一种基于禁忌搜索的全局最优化模糊聚类算法

朱 毅[1,3],杨 航[2],吕泽华[1],陈传波[1],邹小威[1]
(1. 华中科技大学软件学院,湖北武汉 430079; 2. 深圳市腾讯计算机系统有限公司,广东深圳 518000;
3. 武汉华中时讯科技有限责任公司,湖北武汉 430079)

摘 要: 模糊C均值(FCM)算法是一种基于贪心思想的迭代算法,算法沿迭代序列收敛到一个极小值,但存在搜索能力弱、易陷入局部最优的缺点.本文提出了一种基于禁忌搜索的模糊聚类算法,该算法在一个解的邻域内使用禁忌搜索,并采用了基于FCM局部收敛性质的长期表禁忌策略,保证在不断移动搜索起点的同时避免重复搜索;其次使用混沌优化思想与动态步长策略来提升算法的全局搜索能力,以达到获取全局最优解的目的.实验结果表明,改进算法极大地提高了聚类准确率,并具有良好的稳定性,与群智算法和遗传算法的优化相比也具有一定的优势.

关键词: 模糊C均值(FCM)算法; 禁忌搜索; 全局最优
中图分类号: TP319 **文献标识码:** A **文章编号:** 0372-2112 (2019)02-0289-07
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.02.005

A Global Optimization Fuzzy Clustering Algorithm Based on Tabu Search

ZHU Yi[1,3], YANG Hang[2], LYU Ze-hua[1], CHEN Chuan-bo[1], ZOU Xiao-wei[1]
(1. *Huazhong University of Science & Technology, Wuhan, Hubei* 430079, *China*;
2. *Shenzhen Tencent Computer Systems Company Limited, Shenzhen, Guangdong* 518000, *China*;
3. *Sencent Technology (Wuhan) Co., Ltd. Wuhan, Hubei* 430079, *China*)

Abstract: The fuzzy c-Means algorithm is a kind of iterative algorithms based on greedy algorithms. It converges to a local minimum value along the iteration sequence, yet it has the insufficient searching ability and can easily fall into local optimum solution. This paper, based on tabu search, introduces a fuzzy clustering algorithm. It uses tabu search in a solution's neighborhood and adopts the tabu strategy of long-term tabu lists based on the local convergence of FCM, which guarantees to move the search starting point constantly and avoids repeated searching. In addition, chaos optimization and dynamic step strategies are utilized to strengthen its global search ability in order to achieve global optimal solution. Experimental results show that this algorithm improves the accuracy of clustering considerably and has great stability. Compared with group-wise algorithm and genetic algorithm, this algorithm also has some advantages.

Key words: fuzzy c-means; tabu search; global minimum

1 引言

模糊C均值(Fuzzy C-Means, FCM)算法是聚类分析中的常用算法,该算法从K-means算法(Hard C-Means, HCM)出发,根据模糊理论[1]的思想引入隶属度的概念,对样本与聚簇之间的隶属关系进行模糊化,允许样本按程度分属于多个聚簇,从而提升了算法的准确度与性能.FCM算法对样本点与聚簇中心之间的距离进行度量,以求解到最小的类内距离作为求解目标,目标函数表示为隶属度加权的类内距离总和,建模后聚类问题抽象为带有约束条件的函数最优化求解问题,Bezdek和Hathaway从数学上证明了FCM算法的全局收敛性[2].

FCM算法比HCM算法有更强的聚类性能,但仍存

收稿日期:2017-05-26;修回日期:2018-10-26;责任编辑:马兰英
基金项目:中央高校基本科研业务费资助(No. HUST:2017KFYXJJ226)

在如下问题:(1)FCM 中聚类数 c 需要指定,而对于未知问题,很难对 c 给出指导性的值;(2)由于算法基于固定的距离度量方式,FCM 只善于发现某一具体形状的聚簇,当数据集中簇的形状不固定或形状与选择的距离度量方式不匹配时,即使目标函数取得全局最小值,仍可能出现聚类簇划分错误的情况;(3)由于采用加权均值的方式,算法还易于受到孤立噪声点的影响;(4)FCM 算法采用加权的类内平方距离作为目标函数,采用迭代的方法求精,而类内平方距离为一非凸函数,函数存在多个局部最小值,迭代最终结果对起始点有很强的依赖性,不佳的起始点将导致算法最终收敛到局部最小值.

FCM 依赖初始点的选取、易收敛到局部最优的缺点由目标函数的非凸性所导致,为解决这个问题,通常从以下两个方面优化:(1)优化初始点的选取,使得迭代序列最终停止在目标函数的全局最优解上;(2)从数学的角度上求解非凸函数的全局最小值. 在求解非凸函数的全局最优值方面,国内外许多学者在这方面做出了大量研究,例如:李凯等使用神经网络和复突触神经网络求解可以使目标函数达到最小的问题解,给出了广义熵模糊聚类算法的解法[3];唐成华等利用遗传算法强大的全局搜索能力克服 FCM 迭代时易于陷入局部最优的缺点,并将其应用于异常入侵检测模型的构造中,提高了入侵检测的成功率[4];Katayoon Ahmadi 等人利用 GA 强化 FCM 的隶属度矩阵,优化 FCM 的求解过程,并将算法应用于肝脏血管 CT 图像分割,取得了很好的效果[5];周双利等采用模拟退火算法补充遗传算法在局部搜索上的不足,进一步提升了算法性能[6];余晓东等人使用粒子群算法首先搜索到一个优秀解,再使用直觉模糊核聚类方法进行了强化得到全局最优解并提高收敛速度[7];Zhongxing Zhang 等使用 FCM 强化粒子群搜索结果,在有限的迭代次数下极大的提高了搜索准确率[8];耿宗科等则提出了一种粒子群算法参数自适应调节的算法,提升了收敛速度[9];Franco, D. G. D. B. 等人进一步发展了 FCM 算法,结合 PSO,GA 提出了一种混合模糊聚类算法,极大提升了算法收敛速度[10];皇甫中民等人受鱼群运动模式启发,提出基于全局公告信息引导及模糊 C 均值修正的人工鱼群聚类算法,与传统粒子群算法相比有更强的全局搜索能力[11];TRAN 等人则在人工蜂群算法的基础上融合粒子群算法、遗传算法突变策略,提出混合改进的人工蜂群算法,将其应用于 k 均值数据聚类中,有很好的性能与鲁棒性[12]. 上述算法基本上都通过保留部分劣解而摆脱局部最优,尤其是各种群智算法,兼顾了局部搜索与全局搜索,取得了较好的效果. 但此类算法也增加了算法的复杂性,算法收敛速度大大降低,并且还引入了多个待确定参数与阈值,如何科学甚至自动地设置参数值是这类算法要解决的一大难题.

基于以上分析,本文采用禁忌算法对 FCM 进行优化. 与其它启发式算法相比,禁忌算法更好地模拟了人的思维习惯,算法在较少地依赖额外参数的前提下,也可以得到很好的搜索效果. 良好的记忆过程还可以在搜索中及时剪枝,提高搜索效率.

2 基于禁忌搜索的模糊聚类算法

2.1 禁忌搜索算法

禁忌算法又称禁忌搜索算法(Tabu Search, TS),是一种启发式算法,该算法使用局部搜索的方式来解决数学最优化问题,最早由 Ferd W. Glover 在 1986 年提出[13~15],它模仿了人类的记忆功能,从一个初始解 X 出发,在 X 的邻域内进行局部搜索,记录邻域内的最优解 X',并从 X 迁移到 X',再以 X' 为基准再次进行搜索,直到终止条件被满足,搜索中记录每一个曾经的最优点以防止循环的产生.

2.2 基于禁忌搜索的 FCM 算法

针对 FCM 算法初始敏感,易于收敛到局部最优甚至鞍点的不足,结合 TS 良好“爬坡”能力,以及较强局部搜索能力,本文提出了一种基于禁忌搜索的 FCM 算法(Tabu Search based Fuzzy C-Means, TSFCM). 算法思路如下:基于 FCM 算法得到一组最优解,在其邻域内进行禁忌搜索,对邻域候选解进行 FCM 迭代后判断禁忌,不断更新搜索中心与禁忌表,当搜索次数到达最大次数后将搜索过程中所记忆的最优解作为问题的解.

禁忌表:禁忌算法部分采用了简单短期记忆与长期记忆结合方式,短期表记忆到达点,长期表记忆迭代过程,一旦迭代被记忆的序列捕获则可立即重置迭代.

禁忌长度:禁忌长度可采用固定值或变化值. 使用变化值时,禁忌长度应与当前迭代次数逆相关,搜索后期禁忌的元素被禁忌的时间更短.

邻域:质心 V 为 $c*s$ 矩阵(c 为聚类数,s 为记录属性数),可视为高维空间内点,其邻域通常指其周围半径为 r 的区域,定义为:

$$N(V,r)=\{V1\,|\,V'-V|\leqslant r\} \tag{1}$$

的超球域. 其中半径 r 为搜索步长,影响邻域范围,从而影响“爬坡”的能力. 本文使用了一种变步长的搜索方式,当本次搜索的最优结果与基准比较差时,增加步长,在更大的范围内搜索;本次搜索较优时减少步长,在局部搜索求精. 考虑到搜索的多样化策略,为与当前解产生明显的差异,r 应有给出最小值,防止候选解移动过小,降低全局搜索能力. 本文受到物理中惯性现象的启发,针对超球邻域进行强化,邻域生成时参考上次 FCM 迭代时质心的移动轨迹,在最后一次聚类中心移动的

方向或反方向进行邻域采样. 该策略通过引入聚类中心运动轨迹与目标函数梯度下降信息,将邻域限制在一个更可能发现优秀解的范围内. 受群体初始化方式启发[16],邻域重复采样时引入混沌优化策略,设置混沌序列 $a=\{a_0,a_1,\cdots,a_p\}$,其中 $a_0\in[-1,1]$,序列由混沌自映射函数

$$a_{i+1}=1-2*a_i^2 \tag{2}$$

获得,由混沌序列 a 与步长 r 决定邻域采样聚类中心 V' 相对原点移动的距离.

禁忌判断:禁忌表中记录的是一高维空间点,但在空间中单独禁忌单点是没有意义的. Hathaway 的 FCM 局部收敛性研究[17,18]证明离收敛点足够近的点都将通过迭代收敛到收敛点,因此本算法对局部收敛区域进行禁忌. 另考虑到 V 表示 c 个聚簇中心坐标,簇顺序与聚类效果无关,因此需在禁忌判断时以固定的规则排序,排除次顺序的影响.

算法的具体步骤如下:

Step1　使用 FCM 算法得到问题的一组最优解 (U^*,V^*),U^* 为 $n*c$ 的模糊隶属度矩阵(n 为记录数,c 为聚类数). 记为全局最优解 $(U^{\text{global}},V^{\text{global}})$,将 V^* 加入空的禁忌表.

Step2　根据之前一次 FCM 迭代质心 V 的移动轨迹,在 V 最后一次移动方向(反方向)上,将 V^* 移动 $a*r$ 距离,得到邻域内的一个采样解 V',并记录(a 为混沌序列由式(2)给出,r 为动态搜索步长).

Step3　判断 V' 是否被禁忌,判断前对聚类中心进行排序,排除顺序影响. 计算 U',与禁忌表中的对象 V^* 进行比较,若:

$$\|(U^*,V^*)-(U',V')\|\leqslant\rho,0<\rho<\min_i\left(\sqrt{\sum_{i=1}(u_{ij})^m}\right) \tag{3}$$

$$u_{ij}\in U'(0\leqslant i\leqslant n,0\leqslant j\leqslant c) \tag{4}$$

则采样解被禁忌.

Step4　重复 Step2 和 Step3 p 次,完成邻域采样与禁忌判断,定义队列 $Q=\{V\}$,若 p 个结果全部禁忌,则 $Q=\{V_1,V_2,\cdots,V_p\}$,否则 $Q=\{V'|V'$未禁忌$\}$.

Step5　由 Step4 得到待搜索队列 $Q=\{V^k|k=1,2,\cdots l\}$,依次取 (U^k,V^k) 作初始值,运行 FCM 算法局部搜索得到新解 (U^{k*},V^{k*}),并计算评价指标 A^{k*}(A 越大则解越好),获得 A^{k*} 最大值 A^{local} 以及对应的局部最优解 $(U^{\text{loacl}},V^{\text{local}})$,若 $A^{\text{local}}>A^*$(A^* 为起始解 (U^*,V^*) 对应指标),则步长 $r=r-\Delta r$,否则 $r=r+\Delta r$. A^{local} 与已知全局最优解 $(U^{\text{global}},V^{\text{global}})$ 对应的 A^{global} 比较,若 $(U^{\text{loacl}},V^{\text{local}})$ 更佳,则更新全局最优,无论如何都将 $(U^{\text{loacl}},V^{\text{local}})$ 记为 (U^*,V^*).

Step6　将 (U^*,V^*) 加入禁忌表,并根据禁忌长度更新禁忌表,若到达最大搜索次数,则完成搜索,输出全局最优,否则返回 Step2.

2.3　TSFCM 算法分析

TSFCM 算法使用 TS 思想,增强了 FCM 的全局搜索能力. 影响算法效率的最关键因素为候选域的生成策略,其包括两部分,邻域的定义以及禁忌区域的定义. 邻域生成策略将聚类中心移动轨迹与动态步长结合,能够极大地增强搜索多样性,提升全局搜索能力,提升算法发现优秀解的可能. 禁忌域结合局部收敛特性,赋予禁忌判断一定的预测性,禁忌判断将不仅仅依据当前情况进行,甚至有能力预测迭代收敛后的结果,通过及时剪枝减小计算规模,从而提升搜索效率. 但由于难以严格的得出每个点的局部收敛半径,禁忌判断时统一采用了收敛半径上界,之后结合特赦规则进行了修正.

对于一个拥有 n 条记录,s 个特征值,分为 c 类的数据集,FCM 算法拥有时间复杂度上界 $O(n*s*c*t)$,其中 t 为迭代次数,TSFCM 基于 FCM 的扰动重复,因此具有最大时间复杂度 $O(n*s*c*t*p*T_{\max})$,其中 p 为邻域采样个数,$T_{\max}$ 为 TSFCM 迭代轮数. $p*T_{\max}$ 表示搜索的规模,是影响算法时间复杂度的关键参数. 当策略一定时搜索规模越大越容易发现优秀解,但同时也会引起效率的下降. 由于剪枝策略的存在,在参数合适的情况下,TSFCM 效率略优于单纯的 FCM 重复实验.

3　仿真实验

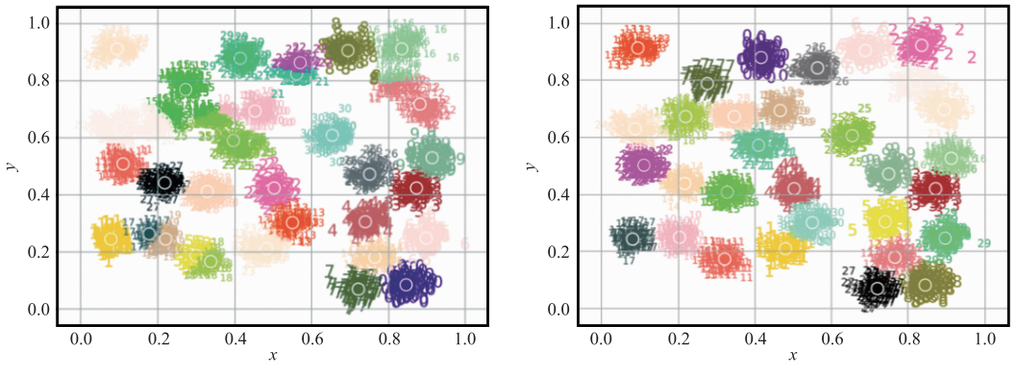
为了验证 TSFCM 算法的有效性与可行性,我们使用 UCI 标准数据库的 PENDIGITS 数据集、CHROAL HARMONY 数据集(简称 CHROAL 集)和 D31 数据集来进行实验. PENDIGISTS 集是一个有标注的用于分类问题的数据集,共有 10 类. 数据集由 30 位作家的 240 个手写数字组成,共包含 7200 条记录,每个记录包含 16 个取值为 0 ~ 100 整数的特征值,表示 8 组不同时间采样时笔的二维坐标. CHROAL 集总共 5665 条记录、102 个和弦类. 每个记录表示从音乐会中提取到的音乐片段,其包含 14 个特征值,表示 12 个音符在该音乐片段中是否出现以及该片段的低音、高音等级. D31 集包含 31 个类共3100 条记录,每个记录有 2 个特征值,表示二维平面上的坐标,每类样本 100 条数据.

实验采用最大搜索次数 $T_{\max}=40$,禁忌长度 $T_{\text{len}}=0.25*T_{\max}$,邻域采样频数 $p=5$. 为了减少样本属性数量级差异,对数据进行归一化处理,初始 V 通过随机方式生成,步长 r 取值范围为 $0.05\leqslant r\leqslant 0.5$,$\Delta r=0.05$. 使用外部指标 FMI(Fowlkes and Mallows Index)衡量聚类分析的准确率,并同时作为 TS 的评价指标 A,FCM 与 TSFCM 中目标函数值表示为 J,其含义为隶属度加权的类内距离和.

实验 1 对 D31 集、CHORAL 集、PENDIGITS 集运行 FCM 算法 k 次，由先验实验（重复对数据进行 FCM 聚类）中频数最高的结果（0.82/0.28/0.51）为起点进行禁忌搜索（$p * T_{max} = k$），分别统计重复进行 FCM 聚类的最佳结果和 TSFCM 聚类的最高频次结果以及运行时间，取普遍解（0.82/0.28/0.51）周围 ±0.005 的解求平均耗时作为获取普遍解的时间消耗，对 TSFCM 收敛时间进行修正（实验如表 1 所示）. FCM 算法与 TSFCM 算法在 D31 集上聚类结果如图 1 所示.

表 1 FCM 与 TSFCM 在 D31、CHORAL、PENDIGITS 上的运行结果

	D31（$k=200, p=5, T_{max}=40$）		CHORAL（$k=75, p=5, T_{max}=15$）		PENDIGITS（$k=50, p=10, T_{max}=5$）	
	FCM	TSFCM	FCM	TSFCM	FCM	TSFCM
J	2.9069178	2.7710301	134.41562	132.667782	1059.7611	1059.0609
FMI	0.9005747	0.9530452	0.313259	0.40702	0.608252	0.6106453
time/s	226.46286	233.805714 + 1.132314285	512.59505	495.51586 + 6.8346	228.83811	138.625904 + 3.882833



(*a*) FCM算法聚类效果图 (*b*) TSFCM算法聚类效果图

图1 FCM算法与TSFCM算法在D31集上聚类结果

在先验实验中 FCM 算法陷入局部最优，几乎不能发现问题的最优解. 而根据本次实验，TSFCM 算法在 D31 集上从准确率 0.82 的普遍解出发，经过 40 轮禁忌搜索后发现了准确率 0.95 的最优解，准确率远远超过了 FCM 算法，从图 1（*b*）上可以看出，该解没有明显的聚簇错误，基本上可以认为到达了全局最优解，但与原算法相比效率上略有下降. 在 CHORAL 集上，TSFCM 算法得到了准确率 40% 的解，准确率上的优化提升明显，但在效率上与原算法相比并没有明显优势. 在 PENDIGITS 集上 TSFCM 算法同样从一个普遍解出发，搜索到问题的更优解，同时还花费了更少的搜索时间. 实验中 TSFCM 最优时搜索到准确率 0.62 的解，准确率较原算法提升 11%. 实验证明 TSFCM 算法与重复 FCM 算法相比，在准确率与运行效率上都有一定的优势. 另外在实验中发现一普遍现象，对 D31 数据集，TSFCM 算法虽然得到了一个更高准确率的解，但目标函数值却反而增高，准确率与目标函数值呈现非同步变化，如表 2 所示.

该问题由距离度量引起. 由于采用了欧式距离，针对椭圆形的聚簇，将其分为 2 簇会得到更小的 J，但该分类与标注不符，所以准确率更低. 该问题不在本文的讨论范围内，将在之后进行更深的研究.

实验 2 参数对 TSFCM 算法性能的影响. 控制 $T_{max}=5$，p 分别为 5、10、15、20，在 D31 集上以普遍解为起点，运行 50 次 TSFCM，统计搜索准确率与耗时，同时与 $T_{max}=10, p=5$ 实验结果进行对比. 结果如表 3 示.

表 2 FCM 中 *J*，FMI 不同步变化情况

	J	FMI
1	3.423342623	0.7440140764
2	3.545779023	0.759438428
3	3.038284103	0.8405597162
4	3.107196866	0.8416278711
5	2.903849084	0.8962414514
6	2.913239402	0.9000018922

表 3 TSFCM 参数对准确率与耗时的影响

		AVG FMI	AVG TIME	MAX FMI
$p=5$	$T_{max}=5$	0.87008593	31.7099186	0.91328127
$p=10$	$T_{max}=5$	0.86641489	64.4160482	0.95316319
$p=15$	$T_{max}=5$	0.86613992	97.8494222	0.95316319
$p=20$	$T_{max}=5$	0.86701606	124.718267	0.9133655
$p=5$	$T_{max}=10$	0.91123054	59.3424727	0.95377849

根据实验结果可得，算法的耗时变化基本上与 $p * T_{max}$ 成正比，而当迭代轮数过低时，单纯的增加邻域采样频率并无法显著的提高算法的搜索能力，这是由于算法在一个邻域内采用相同的步长，轮数过低时变步长的优势无法体现，全局搜索能力受到影响. 相反在采样频率足够的情况下，适当的提升迭代次数可以获得不错的结果.

如图 2 所示,增加采样频率使得发现更好搜索结果的频率增高,但改进有限,增加轮数则获得了较大的提升.

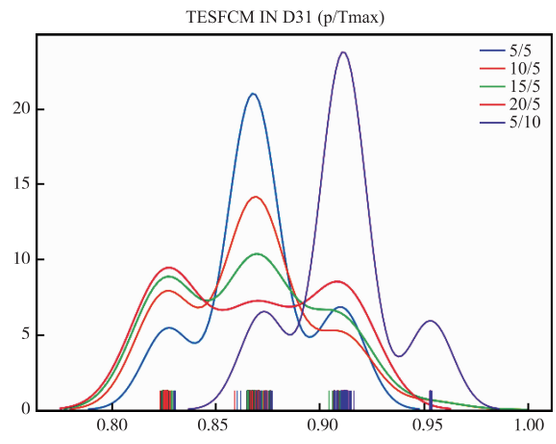


图2　D31上不同参数TSFCM搜索准确率频数

采样频率 p 表征算法对邻域的感知程度,过低会导致对邻域进行误判,遗漏可能的优秀解;过高则会导致冗余计算,降低算法效率. 迭代轮数 T_{max} 表征聚类中心的移动次数,过低导致搜索覆盖区域有限,全局搜索能力差;过高则会在搜索至最优解后继续向其他方向搜索,增加算法耗时. 合适地选择 p 与 T_{max} 可以在效率与准确率间达到平衡.

实验 3　为了评价 TSFCM 算法的改进效果,本实验在 PENDIGITS 集上分别运行 100 次 TSFCM 与 PS-FCM[7](原算法中 PSO 部分及 FCM 部分),PSO-FCM[8]、EFCMIDPSO[9]、AF-FCM[11]、GA-FCM[4](原算法中遗传算法优化 FCM 部分)、EABCK[12],同样使用 FMI 作为搜索准确率,比较各算法的准确率与时间消耗. 为控制变量,各算法的理论时间复杂度被设置为相同值(每次算法运行共进行 100 轮搜索):TSFCM 中 $p=5$,$T_{max}=20$,步长 r 有 $0.1\leqslant r\leqslant 1$,$\Delta r=0.1$. PS-FCM、PSO-FCM 与 EFCMIDPSO 中粒子个数 $p=5$,最大迭代轮数 $T_{max}=20$. PS-FCM 惯性权重 $\omega=0.72$,PS-FCM 惯性权重 $\omega=1$,学习因子 $c_1=c_2=2$,EFCMIDPSO 初始惯性权重 $\omega=0.9$,初始学习因子 $c_1=c_2=2$. AF-FCM 中鱼群数量 $p=2$,每条人工鱼每轮依次进行觅食、聚群、追尾动作,迭代轮数 $T_{max}=17$. GA-FCM 中种群大小 $p=20$,迭代轮数 $T_{max}=5$,交叉概率 $p_c=0.9$,变异概率 $p_m=0.001$. EABCK 中种群数量 $p=6$(其中雇佣蜂数量为 3),迭代轮数 $T_{max}=17$,当 5 次蜜源没有更新时雇佣蜂变为侦察蜂(AF-FCM 与 EABCK 搜索次数为 102 次,对耗时进行修正). 实验统计结果如表 3 所示. 根据实验准确率数据绘制算法准确率频数分布图(图 3)、算法准确率提琴图与箱式图(图 4),提琴图中标记准确率均值,箱式图中标记准确率中位值.

实验结果分析可得,TSFCM 平均收敛速度与 PS-FCM 算法基本处于相同水平,耗时略高 3%. PS-FCM 最大准确率 0.60 略低于 TSFCM 最大准确率 0.63,且平均准确率远低于 TSFCM 算法,由图 4 可以看出 PS-FCM 算法搜索准确率集中在 0.49 ~ 0.51 区间,箱式图舍弃离群点后,其最大值甚至远低于 TSFCM 搜索最差值,PS-FCM 算法虽然与原 FCM 算法相比全局搜索能力有所加强,但其局部搜索能力太弱,在有限的搜索轮次下难以稳定的得到优秀解,TSFCM 算法与其相比有着很好的鲁棒性,与搜索性能.

表 4　TSFCM 等算法百次实验耗时/准确率统计

Alg.	MAX FMI	AVG FMI	AVG TIME/s
PS-FCM	0.603607	0.502857	274.689488
PSO-FCM	0.613519	0.590959	399.326333
EFCMIDPSO	0.632179	0.623086	341.915152
AF-FCM	0.623227	0.539012	419.494606
GA-FCM	0.627672	0.5944946	369.9082268
EABCK	0.619380	0.542453	291.046937
TSFCM	0.63133	0.620369	283.660521

PSO-FCM 算法在 PS-FCM 算法的基础上使用 FCM 算法初始化粒子群,并针对变动粒子进行了 FCM 迭代强化操作,搜索准确率有了很大的提高,且更加稳定的得到较优解. 但准确率均值、中值、最值均低于 TSFCM 算法,且时间消耗大大增加,约增加 40%. 此外,TSFCM 算法只包含少数参数且可以朴素地设置,避免了 PSO 算法中学习因子等参数的确定,还有着更低的调参难度. TSFCM 算法在收敛速度以及搜索准确率上全面占优,算法也更简单.

与 EFCMIDPSO 相比,TSFCM 几乎得到了相同的最优解. 并且在实验中,两种算法的准确率均值、中位数基本相同,可以认为 TSFCM 算法有着几乎和 EFCMIDPSO 相同的全局搜索优化能力. 从算法平均耗时来看,TSFCM 算法有着明显的优势,与 EFCMIDPSO 相比平均耗时减少 21.01%. 基于 FCM 局部收敛性质的禁忌策略极大的提升的算法的收敛速度,且基本没有影响算法的搜索能力.

由图 4 得 AF-FCM 算法的搜索均值、中值都偏低,箱式图中除去前 1/4 离群点后,甚至准确率低于 TSFCM 均值,在同等水平的参数设置下,AF-FCM 算法的全局搜索能力远低于 TSFCM 算法. 从算法效率上来看,AF-FCM 算法收敛速度远慢于 TSFCM 算法,TSFCM 算法与之相比有着全面的优越性.

由表 4 结果可知,GA-FCM 算法凭借遗传变异算子带来的强大全局搜索能力得到了优秀的收敛解,由图 4 可以看出算法有着不错的鲁棒性. 但 TSFCM 算法在准确率的均值、中值、最值上都优于 GA-FCM 算法,收敛解更加稳定地集中在 0.60 ~ 0.62 的范围内,有着更强的鲁棒性. 并且在收敛速度上,TSFCM 算法比 GA-FCM 算法快约 23.3%,极大地减少了搜索时间. 因此 TSFCM 算法优于遗传算法优化的 FCM 算法(GA-FCM).

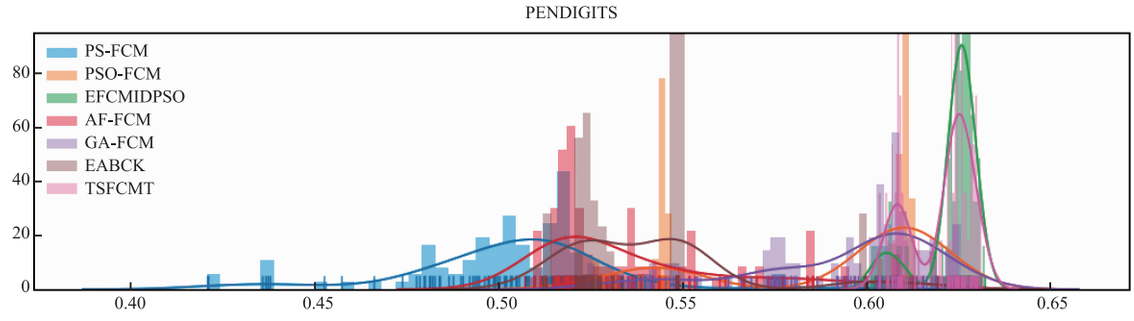


图3 TSFCM等算法百次实验准确率频数分布图

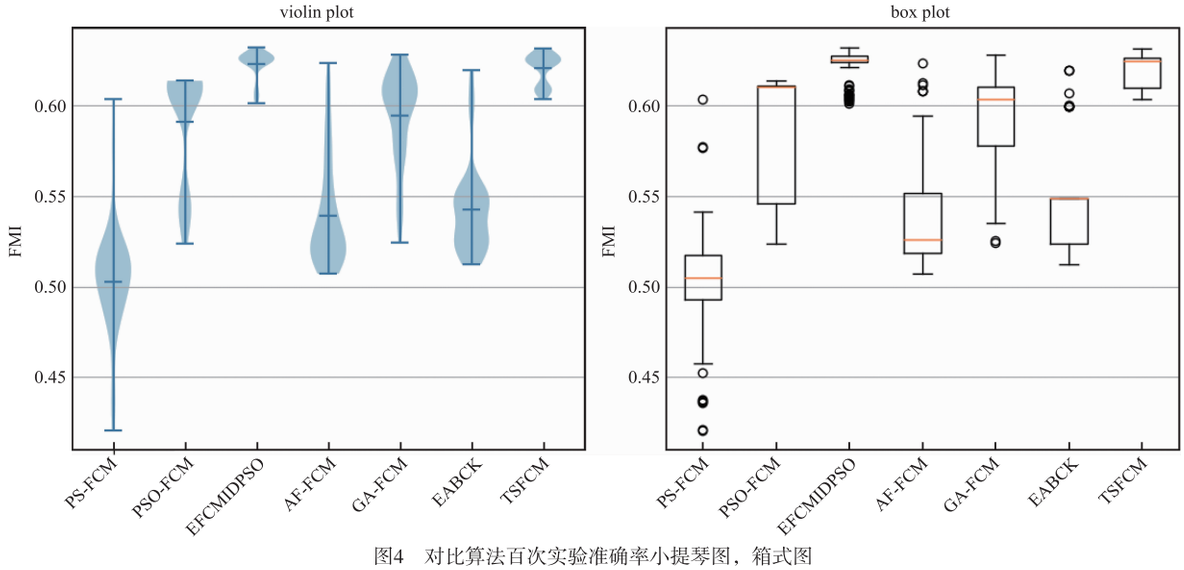


图4 对比算法百次实验准确率小提琴图，箱式图

EABCK 算法结合了粒子群与遗传算法，与 GA-FCM 算法相比收敛速度大大提高，但仍略差于 TSFCM 算法. 然而 EABCK 算法准确率集中分布在 0.52 和 0.55 附近，全局搜索能力差，最优准确率低于 TSFCM 算法，平均准确率远低于 TSFCM 算法. 因此无论从准确率还是收敛速度上，TSFCM 算法都优于该算法.

综上，TSFCM 算法在搜索能力基本与 EFCMIDPSO、GA-FCM 相近的情况下，有着更快的收敛速度；在耗时略优于 EABCK 的情况下，有着更稳定且更强的全局搜索能力，更加难以陷入局部最优；无论在全局搜索能力还是收敛速度上都优于 PS-FCM、PSO-FCM、AF-FCM 算法，且参数设置更为简单. 因此 TSFCM 算法与上述群智算法和遗传算法改进的 FCM 算法相比具有优越性.

4 总结

针对 FCM 算法初始点敏感，易收敛至局部最优的缺陷，本文基于 TS 算法的“爬坡”能力提出了 TSFCM 算法. 算法在禁忌搜索部分使用了变步长的结合历史信息的邻域生成方案，增强了搜索的多样化能力；使用 FCM 迭代作为候选解的强化规则，提升局部搜索能力；使用了基于 FCM 局部收敛性质的长期表改进方案，给予算法一定程度的预测能力而通过剪枝提高了搜索效率. 实验结果显示，改进算法在聚类准确率上有明显的提升并且有很强的鲁棒性. 与粒子群、人工蜂群、人工鱼群等群智算法以及遗传算法优化的 FCM 算法相比，本算法在准确度、收敛速度、鲁棒性等方面都有不同程度的提升. TSFCM 算法在一些细节策略上还有强化的空间，之后在算法效率提升、全局搜索能力提升以及高纬海量数据处理方向可进行进一步的研究，并将算法应用于图像分割领域使其更具实用价值.

参考文献

[1] Zadeh L A. Similarity relations and fuzzy orderings[J]. Inf. Sci. ,1971,3(2):177 - 200.

[2] Bezdek J. C. Aconvergence theorem for the fuzzy ISODATA clustering algorithms[J]. IEEE Trans. Pattern Anal. Mach. Intell. (USA),1980,PAMI-2(1):1 - 8.

[3] 李凯，曹喆. 一种基于神经网络的广义熵模糊聚类算法

[J]. 电子学报,2016,44(8):1881 – 1886.
Li Kai, Cao Zhe. A fuzzy clustering algorithm with generalized entropy based on neural network[J]. Acta Electronica Sinica, 2016, 44(8): 1881 – 1886. (in Chinese)

[4] 唐成华,刘鹏程,汤申生,谢逸. 基于特征选择的模糊聚类异常入侵行为检测[J]. 计算机研究与发展,2015,52(3):718 – 728.
Tang Chenghua, Liu Pengcheng, Tang Shensheng, Xie Yi. Anomaly intrusion behavior based on fuzzy clustering and features selection[J]. Journal of Computer Research and Development, 2015, 52(3): 718 – 728. (in Chinese)

[5] Katayoon Ahmadi, Abbas Karimi, Babak Fouladi Nia, Thomas Desaive. Newtechnique for automatic segmentation of blood vessels in CT scan images of liver based on optimized fuzzy c-means method[J]. Computational and Mathematical Methods in Medicine, 2016, 2016: 1 – 8.

[6] 周双,冯勇,吴文渊. 一种识别关联维数无标度区间的新方法[J]. 物理学报,2015,64(13):36 – 41.
Zhou Shuang, Feng Yong, Wu Wen-Yuan. A novel method to identify the scaling region of correlation dimension[J]. Acta Physica Sinica, 2015, 64(13): 36 – 41. (in Chinese)

[7] 余晓东,等. 基于粒子群优化的直觉模糊核聚类算法研究[J]. 通信学报,2015,36(5):78 – 84.
Yu Xiaodong, et al. Research on PSO-based intuitionistic fuzzy kernel clustering algorithm[J]. Journal on Communications, 2015, 36(5): 78 – 84. (in Chinese)

[8] Zhongxing Zhang. Intrusion detection network based on fuzzy c-means and particle swarm optimization[A]. IEEE Beijing Section. Proceedings of 2014 International Conference on Industrial Engineering and Information Technology[C]. IEEE Beijing Section, 2014. 4.

[9] 耿宗科,等. 基于模糊 c-means 与自适应粒子群优化的模糊聚类算法[J]. 计算机科学,2016,43(8):267 – 272.
Geng Zongke, et al. Fuzzy c-means and adaptive PSO based fuzzy clustering algorithm[J]. Computer Science, 2016, 43(8): 267 – 272. (in Chinese)

[10] David Gabriel de Barros Franco, Maria Teresinha Arns Steiner. Clustering of solar energy facilities using a hybrid fuzzy c-means algorithm initialized by metaheuristics [J]. Journal of Cleaner Production, 2018, 191: 445 – 457.

[11] 皇甫中民,等. 鱼群启发的三维 CAD 模型聚类与检索[J]. 计算机辅助设计与图形学学报,2016,28(8):1373 – 1382 + 1392.
Huangfu Zhongmin, et al. 3D CAD model clustering and retrieval inspired by fish swarm[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(8): 1373 – 1382 + 1392. (in Chinese)

[12] TRAN Dang Cong, WU Zhijian, WANG Zelin, DENG Changshou. A novel hybrid data clustering algorithm based on artificial bee colony algorithm and *k*-means[J]. Chinese Journal of Electronics, 2015, 24(4): 694 – 701.

[13] Fred Glover. Future paths for integer programming and links to artificial intelligence[J]. Computers and Operations Research, 1986, 13(5): 533 – 549.

[14] Fred Glover. Tabu search-part1[J]. ORSA Journal on Computing, 1989, 1(2): 190 – 206.

[15] Fred Glover. Tabu search-part2[J]. ORSA Journal on Computing, 1990, 2(1): 4 – 32.

[16] 朱喜华,李颖晖,李宁,范炳奎. 基于群体早熟程度和非线性周期振荡策略的改进粒子群算法[J]. 通信学报,2014,35(2):182 – 189.
Zhu Xihua, Li Yinghui, Li Ning, Fan Bingkui. Improved PSO algorithm based on swarm prematurely degree and nonlinear periodic oscillating strategy[J]. Journal on Communications, 2014, 35(2): 182 – 189. (in Chinese)

[17] Bezdek J. C. Convergence theory for fuzzy c-means: counterexamples and repairs[J], IEEE Trans Syst Man Cybern. 1987, SMC-17: 873 – 877.

[18] Hathaway R. J. Local convergence of the fuzzy c-means algorithm[J]. Pattern Recognition, 1986, 6: 477 – 480.

作者简介

朱　毅　男,1987 年出生,浙江温州人,2013 年毕业于华中科技大学获得硕士学位,现为 2013 级华中科技大学软件学院博士生,主要从事大数据、深度学习等方面的研究.

杨　航　男,1995 生,河南安阳人,2017 年毕业于华中科技大学软件学院,主要从事聚类分析,图像处理等方面的研究.

吕泽华(通信作者)　男,1976 年出生,湖北宜昌人,2007 年毕业于华中科技大学获博士学位,控制科学与工程专业博士后,现为软件学院副教授,硕士生导师,主要从事模糊推理,图像处理与模式识别,数据挖掘等方面的研究.
E-mail: lvzehua@ hust. edu. cn