多,几乎是在同一代开始收敛,并且后面的迭代过程基本重叠. 而 K-means 算法处于局部最小值,一开始就收敛,说明 K-means

表 2 GI-ABC、ABC、PSO 和 K-means 算法在数据集上聚类效果

		GI-ABC	ABC	PSO	K-means
Wine	Mean	1. 629221E+04	1. 629225E+04	1. 629775E+04	1. 710486E+04
	Std	9. 069716E-02	2. 005034E-01	1. 845086E+01	1. 002663E+02
	Rank	1	2	3	4
Glass	Mean	2. 110551E+02	2. 139051E+02	2. 150150E+02	2. 198327E+02
	Std	2. 007799E+00	1. 053288E+00	4. 284396E+00	2. 435166E-01
	Rank	1	2	3	4
Dermatology	Mean	2. 062112E+03	2. 069844E+03	2. 600486E+03	2. 098601E+03
	Std	1. 595352E+00	3. 656927E+00	9. 129002E+01	2. 843418E+00
	Rank	1	2	3	4
Thyroid	Mean	1. 866648E+03	1. 866735E+03	1. 877187E+03	1. 996255E+03
	Std	7. 919978E-01	4. 692489E-01	1. 274241E+01	7. 046760E-01
	Rank	1	2	3	4

算法对初始的聚类中心非常敏感. 4 种算法在 Glass 数据集上的收敛性分析图表明,除了 K-means 的收敛速度最快外,其次就是 GI-ABC 算法,都是在 500 代左右开始收敛. 4 种算法在 Dermatology 数据集上的收敛性分析表明,GI-ABC 算法的收敛速度和 K-means 的相当,比其它算法的收敛速度要快. 4 种算法在 Thyroid 数据集上的收敛性分析表明除 K-means 算法外,其它算法的收敛速度几乎差不多.

6 结束语

基于全局信息的人工蜂群算法引进了全局最优信息来增强跟随蜂的搜索策略,能够更高效地寻找高质量的食物源. 在雇佣蜂的搜索策略中,添加了条件选择来使用聚类的中间结果. 对于聚类问题,中间的聚类结果可以被使用并且正反馈给算法,两两相互促进;另外,GI-ABC 算法虽然多了一步迭代过程,但是其复杂度和 ABC 算法的复杂度是一样的. 综上可得,GI-ABC 算法并没有以时间代价来获得性能上的提高,而且是优于其他三种算法的.

参考文献:

[1] Omran M G H, Salman A, Engelbrecht A P. Dynamic clustering using particle swarm optimization with application in image segmentation[J]. Pattern Analysis and Applications, 2006, 8(4): 332-344.

[2] Evangelou I E, Hadjimitsis D G, Lazakidou A A, et al. Data mining and knowledge discovery in complex image data using artificial neural networks[C]// Proc. Workshop Complex Reason. Geogr. Data, Paphos, 2001,Japan, Sendai, 2001:15-30.

[3] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM computing surveys (CSUR), 1999,

31(3): 264-323.

[4] Zhao Y, Karypis G. Empirical and theoretical comparisons of selected criterion functions for document clustering[J]. Machine Learning, 2004,55(3): 311-331.

[5] Zhang C, Ouyang D, Ning J. An artificial bee colony approach for clustering[J]. Expert Systems with Applications, 2010, 37(7): 4761-4767.

[6] Gao Weifeng, Liu Sanyang. Improved artificial bee colony algorithm for global optimization[J]. Information Processing Letters, 2011,111(17): 871 - 882.

[7] Yan X H, Zhu Y L, Zou W P, et al. A new approach for data clustering using hybrid artificial bee colony algorithm [J]. Neuro computing, 2012, 97 (1): 241-250.

[8] Niknam T. An approach based on particle swarm optimization for optimal operation of distribution network considering distributed generators[C]// Proceedings of the 32nd Annual Conference on IEEE Industrial Electronics, IECON 2006. France, Paris, 2006: 633-637.

[9] Lichman M. UCI machine learning repository[EB/OL]. [2016-06-01]. URL <http://archive.ics.uci.edu/ml>.

[10] Kennedy J, Eberhartr. Particle swarm optimization [C]// Proceedings of IEEE international conference on neural networks. Australia, Party, 1995: 1942-1948.

作者简介:

邓玉婷 女,(1990-),硕士. 研究方向为群体智能、机器学习、数据挖掘. E-mail:dengyuting2015@qq.com.

宋 威 男,(1981-),博士,副教授. 研究方向为数据挖掘、模式识别、信息检索、人工智能.

马 伟 男,(1988-),硕士. 研究方向为进化计算.