跳跃跟踪SSA交叉迭代AP聚类算法

黄 鹤[1,2],李文龙[1,2],杨 澜[1*],王会峰[1],高 涛[1],陈 婷[1]
(1. 长安大学,陕西西安710064;2. 西安市智慧高速公路信息融合与控制重点实验室,陕西西安710064)

摘 要: 针对传统近邻传播聚类算法以数据点对之间的相似度作为输入度量,由于需要预设偏向参数p和阻尼系数λ,算法精度无法精确控制的问题,提出了一种跳跃跟踪麻雀搜索算法优化的交叉迭代近邻传播聚类方法. 首先,针对麻雀搜索算法中发现者和加入者位置更新不足的问题,设计了一种跳跃跟踪优化策略,通过考虑偏好阻尼因子的跳跃策略设计大步长更新发现者,增加麻雀搜索算法的全局勘探能力和寻优速度,加入者设计动态小步长跟踪领头雀更新位置,同时,利用自适应种群划分机制更新发现者和加入者的比重,增加算法的后期局部开发能力和寻优速度;其次,设计基于扰动因子的Tent映射,在此基础上增加3个参数,使映射分布范围增大,并避免了陷入小周期点和不稳周期点;最后,引入轮廓系数作为评价函数,跳跃跟踪麻雀搜索算法自动寻找较优的p和λ,代替手动输入参数,并融合基于扰动因子的Tent映射优化近邻传播算法,交叉迭代确定最优簇数. 使用多种算法聚类University of California Irvine数据集的10种公共数据集,仿真结果表明,本文提出的聚类算法与经典近邻传播算法、基于差分改进的仿射传播聚类算法、基于麻雀搜索算法优化的近邻传播聚类算法和进化近邻传播算法相比具有更优的搜索效率以及聚类精度. 对国家信息数据进行了聚类分析,提出的方法更加准确有效合理,具有较好的应用价值.

关键词: 近邻传播聚类;改进Tent映射;改进麻雀搜索算法;轮廓系数;聚类数据集

基金项目: 国家重点研发计划项目(No.2021YFB2501200);国家自然科学基金面上项目(No.52172379,No.52172324);陕西省重点研发计划项目(No.2021SF-483);陕西省博士后科研项目(No.2018BSHYDZZ64);中央高校基本科研业务费资助项目(No.300102240203)

中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0977-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220209

Jump Tracking SSA Hybrid Iterative AP Clustering Algorithm

HUANG He[1,2], LI Wen-long[1,2], YANG Lan[1*], WANG Hui-feng[1], GAO Tao[1], CHEN Ting[1]
(1. *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*;
2. *Xi' an Key Laboratory of Intelligent Expressway Information Fusion and Control*, *Xi'an*, *Shaanxi* 710064, *China*)

Abstract: Aiming at the problem that the traditional affinity propagation (AP) clustering algorithm takes the similarity between data points as the input measure, and the accuracy of the algorithm cannot be accurately controlled due to the need to preset the preference (p) and the damping coefficient (λ), a jump tracking sparrow search algorithm (JTSSA) optimized hybrid iterative AP clustering method (JTSSA-AP) is proposed. Firstly, in order to solve the problem of insufficient update of the position of the producers and the scroungers in sparrow search algorithm (SSA), a jump tracking optimization strategy is designed. By considering the preference factor, the jump strategy updates the producers in a large step, which increases the global exploration ability and optimization speed of SSA algorithm. The scroungers dynamically track the update position of the leading sparrow in a small step, and uses the adaptive population division mechanism to update the proportion of the producers and the scroungers, which increases the late local development ability and optimization speed of the algorithm. Secondly, on the basis of the original Tent mapping, the disturbance factor is added and three parameters are added, so that the mapping distribution range is increased and the small periodic point and unstable periodic point are avoided. Finally, the silhouette index is introduced as the evaluation function, JTSSA is designed to automatically find better p and λ instead of manual input parameters, the tent map based on disturbance factor optimize the AP clustering, and the optimal number of clusters is determined by hybrid iteration. Multiple algorithms are utilized to cluster the 10 public datasets of the university of California Irvine dataset. Simulation results indicate that the proposed clustering algorithm in this paper ex-

收稿日期:2022-03-01;修回日期:2022-11-02;责任编辑:梅志强
*通讯作者:杨澜

hibits superior search efficiency and clustering accuracy compared to the AP algorithm, the AP clustering algorithm based on differential evolution, the AP clustering algorithm optimized by SSA, and the evolutionary affinity propagation. Cluster analysis is conducted on country data, and the proposed method demonstrates greater accuracy, effectiveness, and rationality, showcasing considerable practical value.

Key words: affinity propagation; improved tent mapping; improved sparrow search algorithm; silhouette index; cluster datasets

Foundation Item(s): National Key Research and Development Program (No.2021YFB2501200); General Program of the National Natural Science Foundation of China (No.52172379, No.52172324); Key Research and Development Program of Shaanxi Province (No. 2021SF-483); Postdoctoral Scientific Research Project of Shaanxi Province (No. 2018BSHYDZZ64); Central Universities Basic Research Funding Projects (No.300102240203)

1 引言

聚类是机器学习领域中的一种统计方法[1],旨在根据不同对象数据之间的内在特征,将相似度较大的数据样本划分到同一簇. 聚类方法应用于诸多领域,如模式识别[2]、图像分割[3]、文献检索[4]等. 传统聚类方法很多,各有局限性. K-Means 聚类需要人工规范初始聚类中心和数目,对噪声和离群点较为敏感[5];模糊 C 均值聚类对初始值的依赖性较高,容易陷入局部极小点[6];基于密度的 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法存在时间复杂度高、不适用于大型数据集等缺点[7]. Frey 等[8]在 Science 上提出了一种近邻传播聚类算法(Affinity Propagation,AP),不需要事先确定聚类数目,相对于以上聚类方法,处理速度较快,目前已被应用到多个领域,如人脸识别、建筑工程、故障诊断与处理等[9 10]. AP 算法中偏向参数 p 和阻尼系数 λ 需要人为设定. 其中,p 决定了聚类效果,越大则聚类出的聚类数目越多,反之亦然;λ 影响聚类过程的震荡程度,增大 λ 会减小聚类速度,减小 λ 则会增大震荡. 近年来利用群智能优化算法[11,12]优化 AP 聚类,确定合适的 p 和 λ,成为研究热点,广受研究学者关注[13~15]. 文献[13]将粒子群算法与 AP 结合,鲁棒性和稳定性得到了提升;文献[14]利用一种改进的果蝇优化算法优化选择 AP 的偏好参数;文献[15]引入烟花爆炸的思想来平衡 AP 的全局和局部搜索能力. 以上算法在避免局部最优和提高全局搜索能力方面做了一些改进,但搜索效率和精度仍有提升空间. 麻雀搜索算法(Sparrow Search Algorithm,SSA)是 2020 年由 XUE[16]提出的一种新的群智能优化算法,模拟了麻雀觅食的过程,具有收敛速度快,适应性强,模型易修改等特点. 因此,本文利用 SSA 在参数寻优中的优点,设计了一种基于跳跃跟踪的 SSA 优化方法(Jump Tracking Sparrow Search Algorithm, JTSSA),并与 AP 算法交叉迭代(JTSSA optimized hybrid iterative AP clustering, JTSSA-AP),对 p 和 λ 进行优化,有效提高了聚类精度与迭代速度.

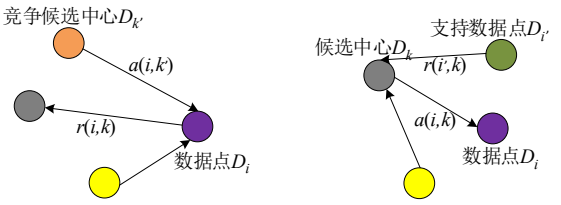
2 算法原理

2.1 AP 算法

AP 算法通过数据点之间传递的消息搜寻聚类中心. 假设对样本集合 $D=\{D_1,D_2,\cdots,D_n\}$ 聚类,首先将所有样本看作潜在的聚类中心,计算 D 中数据点之间的相似度矩阵 $\boldsymbol{S}$ 作为输入,选用欧式距离作为测量指标,则任意两点的相似度为:

$$s(i,k)=-\left\| D_i-D_k \right\|^2,\ (i,k=1,2,\cdots,n;\ i\neq k) \tag{1}$$

$\boldsymbol{S}$ 矩阵对角线之外的元素可通过式(1)计算得到,对角线元素 $s(i,i)$ 的值由 $\boldsymbol{S}$ 矩阵非对角线元素求得,记为偏向参数 p,一般为所有非对角元素数值的中值. 将 $\boldsymbol{S}$ 矩阵作为输入,计算吸引度矩阵 $\boldsymbol{R}$ 和归属度矩阵 $\boldsymbol{A}$. 其中,$r(i,k)$ 反映了 D_k 适合作为 D_i 类中心积累的证据,$a(i,k)$ 反映了 D_i 选择 D_k 作为其类中心合适程度所积累的证据,$\boldsymbol{R}$ 和 $\boldsymbol{A}$ 之间的传递更新过程如图 1 所示.



(a) 吸引度的传递 (b) 归属度的传递

图 1 传递更新过程

近邻传播迭代结束后,D_i 选择 D_k 作为其类中心满足:

$$\arg_k \max(a(i,k)+r(i,k)) \tag{2}$$

初始化 $r(i,k)$ 和 $a(i,k)$=0. 在 AP 的迭代过程中,$r(i,k)$ 和 $a(i,k)$ 的计算公式为:

$$r(i,k)=s(i,k)-\max_{k'\neq k}[a(i,k')+s(i,k')] \tag{3}$$

$$a(i,k)=\begin{cases}\min\{0,r(k,k)+\sum\limits_{i'\neq i}\max[0,r(i',k)]\}, & i\neq k\\ \sum\limits_{i'\neq i}\max[0,r(i',k)], & i=k\end{cases} \tag{4}$$

由于AP聚类过程中易发生振荡,因此添加阻尼系数λ,一般在[0.5,1]中取值. 迭代更新时,$\boldsymbol{R}$和$\boldsymbol{A}$矩阵由当前值和上一步的值加权计算而得. 第t次迭代的计算公式为:

$$\boldsymbol{R}_t=(1-\lambda)\times\boldsymbol{R}_t+\lambda\times\boldsymbol{R}_{t-1} \tag{5}$$

$$\boldsymbol{A}_t=(1-\lambda)\times\boldsymbol{A}_t+\lambda\times\boldsymbol{A}_{t-1} \tag{6}$$

AP算法的具体流程如图2所示.

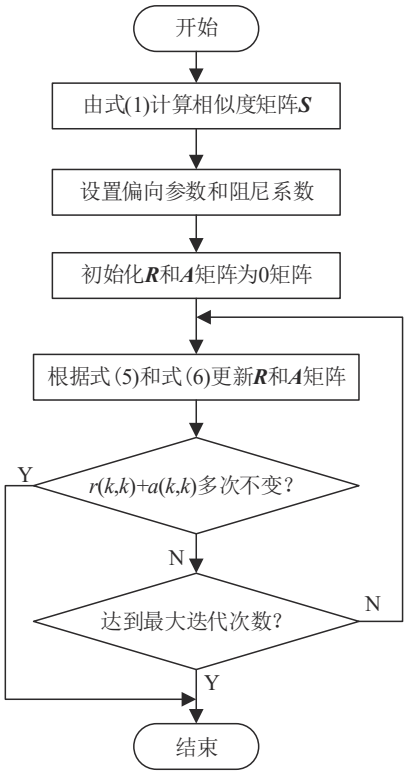
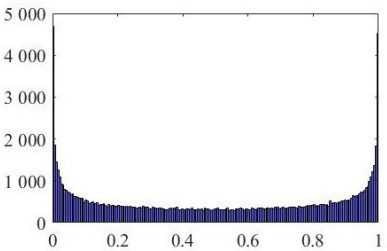


图2　AP算法流程

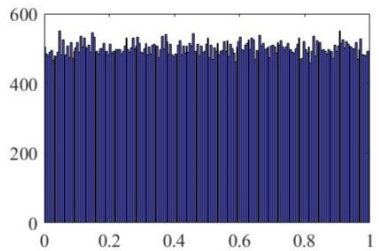
2.2　基于扰动因子的Tent映射

混沌系统是由于内在不确定性而产生外在复杂和随机表现的系统. 随机初始化可在一定程度保证麻雀种群均匀分布,但个别麻雀可能质量较低,而混沌映射产生的序列具有一般性、遍历性和非线性等特点. 因此,利用混沌映射对麻雀种群进行初始化表现更好,可以获得分布均匀,更有利于聚类的初始种群. 常见的映射主要包括Logistic、Tent、Gussian和Circle映射等. 10 000次迭代次数的混沌状态下,上述四种映射空间散布如图3所示. Logistic映射的分布不均匀,其分布为两头大中间小;Tent映射相比其它3种映射分布更均匀;Gussian映射分布呈现出从0到1递减的趋势;Circle与Logistic映射相反,其分布是两头小中间大.

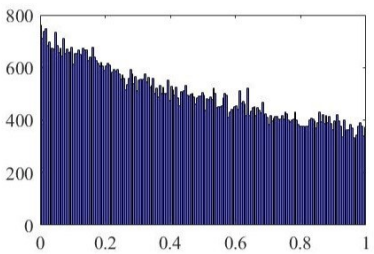
Tent映射分布更均匀,但存在参数少、小周期点、不稳周期点、满映射范围小等缺点. 因此,设计了一种



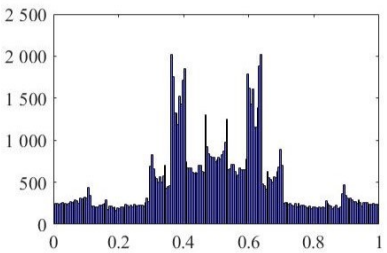
(*a*) Logistic map



(*b*) Tent map



(*c*) Gussian map



(*d*) Circle map

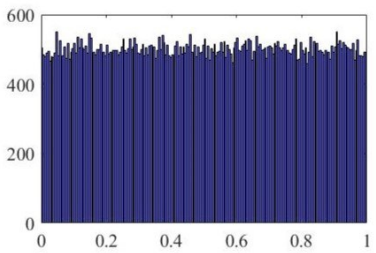
图3　混沌映射空间分布图

基于扰动因子的Tent映射,并增加了2个混沌参数和1个幅度参数,记作RTent,即:

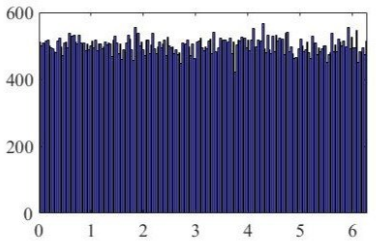
$$\boldsymbol{X}_i=\begin{cases}\psi_1\times\boldsymbol{X}_i+s\times\cos(\boldsymbol{X}_i),0<\boldsymbol{X}_i<\dfrac{\gamma}{2}\\ \psi_2\times(\gamma-\boldsymbol{X}_i)+s\times\cos(\boldsymbol{X}_i),\dfrac{\gamma}{2}\leqslant\boldsymbol{X}_i\leqslant\gamma\end{cases} \tag{7}$$

式中,$\boldsymbol{X}_i$为种群矩阵中第i个个体,s为(0,1)的随机数,$s\times\cos(\boldsymbol{X}_i)$为扰动因子,$\psi_1$,$\psi_2$为混沌参数,$\gamma$为幅度参数. 当$\psi_1$,$\psi_2$保持不变,通过不断增大$\gamma$测试可知,随着$L$增加,满映射幅度的比例也在增加,所以理

论上γ的取值范围可取$(0,+\infty)$；保持γ和ψ_2不变，ψ_1增加，满映射幅度的比例也在增加；保持γ和ψ_1不变，ψ_2增加，满映射幅度的比例不变，但是混沌映射的大小随ψ_2增加而减小. 图4(a)为Tent映射区在区间[0, 1]的分布图，数据标准差为93.48；图4(b)为基于扰动因子的Tent映射RTent，在区间[0, 6.28]的分布柱状图，此时ψ_1=2，ψ_2=1.5，γ=6.28，数据标准差为99.12.



(a) Tent map



(b) RTent map

图4 混沌映射空间分布图

改进后的分布图保留了原来的均匀特性，加入扰动因子则避免了陷入小周期点和不稳周期点状态，而通过增加3个参数能够使映射值分布在更大的范围内，应用场景更加灵活.

2.3 SSA算法

2.3.1 初始化

SSA[16]中麻雀种群分为觅食发现者，抢食物的加入者，两者角色可以互换. 同时还选择了一定比例的麻雀作为侦察的警戒者，一有危险便飞向别处. 麻雀种群在d维空间内n只麻雀的位置矩阵$\boldsymbol{X}$及相应的适应度矩阵$\boldsymbol{F}_X$表示如下：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \tag{8}$$

$$\boldsymbol{F}_X=\begin{bmatrix} f([x_{11},x_{12},\cdots,x_{1d}]) \\ f([x_{21},x_{22},\cdots,x_{2d}]) \\ \cdots \\ f([x_{n1},x_{n2},\cdots,x_{nd}]) \end{bmatrix} \tag{9}$$

式中，x_{ij}为j维空间内第i只麻雀的位置，$f([x_{i1},x_{i2},\cdots,x_{id}])$表示第$i$麻雀的适应度值.

2.3.2 麻雀位置更新机制

更新机制主要可分为发现者的觅食过程，加入者的抢食过程和警戒者的检测过程如下.

(1)觅食过程：迭代寻优过程中，麻雀种群中的发现者负责觅食和指导整个种群移动，发现者位置更新如下：

$$\boldsymbol{X}_i^{t+1}=\begin{cases} \boldsymbol{X}_i^t\cdot\exp\left(\dfrac{-i}{\alpha\times\mathrm{iter}_{\max}}\right), R_2<\mathrm{ST} \\ \boldsymbol{X}_i^t+Q\times\boldsymbol{L}, R_2\geqslant\mathrm{ST} \end{cases} \tag{10}$$

式中，$\boldsymbol{X}_i^t$表示种群中第t代中第i个个体，α为(0,1)中的均匀随机数，$\mathrm{iter}_{\max}$为SSA最大迭代次数，Q为一个标准正态分布随机数，$\boldsymbol{L}$是一行多维的全一矩阵. R_2为[0, 1]区间上的均匀随机数，ST为警戒阈值，一般取值为0.6. 可以看出，当R_2<ST，发现者的每一维都在缩小，这对求解测试函数非常有效，但对于大多数求解非0的实际应用反而降低了搜索能力.

(2)抢食过程：加入者为除去发现者适应度较差的一些个体，设计位置更新公式如下：

$$\boldsymbol{X}_i^{t+1}=\begin{cases} Q\cdot\exp\left(\dfrac{\boldsymbol{X}_{\mathrm{w}}^t-\boldsymbol{X}_i^t}{i^2}\right), & i>\dfrac{n}{2} \\ \boldsymbol{X}_{\mathrm{P}}^t+\left|\boldsymbol{X}_i^t-\boldsymbol{X}_{\mathrm{P}}^t\right|\cdot\boldsymbol{E}^+\cdot\boldsymbol{L}, & i\leqslant\dfrac{n}{2} \end{cases} \tag{11}$$

式中，$\boldsymbol{X}_{\mathrm{P}}^t$是第$t$代种群发现者占据的最佳位置，$\boldsymbol{X}_{\mathrm{w}}^t$表示当前全局最差位置，$\boldsymbol{E}$为只含1或−1元素($1\times d$)的矩阵，定义矩阵运算$\boldsymbol{E}^+$的计算方式如下：

$$\boldsymbol{E}^+=\boldsymbol{E}^{\mathrm{T}}\left(\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\right)^{-1} \tag{12}$$

(3)检测过程：在麻雀种群随机选取一定比例的个体作为警戒者，则这些个体对全体的影响如下：

$$\boldsymbol{X}_i^{t+1}=\begin{cases} \boldsymbol{X}_{\mathrm{B}}^t+\beta\cdot\left|\boldsymbol{X}_i^t-\boldsymbol{X}_{\mathrm{B}}^t\right|, & f_i\neq f_{\mathrm{G}} \\ \boldsymbol{X}_i^t+\xi\cdot\left(\dfrac{\boldsymbol{X}_i^t-\boldsymbol{X}_{\mathrm{B}}^t}{\left|f_i-f_{\mathrm{W}}\right|+\varepsilon}\right), & f_i=f_{\mathrm{G}} \end{cases} \tag{13}$$

式中，$\boldsymbol{X}_{\mathrm{B}}^t$表示当前全局最优位置；$\beta$是符合正态分布的步长控制参数，均值为0，方差为1；ξ是[−1, 1]的随机数；f_i表示麻雀当前位置的适应度；f_{G}与f_{W}分别为全局最优和最差适应度；ε为不为0的极小值，设置为10^{-8}，防止分母为0.

2.4 总体方案

为改善传统AP聚类算法以数据点对之间的相似度作为输入度量，需要预设偏向参数p和阻尼系数λ，算法精度无法精确控制的问题，利用改进麻雀搜索算法寻优，其中轮廓系数为连接改进麻雀搜索算法和AP算法的桥梁，如图5为总体方案框图.

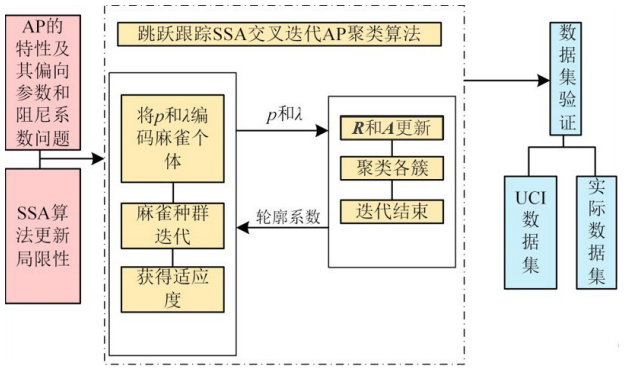


图5　总体框架图

3　基于跳跃跟踪的SSA优化策略

麻雀的后肢较短,胫部跗骨和跗部趾骨之间没有关节曰,不能弯曲,没有能力用后肢在平地上行走,只能做频繁地跳跃运动,因此,设计改进SSA的收敛方式应该是跳跃式的,有利于优化AP聚类时寻找最优的偏好参数p. 但SSA中的麻雀位置更新具有向原点收敛的趋势,对于求解非原点问题存在一定局限性,而p一般不取0值. 这里需要结合跳跃式收敛,对SSA作进一步改进.

3.1　考虑偏好阻尼因子的跳跃策略

将AP聚类的偏好参数和阻尼系数结合,设计成SSA算法跳跃策略的偏好阻尼因子,这样通过改进SSA与AP参数的交叉迭代来优化AP聚类算法,进一步提升效率. 由更新式(10)可知,发现者的位置更新存在收敛于0的趋势,而加入者和警戒者的收敛趋于最优解,这导致了SSA寻找最优解时易收敛于0,而AP的偏好参数和阻尼系数的值不为0,SSA在寻优时易远离最优解从而陷入局部最优无法跳出. 这里设计考虑偏好阻尼因子的跳跃策略更新发现者来完善SSA算法,增加全局勘探能力. 首先在发现者更新公式加入大步长更新策略,使迭代更新时发现者能在寻优区间内大步长的全局勘探;同时为增加算法的寻优速度,引入偏好阻尼因子,设计为初始化麻雀中最优的麻雀. 这里新的发现者更新公式修正为:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}\dfrac{1}{\rho}\times\boldsymbol{X}_i^t\cdot\exp\left(\dfrac{-i}{\alpha\times\mathrm{iter}_{\max}}\right)+\rho\times\omega\times\mathrm{Gbest0}, R_2<\mathrm{ST}\\ \boldsymbol{X}_i^t+\rho\times\mathrm{rand}(\mathrm{Lb},\mathrm{Ub}), R_2\geqslant\mathrm{ST}\end{cases}\tag{14}$$

$$\rho=\exp\left(\frac{-t}{|Q|\times\mathrm{iter}_{\max}}\right)\tag{15}$$

式中,ρ为权重因子,ω为(−1,1)上均匀分布的随机数,Gbest0为初始麻雀种群中的最优麻雀,Lb和Ub分别为麻雀种群的上下界.

3.2　动态跟踪领头雀策略

发现者的更新采用大步长的机制,不利于SSA的后期局部开发. 算法中加入者追随发现者向最优解迭代,但加入者的更新为标准正态分布随机数和以自然对数为底数的指数函数的积,有向0收敛的趋势. AP聚类中各数据集的簇数对应偏好参数的一个区间,当发现者寻得一个合适的大区间时,需要加入者在发现者周围局部开发,找到最优的p. 通过本文作者观察,自然界中的麻雀群在觅食通常都会由一个体型稍大、基因更优秀的个体充当领头雀. 鉴于此,本文设计了基于动态跟踪领头雀的加入者更新策略,使得加入者小步长的在领头雀周围动态开发,增强算法的后期局部开发能力. 新的加入者更新公式为:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}g_{i,\mathrm{P}}+\sin\left(\mathrm{rand}(0,2\pi)\right)\times\left|\boldsymbol{X}_{\mathrm{P}}^t-\boldsymbol{X}_i^t\right|, i>\dfrac{n}{2}\\ \boldsymbol{X}_{\mathrm{P}}^t+\left|\boldsymbol{X}_i^t-\boldsymbol{X}_{\mathrm{P}}^t\right|\cdot\boldsymbol{E}^+\cdot\boldsymbol{L}, i\leqslant\dfrac{n}{2}\end{cases}\tag{16}$$

式中,$g_{i,\mathrm{P}}$为$\boldsymbol{X}_i$和$\boldsymbol{X}_{\mathrm{P}}$的距离,将$i>n/2$时向0收敛改为向发现者搜索到的最优麻雀即领头雀小步长跟踪收敛.

3.3　麻雀种群自适应划分策略

假设SSA中麻雀种群规模为n. 在迭代更新中发现者所占比重PD=0.6且保持不变,大比重的发现者有益于全局探索,但是每次迭代更新中大比重的发现者更新结束计算聚类适应度会增长寻优的时间,而且不利于后期的局部开发. 为解决这一问题,本文设计了发现者和加入者自适应划分的策略,更新开始时PD=0.6,发现者的数量为n×PD. 在迭代更新的过程中,使发现者的数量呈现递减的趋势,加入者相应采取递增的趋势,使改进SSA在迭代早期增加全局探索能力,提高迭代更新的速度,后期侧重于大比重的加入者的局部开发能力,提升寻优精度.

$$\mathrm{PD}_t=0.6-(0.6-0.2)\times\frac{2}{\pi}\times\arcsin\frac{t}{\mathrm{iter}_{\max}}\tag{17}$$

式中,PD_t为第t次迭代发现者所占比重,发现者为Round(n×PD),Round()表示取整;可知,PD随着迭代次数的增加非线性减少,动态改变麻雀种群中发现者和加入者所占的比重. JTSSA算法的具体流程如图6所示.

3.4　优化策略的测试

为了验证算法性能,JTSSA与SSA[16]、MFO(Moth-Flame Optimization)[17]、SMA(Slime Mould Algorithm)[18]和BES(Bald Eagle Search)[19]在7种基准测试函数[20]上比较. 设置各群智能算法种群规模为20,迭代次数为500. 测试中,每种算法运行30次,得到均值与标准差如表1所示,测试函数空间表示(三维展示)以及适应度收敛曲线如图7所示.

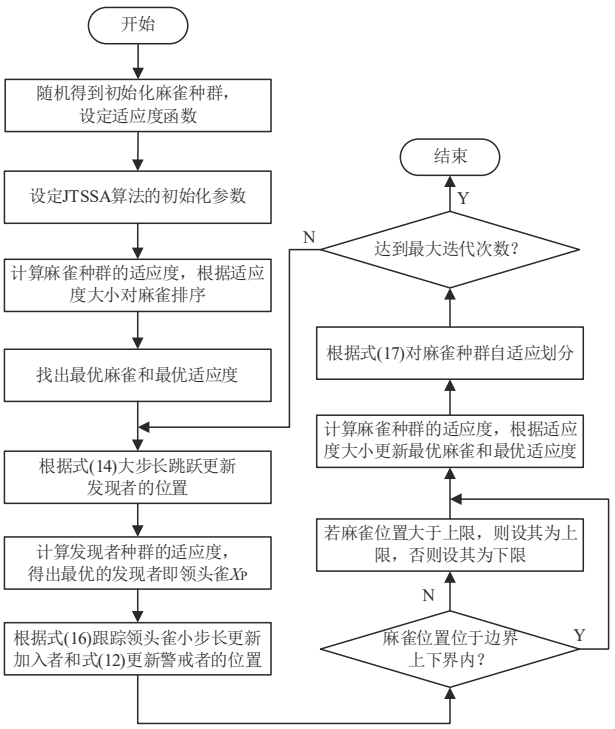


图6 JTSSA算法流程

表1和图7可以看出，与其它算法相比，JTSSA具有更快的收敛速度，效率更高，总体性能较优. 大步长机制以及动态跟踪策略的引入显著改善了SSA的收敛趋势和全局寻优能力.

4 JTSSA与AP的互补迭代

4.1 问题描述

$D=\{D_1,D_2,\cdots,D_n\}$是一个d维的数据集，目标是找到一个集合$C=\{C_1,C_2,\cdots,C_k\}$，$1<k<n$，集合C是D的k个聚类中心. C满足如下要求：

$$\begin{cases} C_i\neq\varnothing, i=1,2,\cdots,k \\ \bigcup_{i=1}^{k} C_i=D \\ C_i\cap C_j=\varnothing, i,j=1,2,\cdots,k\ ;\ i\neq j \end{cases} \tag{18}$$

AP算法可以描述为：

$$[\text{idx},\text{dpsim},\text{expref}]=\text{AP}(\boldsymbol{S},p,\lambda) \tag{19}$$

式中，idx为数据点的分类标签，dpsim为数据点与其聚类中心的相似度总和，expref为聚类中心的偏好值之和.

$\boldsymbol{S}$通过D可以得到，故式(19)可以描述为：

$$[\text{idx},\text{dpsim},\text{expref}]=\text{AP}'(D,p,\lambda) \tag{20}$$

定义C^*是C中最佳的集合，C^*可以描述如下：

$$C^*=\underset{C}{\arg}\ \text{Optimize}\ f(D,C) \tag{21}$$

式中，$f(\cdot)$是一个统计数学函数，可以基于距离划分的模式量化聚类中心的好坏. arg Optimize为对其后的函数择优选择，通过式(21)可以得到最优的聚类中心集合.

C通过idx映射得到：

$$C=g(\text{idx}) \tag{22}$$

式中，$g(\cdot)$是一个映射函数，可以从idx中计算得到相应的聚类中心. 通过式(19)~(21)可以得到：

$$C^*=\underset{g(\text{AP}'(D,p,\lambda))}{\arg}\ \text{Optimize}\ f(D,g(\text{AP}'(D,p,\lambda))) \tag{23}$$

假设p^*和λ^*是找到最优聚类数目的偏好参数，当找到最优聚类数目，满足：

$$p^*,\lambda^*=\underset{p,\lambda}{\arg}\ \text{Optimize}\ f(D,g(\text{AP}'(D,p,\lambda))) \tag{24}$$

4.2 聚类评价指标

轮廓系数(Silhouette Index，SI)具有相对较好的评价能力，被广泛使用，能够反映聚类数据集的类内紧密程度和类间区分程度. 本文使用SI评价JTSSA-AP算法聚类的效果. 假设数据集D被划分成了k类，D_i是其中一个样本数据，D_i的SI为：

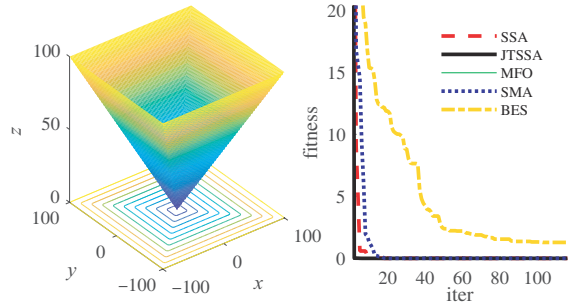
$$\text{SI}(D_i)=\frac{b(D_i)-a(D_i)}{\max\{b(D_i),a(D_i)\}} \tag{25}$$

式中，$a(D_i)$为样本D_i与其所属中心C_i中其它样本的平均相似度；$d(D_i,C_j)$为样本D_i与其它中心C_j的所有样本的平均相似度；$b(D_i)=\min\{d(D_i,C_j)\}$. 计算公式如下：

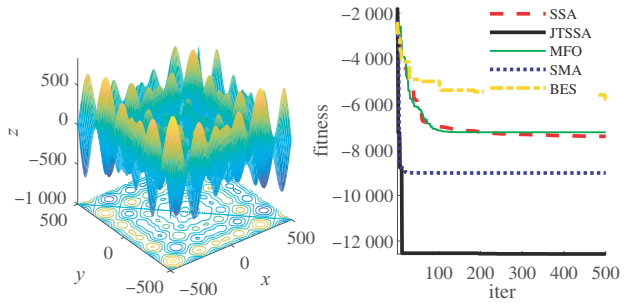
$$a(D_i)=\frac{1}{m_i-1}\sum_{q=1,q\neq i}^{n_i}\left\|D_i-D_q\right\| \tag{26}$$

表1 算法测试结果比较

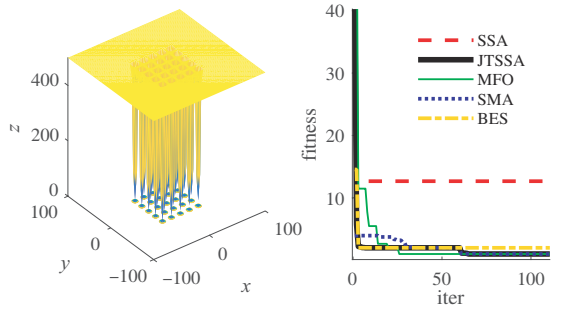
ID	本文算法		SSA		MFO		SMA		BES	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Schwefel 2.21	8.94×10^{-16}	9.32×10^{-12}	2.74×10^{-24}	6.82×10^{-7}	$5.55\times10^{+1}$	3.86×10^{-7}	3.65×10^{-56}	2.45×10^{-5}	3.02×10^{-1}	1.73×10^{-7}
Schwefel	$-1.26\times10^{+4}$	7.16×10^{-8}	$-7.41\times10^{+3}$	3.09×10^{-6}	$-7.23\times10^{+3}$	4.25×10^{-7}	$-9.01\times10^{+3}$	3.57×10^{-7}	$-5.80\times10^{+3}$	7.19×10^{-6}
Foxholes	9.98×10^{-1}	7.21×10^{-9}	$1.27\times10^{+1}$	7.86×10^{-8}	9.98×10^{-1}	9.03×10^{-9}	9.98×10^{-1}	5.69×10^{-8}	9.98×10^{-1}	8.79×10^{-8}
Branin	3.98×10^{-1}	2.36×10^{-8}	3.98×10^{-1}	1.20×10^{-6}	3.98×10^{-1}	7.31×10^{-9}	3.98×10^{-1}	6.32×10^{-8}	3.98×10^{-1}	3.52×10^{-9}
GoldStein-Price	3.00	0.00	3.00	2.21×10^{-8}	3.00	0.00	3.00	0.00	3.00	0.00
Hartman 3	−3.86	3.57×10^{-7}	−3.86	6.27×10^{-7}	−3.86	8.21×10^{-7}	−3.86	3.14×10^{-7}	−3.86	5.31×10^{-6}
Shekel 5	$-1.02\times10^{+1}$	3.21×10^{-8}	$-1.01\times10^{+1}$	2.14×10^{-7}	−2.68	2.53×10^{-7}	$-1.02\times10^{+1}$	2.78×10^{-8}	$-1.02\times10^{+1}$	5.39×10^{-8}



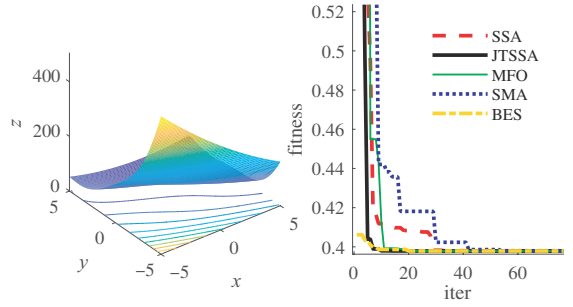
(a) Schwefel 2.21 函数



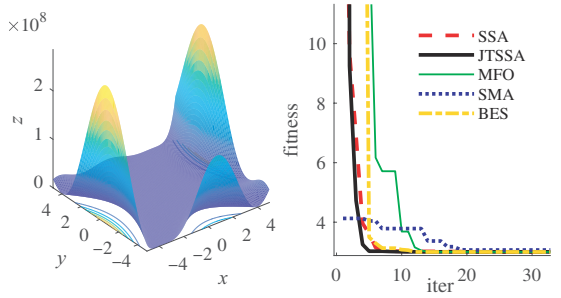
(b) Schwefel 函数



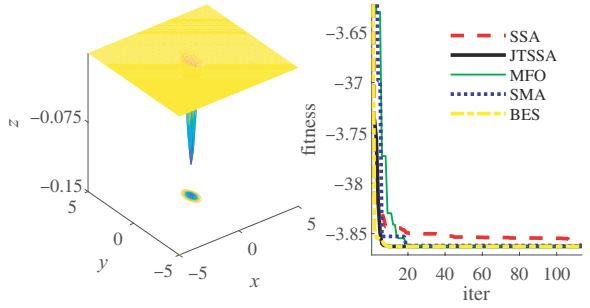
(c) Foxholes 函数



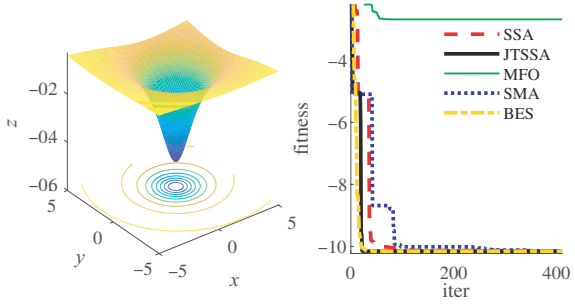
(d) Branin 函数



(e) GoldStein-Price 函数



(f) Hartman 3 函数



(g) Shekel 5 函数

图7　测试函数收敛曲线

式中，m_i为D_i所属类的数量，D_q为D_i所属类的其它样本.

$$b(D_i)=\min_{j=1,2,\cdots,k,j\neq i}\left(\frac{1}{m_j}\sum_{p=1}^{n_j}\left\|D_i-D_p\right\|\right) \tag{27}$$

通过式(25)可以得到一个样本D_i的SI值，进而由SI(D_i)可以计算得到一个类C_i的所有样本的平均SI值，记为Sav(C_i)，最后各个类相加平均计算得到数据集所有样本的平均SI值，记为Sav(C)，该值可以反映聚类结果的好坏，指标值越大，代表类别之间分离程度越大，类内紧密程度越高.

4.3 麻雀种群编码方案

AP聚类需先确定参数p和λ,本文拟采用JTSSA自动选择AP的两个参数,将p和λ作为JTSSA的两个寻优元素,即由p和λ组成麻雀个体$\boldsymbol{X}$.

设种群个数为N,设置有效的搜索空间为$[p_{\min},p_{\max}]$,当数据集聚类时,最大的聚类数目应该是$\sqrt{N}$[21].实验结果表明,当偏好值设置为$\bar{p}$的一半时,AP聚类数始终等于或大于$\sqrt{N}$,$p_{\max}=\bar{p}/2$,$p_{\min}$可以设置为一个更低的值,本文设置为$p_{\min}=\bar{p}\times10$,阻尼系数λ设置为$[0.5,1]$,根据p和λ的区间初始化得到初始麻雀种群.

4.4 适应度函数设计

SI可以用来估计聚类好坏,适应度函数设置为所有样本的平均SI值Sav(C):

$$f=\mathrm{Sav}(C) \tag{28}$$

根据式(20)、(22)、(28)得:

$$f=\mathrm{Sav}(C)=\mathrm{Sav}(g(\mathrm{idx}))=\mathrm{Sav}(g(\mathrm{AP}'(D,p,\lambda))) \tag{29}$$

4.5 JTSSA与AP的互补迭代

JTSSA与AP之间的互补迭代具体过程详见算法1,流程图如图8所示.

算法1 JTSSA-AP

输入:

N:麻雀种群的个数;

$s(i,k)$:相似度;

$(p_{\min},p_{\max})$:p的取值范围;

$(\lambda_{\min},\lambda_{\max})$:$\lambda$的取值范围;

Maxiter:最大迭代次数.

输出:

Gbestp, Gbestλ:最优麻雀即最优的p和λ;

GbestSav:最优麻雀的适应度;

CurveSav:每次迭代的适应度.

迭代过程:

步骤1:根据p和λ的取值范围利用RTent混沌映射初始化得到N只麻雀种群;

步骤2:运行AP算法,得到分类结果,根据式(29)计算适应度;

步骤3:对麻雀种群根据适应度进行排序,找出最优最差麻雀及其适应度,运行JTSSA算法,得到新的麻雀种群;

步骤4:判断是否达到最大迭代次数,或者最优适应度五次不变,是则退出,输出最优麻雀及其对应的适应度,否则返回步骤2.

4.6 复杂度分析

时间复杂度是评价算法性能的一个重要指标.SSA算法的时间复杂度为[22]:

$$T_{\mathrm{SSA}}=O(d+f) \tag{30}$$

式中,d表示空间维度,f为求解适应度所需时间.

JTSSA算法中,参数初始化所需执行时间为η_0,生成扰动因子的时间为η_1,按式(7)生成混沌麻雀的时间

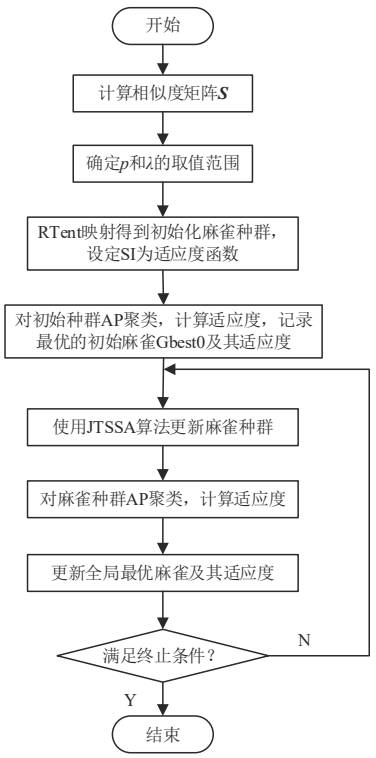


图8 JTSSA-AP算法流程

为η_2,则初始化阶段的时间复杂度为:

$$T_1=O\left(\eta_0+n\left(f+\left(\eta_1+\eta_2\right)\times d\right)\right) \tag{31}$$

麻雀种群中发现者数量为r_1n,随机数ω、Q生成时间为η_3和η_4,计算ρ的时间η_5,按照式(14)进行更新的时间为η_6,则发现者更新阶段时间复杂度为:

$$T_2=O\left(\eta_3+\eta_5+r_1n\left(\left(\eta_4+\eta_6\right)\times d\right)\right) \tag{32}$$

警戒者数量为r_2n,随机数β、ξ生成时间均为η_8,按照式(13)进行更新的时间为η_7,则警戒者更新阶段时间复杂度为:

$$T_3=O\left(r_2n\left(\left(\eta_7+\eta_8+\eta_8\right)\times d\right)\right) \tag{33}$$

加入者数量为$(1-r_1-r_2)n$,产生参数$g_{i,\mathrm{P}}$的时间为η_9,按式(16)进行位置更新的时间为η_{10},则加入者更新阶段时间复杂度为:

$$T_4=O\left(\left(1-r_1-r_2\right)n\left(\left(\eta_9+\eta_{10}\right)\times d\right)\right) \tag{34}$$

假设计算比例参数PD的时间分别为η_{11},综上,JTSSA的时间复杂度为:

$$T=O\left(T_1+\mathrm{iter}_{\max}\left(T_2+T_3+T_4+\eta_{11}\right)\right)=O\left(d+f\right)=T_{\mathrm{SSA}} \tag{35}$$

综上,JTSSA与基本SSA的时间复杂度相同.

AP算法[8]的时间复杂度为$O(N^3)$,按式(29)计算轮廓系数的时间为η_{12},则$f=O(N^3)+\eta_{12}$,JTSSA和AP互补迭代的时间复杂度为:

$$T'=O\left(N^3+d\right) \tag{36}$$

从整体上看,算法的时间复杂度主要和数据集规模有关.

5　实验结果分析与应用

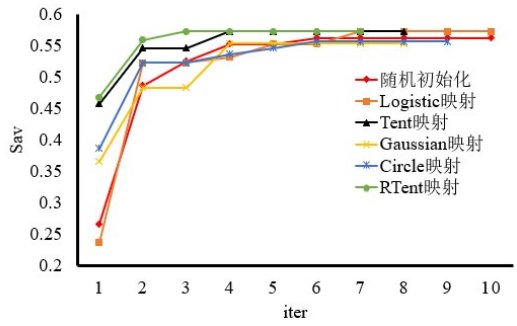
硬件平台为Intel Core i7-11800 H 2.3 GHz CPU、RTX3060 GPU、16 GB内存的计算机,计算软件平台为Matlab R2017b.

5.1　初始化对比

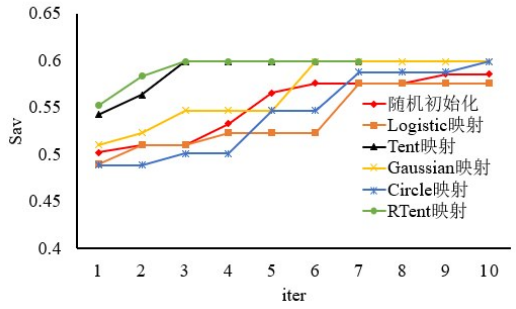
根据p和λ的取值区间,采用随机初始化、四种基本混沌映射和RTent映射生成初始种群,进行JTSSA-AP聚类,选取UCI(University of California Irvine)数据集中的Iris、Wine、Glass、Seeds数据集聚类,数据集的主要特征如表2所示,聚类对比结果图如图9所示.

表2　标准数据集特征

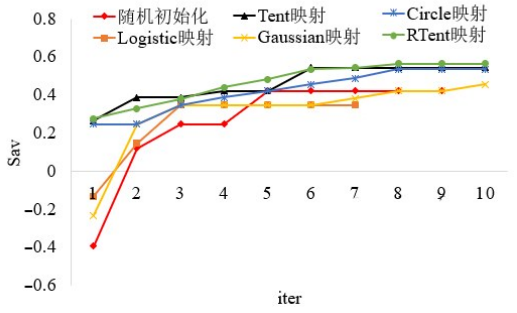
数据集名称	样本数目	属性维数	类别个数
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Seeds	210	7	3



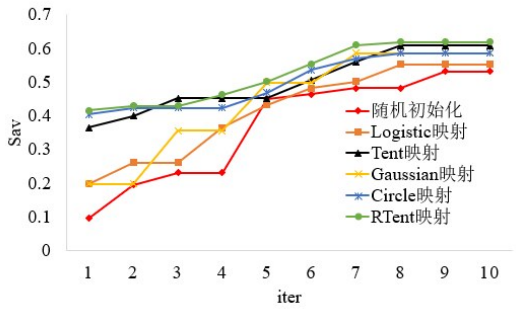
(a) Iris数据集对比



(b) Wine数据集对比



(c) Glass数据集对比



(d) Seeds数据集对比

图9　各混沌映射在各数据集初始化聚类对比

混沌映射相比随机初始化可以有效提高聚类的初始适应度和迭代速度;Tent映射得到的初始麻雀种群具有更优的质量,相比其它初始化算法具有更好的效果.RTent映射引入扰动因子并增加3个参数,应用场景更加广泛灵活,获得比原Tent映射更优的初始种群,故本文初始化麻雀种群采用RTent映射获得.

5.2　p值和聚类中心数目的关系

AP算法选取$\boldsymbol{S}$中的数据中值作为p的数值,讨论p和聚类中心数目的关系,具体步骤如下:

(1)设定p初值为$\boldsymbol{S}$的中值.

(2)AP聚类.

(3)按步长为p的初值对p值减小.

重复步骤(2)、(3),直到聚类中心数目等于2或者达到预设的最大迭代次数.采取上述算法处理Iris数据集,得到的p以及相应的聚类中心的数目k如图10所示.

随着p值减小,k也相应减小,但是p和k并不是一一对应的关系,某一个聚类中心会对应一个区间的p值.所以从大到小定步长减小p寻找最优的k,会增加寻优时间.为了提高寻优效率,本文通过引入SI作为评

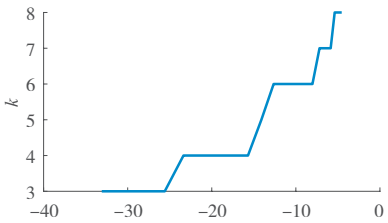


图10 p值及其对应的聚类中心数目k

价函数使用JTSSA在定义好的p区间上跳跃寻找p值.

5.3 算法性能评价

本文采用所有数据的平均SI值即Sav作为评价函数,因为UCI数据集中已知数据标签,这里增加了准确率ACC、调整兰德系数ARI和聚类中心数目k三个评价指标,对不同聚类算法进行效果评估. 其中,ACC表示聚类准确率,ARI衡量两组数据分布的吻合程度. ACC和ARI的取值范围都为[0,1],值越大表示结果越好.

为评价聚类算法效果,将AP[8]、EAP(Evolutionary Affinity Propagation)[23]、DE-AP(AP Clustering Algorithm Based on Differential Evolution)[24]、SSA-AP(AP clustering algorithm based on SSA)、JTSSA-AP对UCI的10种公共数据集进行聚类分析,10种数据集的样本数和类别个数有多有少,属性维数有高有低,基本涵盖了生活中各种真实数据集类型,主要特征如表3所示. 各聚类算法对各数据集聚类运行50次取均值的结果评价如表4所示,将数据集从左到右编码为1~10,图11为对表4各个指标的可视化.

表3 标准数据集特征

数据集名称	Diabetes	Heartstatlog	Bupa	Haberman	Vehicle	Balancescale	Thyroid	Vowel	Aggregation	Ecoli
样本数目	768	270	345	306	846	625	215	871	788	336
属性维数	8	13	6	3	18	4	5	3	2	8
类别个数	2	2	2	2	3	3	3	6	7	8

表4 各聚类算法对各数据集聚类的结果评价

算法	评价指标	数据集									
		Diabetes (2簇)	Heartstatlog (2簇)	Bupa (2簇)	Haberman (2簇)	Vehicle (3簇)	Balancescale (3簇)	Thyroid (3簇)	Vowel (6簇)	Aggregation (7簇)	Ecoli (8簇)
AP	Sav	4.22×10^{-2}	-1.80×10^{-3}	-3.29×10^{-1}	-1.70×10^{-1}	3.74×10^{-2}	-7.03×10^{-2}	-1.14×10^{-1}	1.30×10^{-1}	3.20×10^{-1}	-5.02×10^{-1}
	Acc	4.62×10^{-1}	4.19×10^{-1}	3.10×10^{-1}	4.75×10^{-1}	3.83×10^{-1}	2.30×10^{-1}	4.44×10^{-1}	5.42×10^{-1}	2.98×10^{-1}	3.90×10^{-1}
	ARI	2.09×10^{-1}	2.12×10^{-1}	2.09×10^{-1}	3.07×10^{-1}	2.57×10^{-1}	1.29×10^{-1}	1.85×10^{-1}	1.17×10^{-1}	1.91×10^{-1}	1.54×10^{-1}
	k	49	30	39	31	34	49	28	52	35	31
EAP	Sav	2.06×10^{-1}	4.45×10^{-2}	3.11×10^{-1}	3.55×10^{-1}	1.21×10^{-1}	1.58×10^{-1}	3.31×10^{-1}	4.75×10^{-1}	4.75×10^{-1}	3.59×10^{-1}
	Acc	6.77×10^{-1}	6.46×10^{-1}	5.80×10^{-1}	6.95×10^{-1}	6.29×10^{-1}	7.34×10^{-1}	8.56×10^{-1}	7.81×10^{-1}	9.11×10^{-1}	4.25×10^{-1}
	ARI	2.78×10^{-1}	3.32×10^{-1}	4.06×10^{-1}	4.22×10^{-1}	3.87×10^{-1}	2.90×10^{-1}	5.18×10^{-1}	4.36×10^{-1}	5.25×10^{-1}	2.90×10^{-1}
	k	7	7	4	5	6	15	3	9	7	2
DE-AP	Sav	1.61×10^{-1}	1.84×10^{-1}	1.53×10^{-1}	2.81×10^{-1}	1.70×10^{-1}	1.65×10^{-1}	2.31×10^{-1}	3.04×10^{-1}	4.06×10^{-1}	1.21×10^{-1}
	Acc	6.24×10^{-1}	6.91×10^{-1}	6.03×10^{-1}	6.55×10^{-1}	5.19×10^{-1}	7.50×10^{-1}	7.79×10^{-1}	7.52×10^{-1}	5.95×10^{-1}	8.02×10^{-1}
	ARI	2.59×10^{-1}	2.37×10^{-1}	4.07×10^{-1}	2.55×10^{-1}	3.23×10^{-1}	2.63×10^{-1}	3.93×10^{-1}	2.92×10^{-1}	3.44×10^{-1}	5.54×10^{-1}
	k	11	4	4	9	9	17	5	7	18	6
SSA-AP	Sav	9.05×10^{-2}	2.84×10^{-1}	2.88×10^{-2}	5.47×10^{-2}	2.31×10^{-1}	6.49×10^{-2}	1.36×10^{-1}	1.80×10^{-1}	4.18×10^{-1}	2.21×10^{-1}
	Acc	5.02×10^{-1}	6.78×10^{-1}	5.10×10^{-1}	5.53×10^{-1}	5.25×10^{-1}	6.24×10^{-1}	5.26×10^{-1}	6.85×10^{-1}	6.87×10^{-1}	6.37×10^{-1}
	ARI	2.29×10^{-1}	3.11×10^{-1}	3.09×10^{-1}	2.09×10^{-1}	3.12×10^{-1}	2.51×10^{-1}	2.85×10^{-1}	2.07×10^{-1}	4.32×10^{-1}	3.62×10^{-1}
	k	23	3	19	19	10	19	22	29	14	15
JTSSA-AP	Sav	4.45×10^{-1}	3.56×10^{-1}	3.64×10^{-1}	3.79×10^{-1}	5.42×10^{-1}	2.65×10^{-1}	3.55×10^{-1}	5.56×10^{-1}	4.88×10^{-1}	3.93×10^{-1}
	Acc	8.51×10^{-1}	8.19×10^{-1}	7.80×10^{-1}	7.35×10^{-1}	8.21×10^{-1}	8.44×10^{-1}	8.58×10^{-1}	8.35×10^{-1}	8.87×10^{-1}	8.22×10^{-1}
	ARI	3.34×10^{-1}	3.53×10^{-1}	5.09×10^{-1}	5.03×10^{-1}	4.31×10^{-1}	3.68×10^{-1}	5.30×10^{-1}	4.90×10^{-1}	5.32×10^{-1}	5.40×10^{-1}
	k	2	2	2	2	3	7	3	6	7	9

从表4和图11曲线可以看出,本文算法的Sav指标>EAP>DE-AP>SSA-AP>AP,ACC和ARI指标亦如此,同时,本文算法的聚类簇数更接近各数据集的原簇数. AP算法因为根据矩阵$\boldsymbol{S}$非对角线元素计算的偏好参数p一般选择过大,聚类的簇数最多,各指标均最低;SSA-AP和DE-AP算法因为混合了群智能算法交叉迭代,聚类的簇数相比AP会大幅度降低,但SSA和DE的算法寻优能力较差,易陷入局部最优,因此Sav、ACC和ARI

指标均较差,聚类簇数和各数据集的原簇数差异也较大;EAP算法由于引入了跨时间连接可变节点的因子,增加了聚类的准确性和鲁棒性,Sav指标和聚类簇数优于SSA-AP和DE-AP,但该算法易早熟,Sav指标比本文算法较差,聚类簇数的准确度差异也更大. 以上各指标对比曲线证明本文算法相比其它算法具有更强的寻优能力,通过评价指标和聚类簇数的综合对比可以看出,本文算法能够针对复杂结构化数据集的鲁棒性更好,同时精度较高.

算法对各数据集聚类运行50次取均值得到的p和λ如表5所示. 总体上本文算法的p值<DE-AP<SSA-AP<AP,p决定聚类簇数,由表4可知本文算法可以找到最优簇数,寻得的p值最适合;AP和DE-AP的λ值<本文算法<SSA-AP< EAP. λ决定振荡程度和迭代速度,过小不能减小振荡,过大会减慢迭代速度,本文算法的λ值适中,总是能找到接近最优聚类数的p和λ.

表5　不同算法聚类的p和λ

数据集		Diabetes	Heartstatlog	Bupa	Haberman	Vehicle	Balancescale	Thyroid	Vowel	Aggregation	Ecoli
偏好参数(p)	AP	$-1.03\times10^{+2}$	$-6.53\times10^{+1}$	$-4.06\times10^{+1}$	$-1.48\times10^{+1}$	$-1.72\times10^{+2}$	-3.87	$-1.50\times10^{+1}$	$-6.34\times10^{+2}$	$-1.65\times10^{+1}$	-5.57×10^{-1}
	DE-AP	$-8.73\times10^{+2}$	$-6.90\times10^{+2}$	$-5.55\times10^{+2}$	$-8.31\times10^{+1}$	$-1.62\times10^{+3}$	$-1.66\times10^{+1}$	$-1.64\times10^{+2}$	$-9.38\times10^{+3}$	$-5.17\times10^{+1}$	-3.14
	SSA-AP	$-5.32\times10^{+2}$	$-9.32\times10^{+2}$	$-2.41\times10^{+2}$	$-5.36\times10^{+1}$	$-1.04\times10^{+3}$	$-1.29\times10^{+1}$	$-3.64\times10^{+1}$	$-3.88\times10^{+3}$	$-7.93\times10^{+1}$	-1.16
	JTSSA-AP	$-7.52\times10^{+3}$	$-2.61\times10^{+3}$	$-1.42\times10^{+3}$	$-5.51\times10^{+2}$	$-5.04\times10^{+3}$	$-6.49\times10^{+1}$	$-6.76\times10^{+2}$	$-1.63\times10^{+4}$	$-3.19\times10^{+2}$	$-1.13\times10^{+1}$
阻尼系数(λ)	AP/DE-AP	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	EAP	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
	SSA-AP	8.99×10^{-1}	8.29×10^{-1}	6.99×10^{-1}	6.78×10^{-1}	5.44×10^{-1}	6.22×10^{-1}	5.73×10^{-1}	5.64×10^{-1}	7.43×10^{-1}	6.81×10^{-1}
	JTSSA-AP	5.81×10^{-1}	5.00×10^{-1}	7.70×10^{-1}	5.44×10^{-1}	6.31×10^{-1}	5.52×10^{-1}	5.21×10^{-1}	7.47×10^{-1}	5.43×10^{-1}	6.63×10^{-1}

5.4　国家信息数据聚类分析

实验选取具有代表性的国家信息聚类数据集[25]进行聚类效果测试,数据为167个国家的9个字段,主要特征包括国家的child_mort(每1 000名出生婴儿中5岁以下儿童死亡数)、exports(人均商品和服务出口额,按人均GDP的百分比计算)、health(人均卫生支出总额,按人均GDP的百分比计算)、imports(人均商品和服务进口额,按人均GDP的百分比计算)、Income(人均纯收入)、Inflation(GDP年增长率)、life_expec(如果当前的死亡率模式保持不变,新生儿的平均寿命)、total_fer(如果当前年龄生育率保持不变,每个妇女将要生的孩子数量)、gdpp(人均GDP). 实验目的是通过聚类,找出每一个簇中国家的共同特征和簇间国家的不同特征,以及分析9个因素之间的关系. 数据特征前五行如表6所示,各算法对国家信息数据集聚类的结果如表7所示.

表7表示了各算法对国家信息数据集运行50次取均值的结果. 本文算法得到的聚类簇数是3类,其Sav指标最高,分别比EAP、DE-AP、SSA-AP和AP算法高0.069、0.072、0.104和0.154. 对每个簇的各个指标求均值结果如表8所示,下面通过表中数据分析本文算法聚类簇数为3类的原因.

对表8的各簇的可视化如图12所示.

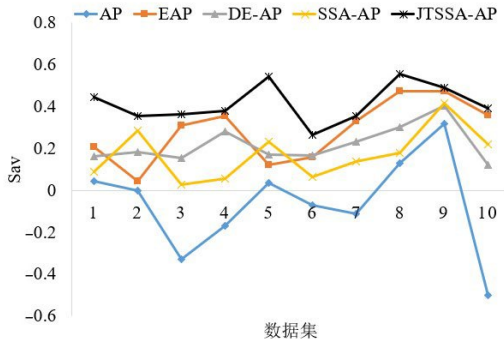
表8中展示了各簇的均值以及总数据集针对各要素的均值对比. 世界上每1 000人中平均死亡人数为38.2,第1簇国家的儿童死亡率最低,相应的进出口率、卫生支出、净收入和GDP都最高,因此,第1簇为发达国家. 同时,由表8和图12还可以得出发达国家的人口老龄化严重,出生率偏低. 第3簇国家儿童人口的死亡率最高,预期寿命、收入和GDP低得多,因此,都为落后国家. 与发达国家相比,这些国家进口较多,出口较少,

表6　country_data数据集前五行

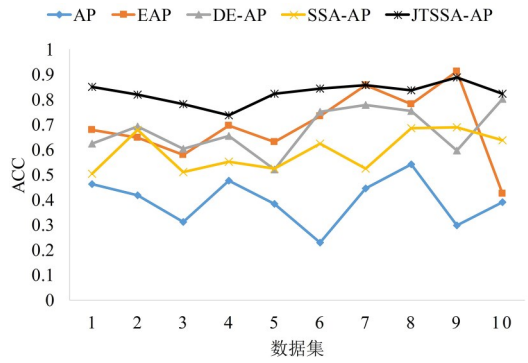
country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	90.2	10	7.58	44.9	1 610	9.44	56.2	5.82	553
Albania	16.6	28	6.55	48.6	9 930	4.49	76.3	1.65	4 090
Algeria	27.3	38.4	4.17	31.4	12 900	16.1	76.5	2.89	4 460
Angola	119	62.3	2.85	42.9	5 900	22.4	60.1	6.16	3 530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19 100	1.44	76.8	2.13	12 200

表7　标准数据集特征

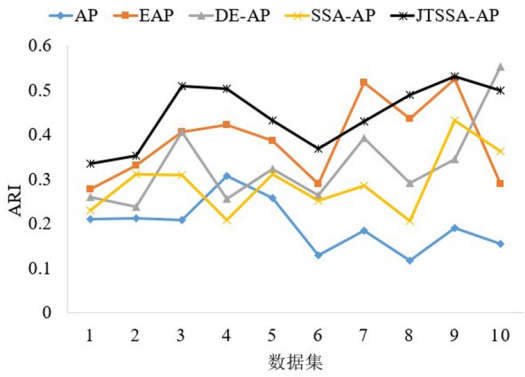
算法指标	AP	EAP	DE-AP	SSA-AP	JTSSA-AP
Sav	1.77×10^{-1}	2.62×10^{-1}	2.59×10^{-1}	2.27×10^{-1}	3.31×10^{-1}
k	15	5	6	9	3
偏向参(p)	$-2.58\times10^{+4}$	—	$-2.05\times10^{+5}$	$-8.98\times10^{+4}$	$-5.53\times10^{+5}$



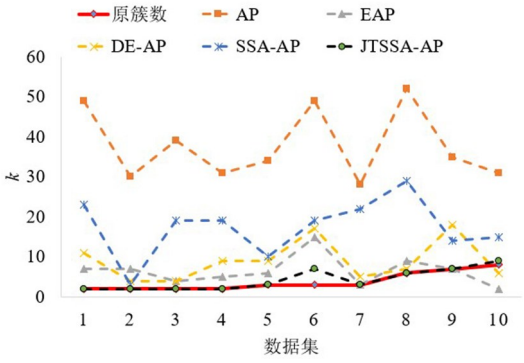
(*a*) Sav值



(*b*) ACC值



(*c*) ARI值



(*d*) *k*值

图11 各算法在各数据集上的聚类指标

表8 国家信息数据聚类的各簇的各因素均值

特征	因素均值			
	k=1	*k*=2	*k*=3	数据集
child_mort	5.59	$1.85\times10^{+1}$	$6.61\times10^{+1}$	$3.82\times10^{+1}$
exports	$5.60\times10^{+1}$	$4.75\times10^{+1}$	$3.04\times10^{+1}$	$4.11\times10^{+1}$
health	8.48	6.79	6.15	6.81
imports	$4.65\times10^{+1}$	$4.79\times10^{+1}$	$4.63\times10^{+1}$	$4.69\times10^{+1}$
Income	$4.90\times10^{+4}$	$1.74\times10^{+4}$	$3.90\times10^{+3}$	$1.71\times10^{+4}$
Inflation	3.85	6.57	$1.03\times10^{+1}$	$7.78\times10^{+1}$
life_expec	$8.01\times10^{+1}$	$7.40\times10^{+1}$	$6.41\times10^{+1}$	$7.05\times10^{+1}$
total_fer	1.89	2.05	4.04	2.94
gdpp	$4.51\times10^{+4}$	$1.03\times10^{+4}$	$1.75\times10^{+3}$	$1.30\times10^{+4}$

这表示第3簇的国家的工业和制造业最不发达,导致了就业率下降、贫困加剧和生活成本的上升. 通过数据分析可以得出,儿童死亡率的高低代表了这个国家的发展程度,第3簇的国家应加强自身的工业和制造业发展,提升就业率,改善生活条件,在降低儿童出生率的同时减少死亡率.

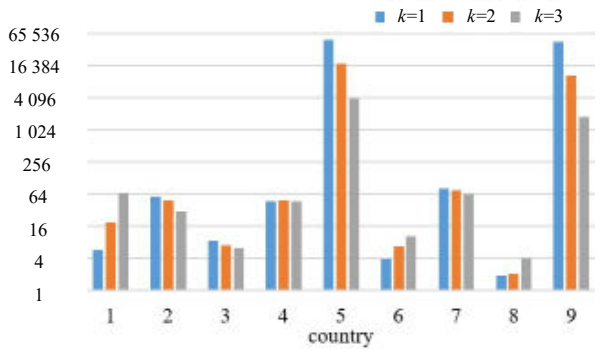


图12 国家信息数据聚类的各簇的各因素对比

6 结论与展望

本文提出了一种基于跳跃跟踪的SSA优化方法JTSSA,引入轮廓系数作为评价指标并与AP算法混合迭代,自动寻找较优的p和λ. 经JTSSA-AP算法聚类10种不同UCI数据集的仿真测试结果表明,提出的聚类算法与经典AP、DE-AP、SSA-AP和EAP方法相比具有更优的搜索效率以及聚类精度,在实际中具有更广阔的应用空间. 通过国家信息数据集聚类验证,本文算法的

Sav指标>EAP>DE-AP> SSA-AP>AP,且聚类簇数是3类,经分析可以得到有效且合理的聚类结果. 将数据集中的国家分为发达,发展中和落后国家,得出了国家发展程度和各因素之间的关系,并针对各簇的国家发展情况提出相应建议,验证了本文算法的应用价值.

本文后续拟针对如下工作进一步开展研究:(1)AP算法的偏向参数和阻尼系数自动更新方面,如何进一步加速,减小计算量,需要更进一步研究;(2)AP算法复杂度比较高,如何聚类结构复杂和大规模的数据,需要进一步研究数据压缩方法和AP聚类算法的因子图稀疏化方法,使聚类算法可自动高效地快速执行;(3)算法如何在工程场景中实际应用,将继续深入研究.

参考文献

[1] PENG X, FENG J S, ZHOU J T, et al. Deep subspace clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5509-5521.

[2] CHEN Y W, HU X L, FAN W T, et al. Fast density peak clustering for large scale data based on kNN[J]. Knowledge-Based Systems, 2020, 187: 104824.

[3] NASIR A, MASHOR A S, MOHAMED M, et al. Enhanced kmeans clustering algorithm for malaria slide image segmentation[J]. Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 2018, 42(1): 1-15.

[4] YOON J, JOUNG S. A big data based cosmetic recommendation algorithm[J]. Journal of System and Management Sciences, 2020, 10(2): 40-52.

[5] ZHANG G, ZHANG C C, ZHANG H Y. Improved K-means algorithm based on density Canopy[J]. Knowledge-Based Systems, 2018, 145: 289-297.

[6] XI L, ZHANG F B. An adaptive artificial-fish-swarm-inspired fuzzy C-means algorithm[J]. Neural Computing and Applications, 2020, 32(22): 16891-16899.

[7] CHEN Y W, ZHOU L D, BOUGUILA N, et al. BLOCK-DBSCAN: Fast clustering for large scale data[J]. Pattern Recognition, 2021, 109: 107624.

[8] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[9] DAGHER I, MIKHAEL S, AL-KHALIL O. Gabor face clustering using affinity propagation and structural similarity index[J]. Multimedia Tools and Applications, 2021, 80 (3): 4719-4727.

[10] HAN Y M, FAN C Y, GENG Z Q. Energy efficient building envelope using novel RBF neural network integrated affinity propagation[J]. Energy, 2020, 209: 118414.

[11] 黄鹤, 李文龙, 杨澜, 等. DHSSA优化的K均值互补迭代车型信息数据聚类[J]. 汽车工程, 2022, 44(05): 691-700, 729.
HUANG H, LI W L, YANG L, et al. K-means complementary iterative vehicle information data clustering based on DHSSA optimization[J]. Automotive Engineering, 2022, 44(05): 691-700, 729. (in Chinese)

[12] 黄鹤, 李潇磊, 杨澜, 等. 引入改进蝠鲼觅食优化算法的水下无人航行器三维路径规划[J]. 西安交通大学学报, 2022, 56(7): 9-18.
HUANG H, LI X L, YANG L, et al. Three dimensional path planning of unmanned underwater vehicle based on improved manta ray foraging optimization algorithm[J]. Journal of Xi'an Jiaotong University, 2022, 56(7): 9-18. (in Chinese)

[13] LIU Y C, LIU J C, JIN Y C. An affinity propagation clustering based particle swarm optimizer for dynamic optimization[J]. Knowledge-Based Systems, 2020, 195: 105711.

[14] ZHOU R H, LIU Q M, WANG J, et al. Modified semi-supervised affinity propagation clustering with fuzzy density fruit fly optimization[J]. Neural Computing and Applications, 2021, 33(10): 4695-4712.

[15] WANG L M, JI Q, HAN X M. Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity[J]. Tehnicki Vjesnik-Technical Gazette, 2016, 23(2): 425-436.

[16] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[17] MIRJALILI S. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm[J]. Knowledge-Based Systems, 2015, 89: 228-249.

[18] LI S M, Chen H M, Wang M J, et al. Slime mould algorithm: A new method for stochastic optimization[J]. Future Generation Computer Systems, 2020, 111: 300-323.

[19] ALSATTAR H A, ZAIDAN A A, ZAIDAN B B. Novel meta-heuristic bald eagle search optimisation algorithm [J]. Artificial Intelligence Review, 2020, 53(3): 2237-2264.

[20] ZHAO W G, ZHANG Z X, Wang L Y. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications[J]. Engineering Applications of Artificial Intelligence, 2020, 87: 103300.

[21] WANG X H, QIN Z, ZHANG X P. Automatically affinity propagation clustering using particle swarm[J]. Journal of Computers, 2010, 5(11): 1731-1738.

[22] 毛清华, 张强. 融合柯西变异和反向学习的改进麻雀算

法[J]. 计算机科学与探索, 2021, 15(6): 1155-1164.
MAO Q H, ZHANG Q. Improved sparrow algorithm combining cauchy mutation and opposition-based learning [J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(6): 1155-1164. (in Chinese)

[23] ARZENO N M, VIKALO H. Evolutionary affinity propagation[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2017: 2681-2685.

[24] 赵小强, 谢亚萍. 基于差分改进的仿射传播聚类算法[J]. 控制工程, 2018, 25(12): 2115-2119.
ZHAO X Q, XIE Y P. An affinity propagation clustering algorithm based on differential evolution[J]. Control Engineering of China, 2018, 25(12): 2115-2119. (in Chinese)

[25] kokkula Rohan. Country-Data[EB/OL]. (2020-07-17) [2022-02-01]. https://www. kaggle. com/rohan0301/unsupervised-learning-on-country-data.

作者简介

黄 鹤 男,1979年2月出生,现为长安大学教授,博士生导师,主要研究方向是无人机测控、信息融合等.
E-mail: huanghe@chd.edu.cn

李文龙 男,1999年4月出生,现为长安大学硕士研究生,主要研究方向是路径规划、目标跟踪等.
E-mail: 691508561@qq.com

杨 澜 女,1985年7月出生,现为长安大学信息工程学院高级工程师,主要研究方向是车联网.
E-mail: lanyang@chd.edu.cn

王会峰 男,1976年12月出生,现为长安大学教授,博士生导师,主要研究方向是机器视觉与图像处理技术.
E-mail: hfwang@chd.edu.cn

高 涛 男,现为长安大学教授信息工程学院教授. 主要研究方向是图像处理、模式识别、人工智能. 中国电子学会会员编号:E190026033M.
E-mail: gtnwpu@126.com

陈 婷 女,现为长安大学教授信息工程学院副教授. 主要研究方向是无线通信、智能交通系统、模式识别.
E-mail: tchenchd@126.com