DOI:10.3979/j.issn.1673-825X.2020.02.019

基于精英个体划分的变步长萤火虫算法的特征选择方法

刘　磊,罗　蓉,尹　胜
(重庆邮电大学 先进制造工程学院,重庆 400065)

摘　要:针对标准萤火虫算法(firefly algorithm,FA)收敛速度慢及其在解空间内的搜索易陷入局部最优的缺陷,充分考虑萤火虫算法在寻优过程中其种群内个体的差异性,提出一种基于精英萤火虫个体划分的变步长策略,改进后的 FA 在算法迭代中对每代目标值较好的精英萤火虫个体随机增大其移动步长,而对每代目标值较差的非精英个体则线性减小其步长。为适用于特征选择问题,又对 FA 中萤火虫的编码和位置移动进行了离散化定义,给出了基于所提改进型离散 FA(binary firefly algorithm,BFA)的包装式特征选择方法流程。在 UCI 分类数据集上对比测试了所提改进型 BFA 与其他算法在优化特征选择方面的性能。测试结果表明,基于所提改进型 BFA 优化特征选择的效果较好,验证了所提改进策略可有效提升 FA 的优化能力。

关键词:特征选择;离散萤火虫算法(BFA);变步长
中图分类号:TP301.6　　**文献标志码**:A　　**文章编号**:1673-825X(2020)02-0313-09

Feature selection method based on the dynamic step firefly algorithm with the elite individual dipartition

LIU Lei, LUO Rong, YIN Sheng
(School of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China)

Abstract: In order to overcome the shortcomings of the firefly algorithm (FA) such as slow convergence speed and falling into local optimum of solution space, the paper fully considers the individual differences among the population of fireflies when FA is running for optimization. And a dynamic step strategy with the elite firefly individual dipartition is proposed, which randomly increases the moving step of elite firefly individuals with better target values in every algorithm iteration but linearly reduces the moving step of non-elite individuals with poor target values. The coding and positional movement of fireflies of FA are defined in binary mode for the feature selection problem, then the procedure of the wrapped feature selection method based on the improved binary FA (binary firefly algorithm, BFA) is presented. Finally, a comparative test of feature selection optimization ability between the improved BFA and other algorithms is executed on UCI classification datasets. The test result shows the improved BFA has better optimization effect on feature selection than other algorithms, verifying that the proposed improvement strategy can effectively improve the optimization ability of FA.

Keywords: feature selection; binary firefly algorithm (BFA); dynamic step

收稿日期:2019-01-14　**修订日期**:2020-02-26　**通讯作者**:刘　磊　liuleifeichuan@foxmail.com
基金项目:国家自然科学基金(51805066)
Foundation Item: The National Natural Science Foundation of China (51805066)

0 引 言

特征选择(feature selection)是机器学习、数据挖掘任务中,对于高维或冗余特征数据集采取的一种数据预处理方法[1]。特征选择可以移除原数据集中冗余、不相关的噪音特征,从原数据集中获取一个最优的特征子集以提升机器学习模型的预测精度,同时最优特征子集相较于原特征集具有更少的特征数据,利用其进行机器学习模型的训练和预测可以极大地降低运行时间与数据存储方面的开销[2]。机器学习领域中主要有 2 类特征选择方式:过滤式(filter mode)和包装式(wrapper mode)[3]。包装式通常会取得比过滤式更好的特征选择结果,包装式比过滤式占用更长运算时间,在非海量数据规模且机器学习模型固定的情况下宜选用这种方式[4]。本文的特征选择方法为包装式的方法。特征选择算法像基于贪心或完备思想的算法[5]虽可在原始数据集上找到最优特征子集,但这些算法为找到全局最优解需要付出高昂的计算代价[6]。特征选择是一个 0-1 规划的组合优化问题[7],被证明是 NP(non-deterministic polynomial)难题[8],而一些元启发式或生物启发式优化算法已在求解此类 NP 难组合优化问题中表现出了良好的性能[9],其在解决特征选择问题方面的研究[10-12],近年来受到学术界越来越多的关注[13],遗传算法(simple genetic algorithm,SGA)[14-15]与离散粒子群优化(binary particle swarm optimization,BPSO)[16-17]等传统生物启发式算法已成功用于特征选择优化。

萤火虫算法(firefly algorithm,FA)是新型的生物启发式群智能优化算法[18],FA 因具有多目标优化能力[19],使其被国内外学者应用于旅行商问题[20-21]、背包问题[22]、系统效能优化[23]、特征选择[24-27]等组合优化问题的研究中。标准 FA 设置参数少,算法步骤和公式较简单,然而其步长参数取值大小对其性能影响很大[22],标准 FA 采用固定步长设置,这使 FA 在优化求解中全局最优解发现率低、算法搜索易陷入局部最优。对此,文献[27]提出一种随算法迭代次数增加而算法步长逐渐减小的改进策略,然而该步长调整方式忽视了 FA 种群中萤火虫个体在寻优过程中的差异性,无法使个体自适应调整寻优步长,且在算法迭代优化的后期寻优效率降低,这存在着寻优局限性。因此,本文充分考虑了 FA 在寻优过程中其种群内个体的差异性,在遵循标准 FA 算法机理的基础上,提出了基于精英个体划分的变步长 FA(elite -individual -dipartition dynamic step firefly algorithm,EDSFA),并将 EDSFA 进行离散化实现,提出 EDSFA 的相应离散化算法 EDSBFA(elite -individual -dipartition dynamic step binary firefly algorithm,EDSBFA)用于机器学习任务中的特征选择优化,实验表明,在优化特征选择方面 EDSBFA 的总体性能优于固定步长 BFA(binary firefly algorithm)、文献[27]所提变步长 BFA 以及传统启发式优化算法 SGA 和 BPSO。

1 精英个体划分的变步长 FA

1.1 标准 FA

标准 FA 算法原理:萤火虫种群中的每只萤火虫个体的位置代表待求解问题的一个可行解,即整个种群为原问题的解空间,萤火虫自身亮度与待求解问题的目标函数值相关,即萤火虫亮度越强,对应可行解的目标函数值越佳。当 FA 算法运行时,荧光亮度较弱的萤火虫被荧光亮度较强的萤火虫吸引,并向其所处位置移动,随着算法的不断迭代,种群中原先亮度较弱的萤火虫持续不断地移向亮度较强个体所处位置,最终种群中大多数萤火虫将聚集在一个或多个亮度较强的萤火虫周围,当萤火虫间没有亮度差异时算法收敛,而最亮萤火虫所处位置代表问题的最优解[19]。

在标准 FA 中,萤火虫个体遵循以下 3 条规则:①所有个体无性别区分;②个体吸引力与自身亮度相关,较弱亮度的个体会被较强亮度的个体所吸引而向其移动,最强亮度的个体将在解空间中随机移动;③待求问题的可行解的目标函数值通常作为个体的亮度值。

1)萤火虫的相对亮度为

$$I(\gamma) = I_0 e^{-\gamma r^2} \tag{1}$$

(1)式中:r 为萤火虫之间的距离;γ 为光强吸收系数,考虑了光在传播过程中因空气等介质会有所损耗,可设为常数;I_0 为 $r=0$ 时萤火虫的荧光亮度,即自身荧光亮度,其值与目标函数值有关,目标函数值越优,自身荧光亮度越强。

2)萤火虫的吸引力为

$$\beta(r) = \beta_0 e^{-\gamma r^2} \tag{2}$$

(2)式中,β_0 为 $r=0$ 时萤火虫的自身吸引力。

3)萤火虫之间的距离为

$$r_{ij} = \| \boldsymbol{x}_i - \boldsymbol{x}_j \| = \sqrt{\sum_{k=1}^{D} (x_{ik} - x_{jk})^2} \tag{3}$$

(3)式中:若两萤火虫个体 i 和 j 分别位于 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$,它们之间的距离可用欧式距离 r_{ij} 表示;D 为问题的解空间维度;x_{ik},x_{jk} 分别表示第 i 与 j 只萤火虫的位置向量 $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ 中第 k 维的分量。

萤火虫的位置移动为

$$\boldsymbol{x}_i(t+1)=\boldsymbol{x}_i(t)+\beta_0*\mathrm{e}^{-\gamma r_{ij}^2}(\boldsymbol{x}_j(t)-\boldsymbol{x}_i(t))+\alpha*\boldsymbol{\varepsilon} \tag{4}$$

发光亮度较弱的萤火虫 i 将移向发光亮度较强的萤火虫 j,从而 i 的位置向量发生更新。

(4)式中,t 指算法迭代计数。等号右端由 3 项因式构成:第 1 项$\boldsymbol{x}_i(t)$表示萤火虫 i 当前所在解空间位置;第 2 项表示萤火虫 i 向萤火虫 j 移动的距离;第 3 项 $\alpha\cdot\boldsymbol{\varepsilon}$ 为 i 向 j 移动的过程中伴有的随机扰动,避免萤火虫过早陷入局部最优,其中,α 为随机移动步长,其取值为 0~1,$\boldsymbol{\varepsilon}$ 是由高斯分布、均匀分布或莱维飞行[28]所得到的随机数向量。

1.2　基于精英个体划分的变步长策略

标准 FA 中萤火虫位置移动方式如(4)式,其中,移动步长 α 的引入是为了扩大萤火虫的移动视野,提高萤火虫种群的多样性,避免算法早熟收敛。步长参数 α 取值大小应与具体解空间的搜索范围相关,在较小搜索范围内 α 取值过大可能导致算法无法收敛;取值过小,随机移动距离极小,无法增加种群多样性,不能起到扩大搜索范围或扩大萤火虫移动视野、避免算法过早收敛的作用[22]。在标准 FA 中所有萤火虫个体的步长取统一固定值,不能根据实际搜索情况进行调节,这具有很大的局限性。文献[27]提出一种步长 α 随算法迭代次数增加而动态减小的变步长 FA,相比标准 FA 扩大了步长 α 的调节范围,在算法运行初始阶段能增强算法的随机搜索能力,跳出局部最优,算法迭代后期搜索的随机性减弱,促使算法逐步收敛。该变步长改进策略没有考虑萤火虫个体在寻优搜索中的差异性,在算法搜索中为所有个体设置一样的步长,在算法迭代后期由于步长值被减到很小,难以使已处在局部最优位置的大部分萤火虫更进一步扩大寻优视野,这极易使整个萤火虫种群的搜索陷入局部最优陷阱,另外算法跌代后期整个种群采取统一的小步长值也降低了算法收敛的效率。对所有萤火虫个体采取大步长,能增强 FA 对解空间的全局搜索能力,但易跳过全局最优而求解精度低;对所有个体采取小步长,能增强 FA 在局部解空间的探索能力,但易使萤火虫寻优陷入局部最优陷阱,且会降低算法求解效率。所以为每只萤火虫个体的移动单独设置动态变化的步长是必要的。本文从萤火虫种群的精英个体划分角度出发,提出一种变步长策略以提高 FA 的寻优能力。

为了兼顾算法在解空间的全局搜索和局部探索能力,同时充分考虑每只萤火虫的性能差异,本文提出如下的步长调整策略:对每轮次算法迭代中种群里性能较好的精英个体增大其步长,保持其全局移动视野以搜索更大区域,提高了种群的多样性,进而降低搜索陷入局部最优陷阱的风险;而对于当前迭代轮次中的非精英个体减小其步长,保持其在局部解空间的探索能力,小步长值对精确搜索有利,能提高算法求解精度和促使种群逐步收敛。基于精英个体划分的变步长设置具体表示为

$$\alpha_i(t+1)=\begin{cases}\alpha_i(t)+\alpha_0\cdot a/\mathrm{Max}Gen, & I_i>\theta\cdot I_{\mathrm{best}}\\ \alpha_i(t)-\alpha_0/\mathrm{Max}Gen, & \text{其他}\end{cases} \tag{5}$$

$$\alpha_i(t+1)=\alpha_0,\ \alpha_i(t)>1 \tag{6}$$

(5)—(6)式中:$\alpha_i(t+1)$表示第 i 只萤火虫个体在第 $t+1$ 次算法迭代时的步长;MaxGen 指设置的算法最大迭代次数;α_0 表示初始统一步长值;a 为步长增大的随机加速度;a 取值为[1, MaxGen/2]的随机整数;I_i 为个体 i 的荧光亮度;I_{best}为当前种群中最佳个体的荧光亮度;θ 为用于划分精英个体的阈值,θ 为[0.85,0.95]的常数值,由(5)式知当个体 i 的亮度大于当前种群中最佳个体亮度值的 θ 倍时,便视其为精英萤火虫个体,增大其步长值,否则线性减小其步长值,如果 i 的步长经过增大调整后已大于 1,则按(6)式所示将其步长重置为初始步长值。对于(5)式中步长增大的随机加速度 a 的引入可使萤火虫个体间步长的扩大趋于多样化、随机性,这也促进了种群的多样性变化,避免了种群的早熟收敛,另外 a 可结合具体解空间的搜索范围进行滑动窗口取值,使步长调节更具灵活性。

2　基于 EDSBFA 的特征选择方法

2.1　萤火虫位置向量编码

特征选择实际上是从原始数据集的 M 个数据特征中选择 N 个特征后组成一个特征子集($M>N$),进而用该特征子集中的特征优化后续的数据分析处

理,对于每个待选特征而言只存在“入选”或“落选”2 种状态,因此,可将萤火虫个体的位置向量每一维的索引序号对应于原数据集各维特征的索引序号,并将个体向量编码为每一维元素仅为‘0’或‘1’的二元离散向量,其中,元素为 0 表示落选,元素为 1 表示入选,向量维度等于原数据集特征维度。例如萤火虫 i 当前位置向量为 $\boldsymbol{X}_i=[0,1,0,1,0,1,0,1,1,0]$,其向量维度为 10,表示原数据集的 10 个数据特征中入选特征的索引序号分别为“2,4,6,8,9”,该索引序号所对应特征即被选择。

2.2 目标函数定义

特征选择的目标是从原始数据集选择一个特征数量较少的特征子集,利用该特征子集中的特征做数据挖掘,使机器学习模型获取更好的预测准确率。目标函数应考虑选择的特征数量和预测准确率这 2 个因素,当所选特征数量越少和模型预测准确率越高时,目标函数值越优,本文中萤火虫的自身亮度等于萤火虫位置向量的目标函数值,所以目标函数值越优,萤火虫的亮度越强。目标函数定义为

$$f(accu,num\mid \boldsymbol{X}_i)=\frac{10^3}{10^4\times(1-accu)+k\times num} \tag{7}$$

(7)式中:$\boldsymbol{X}_i$ 为萤火虫个体 i 的位置向量;$accu$ 为个体向量对应特征子集的预测准确率;num 为该特征子集中入选特征的数量;k 为特征数量 num 的权重,本文实验中 $k=0.5$。

2.3 EDSFA 离散化为 EDSBFA

由于特征选择问题是组合优化问题,属于离散优化范畴,本文萤火虫的位置向量编码是 0,1 二元离散编码,标准 FA 中对萤火虫的距离、位置移动的定义只适用于连续优化领域而不适用于离散优化,故须对萤火虫的距离、位置移动做适用于离散化操作的重新定义。对于提出的基于精英个体划分的变步长萤火虫算法(EDSFA),本文采用了与文献[27]一样的萤火虫算法离散化方式,将 EDSFA 离散化为 EDSBFA。

定义 1 萤火虫 i 与 j 之间的距离。由于本文萤火虫位置向量被编码为 0,1 二元离散向量,所以 2 个萤火虫之间的距离或差距使用汉明距离描述比使用原 FA 中的欧式距离更适合,汉明距离准确刻画了 2 个向量的差异性,比欧式距离的计算开销少,提高了算法运行效率。两萤火虫个体间的归一化汉明距离定义为

$$r_{i,j}=1-\frac{\sum_{k=1}^{d}\mid x_{ik}\oplus x_{jk}\mid}{d} \tag{8}$$

(8)式中:⊕指 XOR 异或操作;d 为个体向量的维度。萤火虫之间的吸引力 β 通过(2)式计算,其中距离 r 使用上述定义的归一化汉明距离。

定义 2 萤火虫的离散化移动。当萤火虫 i 向亮度更强更有吸引力的萤火虫 j 移动时,个体 i 的位置向量每一维元素值将做决策是否发生改变,本文将个体位置向量中每一维元素值的改变分 2 步进行:①吸引移动,如(9)式,对应于(4)式右端第 2 项因式所作操作;②随机游走,如(10)式,对应于(4)式右端第 3 项因式所作操作。其中,$rand(0,1)$ 指 0~1的随机数,α_i 为个体 i 的当前步长,v_{ik} 是个体 i 的位置向量的第 k 维分量在“吸引移动”后的中间变量。

$$v_{ik}=\begin{cases}x_{jk}, x_{ik}\neq x_{jk}, \beta>\mathrm{rand}(0,1)\\ x_{ik}, \text{其他}\end{cases} \tag{9}$$

$$x_{ik}=\begin{cases}1-v_{ik}, \alpha_i>\mathrm{rand}(0,1)\\ v_{ik}, \quad \text{其他}\end{cases} \tag{10}$$

2.4 基于 EDSBFA 的包装式特征选择方法流程

包装式特征选择,使用优化算法从原始数据集中选择特征,产生一系列待评价特征子集,将各个特征子集对应的特征数据送入机器学习分类器进行分类器的训练,并用训练好的分类器做分类预测,经过数次迭代优化,最终将能使分类器获得最好精度和泛化能力的特征子集输出。基于精英个体划分的变步长离散 FA(EDSBFA)的包装式特征选择方法流程如图 1。

3 实验与分析

3.1 实验设置

使用 python3 编程语言分别实现了本文提出的基于精英个体划分的变步长离散萤火虫算法 EDSBFA、固定步长离散萤火虫算法 BFA、文献[27]中所提变步长离散萤火虫算法 IBFA,EDSBFA 中除精英个体划分阈值 θ 和步长增大的随机加速度 a 之外,初始步长值 α_0、其他参数设置和 BFA 一样均同 IBFA[27]一致,如表 1。表 1 中,所有算法的最大迭代次数 Max$Gen=100$,t 为当前迭代计数,Randint 表示随机整数。

实验所用机器学习分类器与文献[14,16]一

致,为 k-近邻算法(k-nearest neighbor, KNN)分类器,其中,$K=1$,将 python 的 Sklearn 库中 KNN 应用接口与 EDSBFA,BFA,IBFA 的算法代码进行包装融合,分别开发了基于 EDSBFA,BFA,IBFA 的包装式特征选择程序 EDSBFA-KNN, BFA-KNN, IBFA-KNN,并使用文献[14,16]中测试所用到的 7 个 UCI 分类数据集对所开发的特征选择程序进行了实验测试,为了保证测试结果的客观准确,由萤火虫算法产生的特征子集使用 KNN 分类器的 10 折交叉验证方式进行评价,且特征选择程序在每个数据集上单独运行各 30 次,实验所得平均分类准确率、选择特征的平均数量与文献[14]中基于遗传算法(simple genetic algorithm,SGA)和文献[16]中基于离散粒子群算法(binary particle swarm optimization,BPSO)优化特征选择所得 KNN 分类结果以及与单一 KNN 分类结果进行了对比分析。此外,在 UCI 高维度数据集 LSVT(LSVT 为医疗语音数据集,共 126 个分类样本,每个样本多达 309 个特征)上进一步测试了 EDSBFA,BFA,IBFA 的性能,每个算法程序单独运行 20 次。实验环境为 Win10 系统 ,Inter (R) Core (TM) i3-3120 CPU 2.50 GHz, 8.0 GB 内存的 PC 机。

表 1　参数设置

Tab.1　Parameter setting

算法	种群规模	其他参数值
BFA	30	$\beta_0=1.0,\gamma=1.0,\alpha=0.5$
IBFA[27]	30	$\beta_0=1.0,\gamma=1.0,\alpha=0.5-0.5\cdot t/\text{MaxGen}$
EDSBFA	30	$\beta_0=1.0,\gamma=1.0$, $\alpha_0=0.5,\theta=0.899$,对前 7 个数据集 $a=\text{Randint}\in[10,25]$,对 LSVT 数据集 $a=\text{Randint}\in[1,10]$

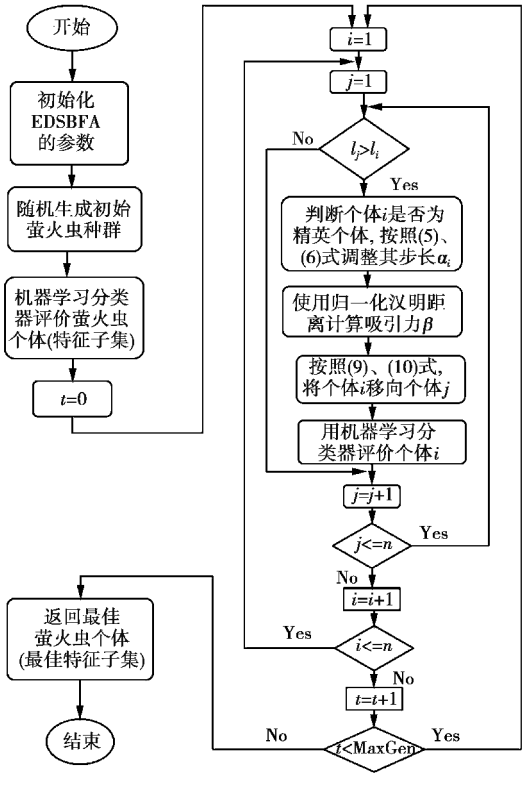

图 1　基于 EDSBFA 的包装式特征选择方法流程

Fig.1　EDSBFA-based wrapped feature selection process

3.2　结果分析

在 7 个 UCI 分类数据集上,包装式特征选择方法 EDSBFA-KNN, IBFA-KNN, BFA-KNN, SGA-KNN[14], BPSO-KNN[16] 和单一 KNN 取得的平均分类预测准确率与特征数量如表 2,其中,D 为数据集特征维度;N 为各方法选择的平均特征数量;A 为各方法获得的平均分类准确率,最少的特征数量与最高的分类准确率为加粗显示的数值。由表 2 可知,除 SGA-KNN 在数据集 Segmentation, WDBC 上取得的分类准确率比使用单一 KNN 的分类结果差之外,SGA-KNN 在其他 5 个数据集上的分类准确率比单一 KNN 要高,而 BPSO-KNN, BFA-KNN, IBFA-KNN, EDSBFA-KNN 在全部数据集上均取得了比单一 KNN 更高的分类准确率,说明了基于优化算法(SGA,BPSO,BFA,IBFA,EDSBFA)构造最优特征子集,KNN($K=1$)利用最优特征子集的特征数据进行分类预测可以显著改善其在数据集上的预测精度。Vowel 数据集的特征维度较小($D=10$),数据集较简单, SGA-KNN, BPSO-KNN, BFA-KNN, IBFA-KNN, EDSBFA-KNN 均取得了 99% 以上的分类精度,而 IBFA-KNN,EDSBFA-KNN 在 Wine 数据集上的分类准确率和选择的特征数量相同且比其他方法取得的结果要优,此外 EDSBFA-KNN 在 Vehicle,Segmentation,WDBC,Ionosphere,Sonar 数据集上均取得了最高的分类准确率,且其在大多数的数据集上选择的特征数量偏少。综合结果表明本文所提算法 EDSBFA 优化特征选择的能力要强于其他对比算法。

表2 各方法在数据集上的平均分类准确率与平均选择特征数量

Tab.2 Average values of classification accuracy and number of selected features through each method

数据集	SGA-KNN		BPSO-KNN		BFA-KNN		IBFA-KNN		EDSBFA-KNN		KNN	
	N	*A*/%	*N*	*A*/%	*N*	*A*/%	*N*	*A*/%	*N*	*A*/%	*N*	*A*/%
Vowel(*D*=10)	8	99.70	9	99.49	8	99.70	8	99.70	8	99.70	10	82.17
Wine(*D*=13)	8	95.51	8	98.88	8	98.36	8	99.44	8	99.44	13	94.40
Vehicle(*D*=18)	7	72.97	11	74.7	10	74.36	11	75.36	8	75.42	18	69.71
Segmentation(D=19)	11	92.95	11	97.88	9	97.66	10	97.92	9	98.17	19	93.89
WDBC(*D*=30)	12	93.95	13	97.72	11	97.52	14	98.33	13	98.42	30	94.89
Ionosphere(*D*=34)	7	94.70	10	93.73	15	93.30	11	94.91	9	95.06	34	87.22
Sonar(*D*=60)	24	95.49	32	92.79	28	89.62	20	87.61	26	95.99	60	69.75

表3 EDSBFA,IBFA,BFA 在 LSVT 上的测试结果

Tab.3 Test results of EDSBFA, IBFA, BFA for LSVT

数据集	BFA-KNN			IBFA-KNN			EDSBFA-KNN		
	N	*A*/%	*T*/s	*N*	*A*/%	*T*/s	*N*	*A*/%	*T*/s
LSVT (*D*=309)	151	86.64 (87.99)	1 016	138	95.51 (97.5)	1 934	105	95.97 (96.73)	804

在拥有高维度特征的 LSVT 数据集上进一步测试了 EDSBFA,IBFA,BFA 的性能,实验结果如表3,其中,*D*,*N*,*A* 的含义同表2,*T* 指算法程序运行时间,单位为 s。由于 LSVT 的候选特征空间相比前7个测试数据集大得多,搜索算法早熟收敛的风险相对较小,故 EDSBFA 的步长增大的幅度不宜过大,其随机加速度 a 应取偏小值,其取值为[1,10]的随机整数。由表3中 *T* 列数据知,EDSBFA-KNN 比 BFA-KNN、IBFA-KNN 要省分别近21%和58%的时间开销,由表3中 *A* 列(括号内为取得的最高分类准确率)数据知,BFA-KNN 的分类精度远不如 IBFA-KNN,EDSBFA-KNN 的分类精度,其最高分类准确率没有达到90%以上,而 IBFA-KNN,EDSBFA-KNN 的平均分类准确率都已达到了95%以上,且 IBFA-KNN 的最高分类准确率大于 EDSBFA-KNN 的最高分类准确率,但 EDSBFA-KNN 的平均分类准确率要高于 IBFA-KNN 的平均分类准确率,且 EDSBFA-KNN 选择的最优特征数量最少,以上表明在高维度数据集上基于 EDSBFA 优化特征选择的精度要比基于 BFA,IBFA 的更优、更稳定,且 EDSBFA 算法运行效率更高。图2~图7分别是 LSVT 测试中 BFA,IBFA,EDSBFA 的目标函数值(荧光亮度)、分类准确率、选择的特征数量在100次算法迭代搜索过程中的变化情况。如图2,BFA 在迭代中步长固定,无法自适应调节,导致搜索趋于随机化,整个种群难以持续获得更好的解,而使算法陷入局部最优或是无法收敛。图3中,IBFA 在迭代中整个种群虽能朝着优化方向进行搜索,可 IBFA 未区分精英与非精英萤火虫个体,这使萤火虫个体在解空间中无法根据自身状况进行差异化搜索,这使找到更好解的效率变低。从图4可知,EDSBFA 在迭代过程中最佳萤火虫个体的目标函数值(荧光亮度)和整个萤火虫种群的平均目标函数值都在持续平稳地获得提升,在迭代25次左右就取得比 BFA,IBFA 迭代100次时更高的目标值,且随着迭代优化的继续呈现出算法收敛趋势(最优值等于平均值时算法收敛),EDSBFA 的收敛速度快于 BFA,IBFA,且最终获得更优的解。

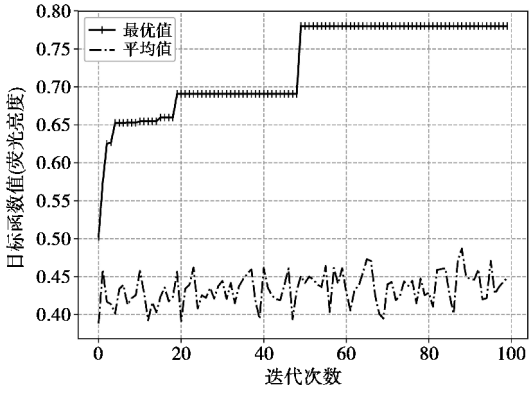

图2 BFA 的目标函数值在算法迭代中的变化

Fig.2 Change of the objective function value of BFA in 100 times iteration

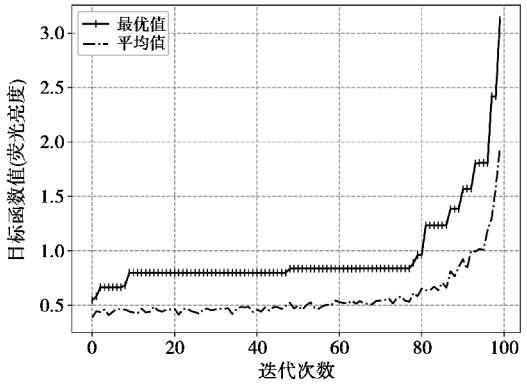

图 3　IBFA 的目标函数值在算法迭代中的变化

Fig.3　Change of the objective function value of IBFA in 100 times iteration

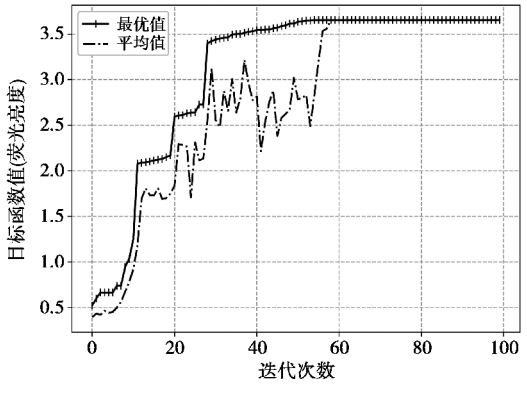

图 4　EDSBFA 的目标函数值在算法迭代中的变化

Fig.4　Change of the objective function value of EDSBFA in 100 times iteration

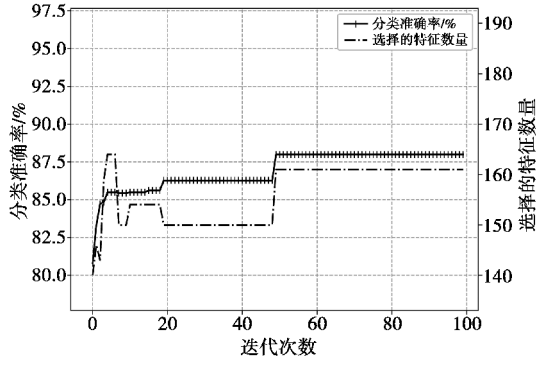

图 5　BFA 分类准确率、选择特征数量在迭代中的变化

Fig.5　Value of classification accuracy and number of selected features in BFA iteration

从图 5~图 7 可知,BFA 在迭代的后期由于算法陷入局部最优或是搜索偏向随机化导致难以取得更优的分类准确率和特征子集。IBFA 在迭代中可持续获得更高的分类准确率,但过程较缓慢,且选择的最优特征数量波动性大。EDSBFA 随着迭代能持续且快速地获得更高的分类准确率和含更少数量特征的特征子集,直至算法收敛,其总体性能明显优于 BFA,IBFA。

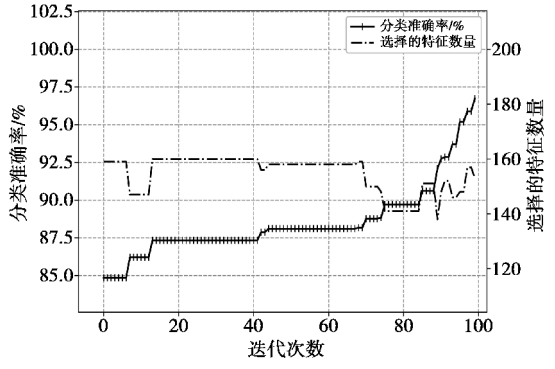

图 6　IBFA 分类准确率、选择特征数量在迭代中的变化

Fig.6　Value of classification accuracy and number of selected features in IBFA iteration

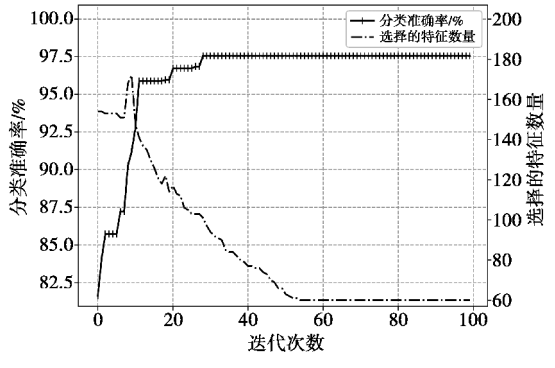

图 7　EDSBFA 分类准确率、选择特征数量在迭代中的变化

Fig.7　Value of classification accuracy and number of selected features in EDSBFA iteration

4　结束语

本文提出一种新的改进型萤火虫算法 EDSFA,并通过重新定义萤火虫距离和移动方式,给出了改进算法的离散化实现 EDSBFA,以适用解决特征选择问题,将 EDSBFA 与 KNN 分类器结合以包装式实现了特征选择优化,在 UCI 数据集上的实验表明,本文所提算法 EDSBFA 在优化特征选择效果和运行效率上性能优越,所提改进萤火虫算法使用基于精英个体划分的变步长策略,考虑了萤火虫个体差异性而进行自适应寻优,算法兼顾了解空间的全局搜索和局部探索,降低算法陷入局部最优的风险,同时使算法朝着种群优化的方向搜索,保证了算法的快速收敛。本文算法改进没有使用交叉、变异等复杂手段,但取得了理想效果。萤火虫算法的参数取值大小对算法的性能影响很大,不宜使用固定值,未

来在本文变步长改进策略的基础上,将对 FA 的光强吸收系数 γ 进行自适应改进。另外基于启发式算法的包装式特征选择优化方法运行效率低、耗时,未来可研究基于算法并行化的改进方法以提升效率。

参考文献:

[1] 管春.电能质量扰动分类中特征选择问题的研究[J].重庆邮电大学学报(自然科学版),2013,25(4):514-517.
GUAN C. Feature selection in power quality event classification[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2013, 25(4): 514-517.

[2] SU C T, LIN H C. Applying electromagnetism-like mechanism for feature selection [J]. Information Sciences, 2011, 181(5): 972-986.

[3] ESSEGHIR M A, GONCALVES G, SLIMANI Y. Adaptive particle swarm optimizer for feature selection[C]// International Conference on Intelligent Data Engineering and Automated Learning. Berlin, Heidelberg: Springer, 2010: 226-233.

[4] 乔立岩, 彭喜元, 彭宇.基于微粒群算法和支持向量机的特征子集选择方法[J].电子学报, 2006, 34(3): 496-498.
QIAO L Y, PENG X Y, PENG Y. BPSO-SVM wrapper for feature subset selection[J]. Acta Electronica Sinica, 2006, 34(3): 496-498.

[5] WANG C, SHAO M, HE Q, et al. Feature subset selection based on fuzzy neighborhood rough sets[J]. Knowledge-Based Systems, 2016(111): 173-179.

[6] UNLER A, MURAT A. A discrete particle swarm optimization method for feature selection in binary classification problems[J].European Journal of Operational Research, 2010, 206(3): 528-539.

[7] 叶东毅, 廖建坤.基于二进制粒子群优化的一个最小属性约简算法[J].模式识别与人工智能, 2007, 20(3):295-300.
YE D Y, LIAO J K. Minimum attribute reduction algorithm based on binary particle swarm optimization[J]. Pattern Recognition and Artificial Intelligence, 2007, 20(3):295-300.

[8] CAI J, LUO J, WANG S, et al. Feature selection in machine learning: A new perspective[J]. Neurocomputing, 2018(300): 70-79.

[9] 朱云龙, 申海, 陈瀚宁, 等.生物启发计算研究现状与发展趋势[J].信息与控制, 2016, 45(5): 600-614.
ZHU Y L, SHEN H, CHEN H N, et al. Research Status and Development Trends of the Bio-inspired Computation [J]. Information and Control, 2016, 45(5): 600-614.

[10] TABAKHI S, MORADI P. Relevance-redundancy feature selection based on ant colony optimization [J]. Pattern recognition, 2015, 48(9): 2798-2811.

[11] XUE Y, JIANG J, ZHAO B, et al. A self-adaptive artificial bee colony algorithm based on global best for global optimization[J]. Soft Computing, 2018, (22)9: 2935-2952.

[12] RODRIGUES D, PEREIRA L A M, NAKAMURA R Y M, et al. A wrapper approach for feature selection based on bat algorithm and optimum-path forest[J]. Expert Systems with Applications, 2014, 41(5): 2250-2258.

[13] DIAO R, SHEN Q. Nature inspired feature selection meta-heuristics[J].Artificial Intelligence Review, 2015, 44(3): 311-340.

[14] OH I S, LEE J S, MOON B R. Hybrid genetic algorithms for feature selection[J]. IEEE Transactions on pattern analysis and machine intelligence, 2004, 26(11): 1424-1437.

[15] WELIKALA R A, FRAZ M M, DEHMESHKI J, et al. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy[J].Computerized Medical Imaging and Graphics, 2015(43): 64-77.

[16] ZHANG Y, GONG D, HU Y, et al. Feature selection algorithm based on bare bones particle swarm optimization [J]. Neurocomputing, 2015(148): 150-157.

[17] 姚旭, 王晓丹, 张玉玺, 等.基于自适应 t 分布变异的粒子群特征选择方法[J].系统工程与电子技术, 2013, 35(6): 1335-1341.
YAO X, WANG X D, ZHANG Y X, et al. Feature selection algorithm using PSO with adaptive mutation-based t distribution [J]. Systems Engineering and Electronics, 2013,35(6):1335-1341.

[18] 郁书好.萤火虫优化算法研究及应用[D].合肥:合肥工业大学,2015.
YU S H. Research on firefly algorithm and its application [D].Hefei: Heifei University of Technology,2015.

[19] YANG X S.Nature-inspired metaheuristic algorithms[M]. Beckington, UK: Luniver press, 2010.

[20] 于宏涛, 高立群, 韩希昌.求解旅行商问题的离散人工萤火虫算法[J].华南理工大学学报(自然科学版), 2015, 43(1): 126-131.
YU H T, GAO L Q, HAN X C. Discrete Artificial Firefly Algorithm for Solving Traveling Salesman Problems[J]. Journal of South China University of Technology(Natural

Science Edition), 2015,43(1):126-131.

[21] 王艳, 王秋萍, 王晓峰.基于改进萤火虫算法求解旅行商问题[J].计算机系统应用, 2018, 27(8): 219-225.
WANG Y, WANG Q P, WANG X F. Solving Traveling Salesman Problem Based on Improved Firefly Algorithm [J].Computer Systems and Applications, 2018, 27(8): 219-225.

[22] 刘艺兰, 徐丽红, 吴丰彦,等.求解 0-1 背包问题的萤火虫算法[J].计算机与现代化, 2014(4):113-117.
LIU Y L, XU L H, WU F Y, et al. Firefly Algorithm for Solving 0-1 Knapsack Problem[J]. Computer & Modernization, 2014(4):113-117.

[23] 范阳涛, 汪民乐, 文苗苗, 等.基于萤火虫算法层次分析法的弹道导弹突防效能分析[J].系统工程与电子技术, 2010, 37(4): 845-850.
FAN Y T, WANG M L, WEN M M. Analysis of ballistic mis-sile penetration effectiveness based on FA-AHP[J]. Systems Engineering and Electronics, 2010, 37(4): 845-850.

[24] ZHANG L, MISTRY K, LIM C P, et al. Feature selection using firefly optimization for classification and regression models[J]. Decision Support Systems, 2018(106): 64-85.

[25] MISTRY K, ZHANG L, SEXTON G, et al. Facial expression recongition using firefly-based feature optimization[C]//2017 IEEE Congress on Evolutionary Computation (CEC). San Sebastian: IEEE, 2017: 1652-1658.

[26] ZHANG L, SHAN L, WANG J. Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion[J].Neural Computing and Applications, 2017, 28(9): 2795-2808.

[27] ZHANG J, GAO B, CHAI H, et al. Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm[J]. BMC bioinformatics, 2016, 17(1): 323.

[28] YANG X S. Firefly algorithm, Levy flights and global optimization[M]//Research and development in intelligent systems XXVI. London: Springer, 2010: 209-218.

作者简介:

刘　磊(1990—),男,山西忻州人,硕士,主要研究方向为智能调度与数据挖掘。E-mail:liuleifeichuan@foxmail.com。

罗　蓉(1979—),女,湖北天门人,讲师,硕士,主要研究方向为智能制造与现代集成制造系统。E-mail: luorong@cqupt.edu.cn。

尹　胜(1976—),男,四川资中人,副教授,博士,主要研究方向为制造系统工程与决策分析。E-mail:yinsheng@cqupt.edu.cn。

(**编辑**:刘　勇)