[文章编号] 1003－4684(2019)05-0064-03

基于多策略改进蝙蝠算法的文本特征选择

侯 乔，陈宏伟

（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 特征选择是文本分类过程的重要处理步骤，在其他分类预处理环节和分类算法确定的条件下，通过传统特征选择方法很难大幅度提高文本分类的准确率。针对此问题，介绍了一个基于改进蝙蝠优化的新的文本特征选择方法，即利用传统的特征选择方法对原始特征进行预选，在此基础上使用高斯局部扰动和自适应调节权重机制改进传统蝙蝠群算法，并以二进制编码形式对预选特征进行优选，分类准确率作为个体的适应度，提出了多策略改进蝙蝠算法的文本特征选择算法 MS-BA，实现对文本特征选择优化模型的高效求解。结果表明，采用 MS-BA 进行特征优选后，其分类准确率得到有效提高。

[关键词] 特征选择；蝙蝠算法；文本分类；多策略改进

[中图分类号] TP391 [文献标识码] A

随着 5G 时代的来临，物联网设备的大量普及，文本数据将进一步增长，而传统的文本预处理[1]会产生大量的原始特征词，包括冗余特征和不相关特征，增加了文本分类所需要的时间并且降低了分类的准确率。常规的这些基于统计和评估函数的方法[2-4]都有其相应的缺点，如何进一步找到最优的特征子集对文本分类有重要意义。

蝙蝠种群[5]在飞行寻优的过程中，所有的个体的位置都会依赖当前的最优个体。如果当前最优蝙蝠个体没有全面搜索，则蝙蝠种群很可能陷入局部收敛，形成早熟；或者算法中蝙蝠种群无论优劣都采用同一种随机飞行搜索方式迭代更新，忽略了个体之间的差异，造成个体的飞行能力下降，种群寻优的精度降低。本文充分利用算法的优势将高斯局部扰动和自适应调节权重迭代机制结合。首先，在常规蝙蝠算法的个体速度迭代方式上，增加自适应调节权重因子，动态调整个体的速度惯性，在迭代前期扩大搜索范围，增强个体全局飞行能力，之后在全局最优值处引入高斯扰动改进个体局部搜索方式，使蝙蝠个体在区域最优值附近有一定的变异扰动能力，避免错过全局最优，以上这些改进，不仅增强了蝙蝠种群在整个飞行空间的寻优能力，还提高了个体在某一区域最优值的扰动能力，避免种群陷入局部，保证了算法的全局收敛精度。

1 蝙蝠算法的文本特征选择

1.1 文本特征选择

文本分类，首先会标明数据集类别，按照一定比例划分数据集为训练集和测试集，之后对划分好的数据集进行文本预处理，根据相应的机器学习算法训练数据集，形成训练模型，然后输入与训练集相同格式的测试集文本，使用训练好的模型定义测试集所属的类别或标签，最后使用相应的评判标准来评价分类的准确率[6]（图 1）。

1.2 蝙蝠算法

蝙蝠算法基本上是针对连续优化问题而设计的。该算法的灵感来自蝙蝠通过回声捕捉猎物的行为。蝙蝠算法有三个主要优点：第一个是频率调谐，第二个是发射率。第三个是它们的音响强度，利用这三个优点，蝙蝠能够改变其飞行的速度和位置，跟随全局最优的蝙蝠在向量空间中寻找最优解。

1）对于已经分好词的训练集，通过 CHI 统计方法，按照 CHI 值降序取前 D 个词，找出种群目前的最优位置记为 X^*。

2）种群迭代。蝙蝠个体飞行按照公式（1）～（3）更新自己的速度和位置。

[收稿日期] 2019－07－05
[基金项目] 国家自然科学基金（61772180）；湖北省自然科学基金（2013CFB020）
[第一作者] 侯 乔（1994－），男，湖北安陆人，湖北工业大学硕士研究生，研究方向为大数据，文本处理
[通信作者] 陈宏伟（1975－），男，湖北武汉人，工学博士，湖北工业大学教授，研究方向为云计算，大数据

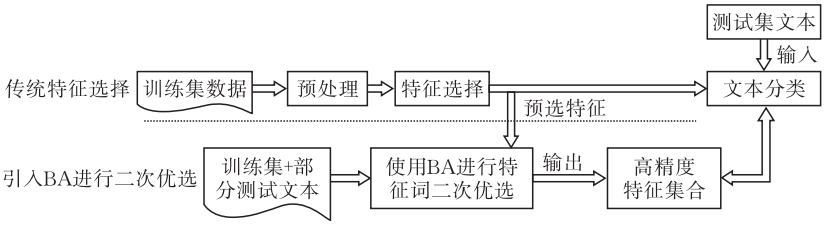


图 1　文本分类基本流程

$$f_i = f_{\min} + (f_{\max} - f_{\min}) \cdot \text{rand} \tag{1}$$

$$V_i^t = V_i^{t-1} + (X_i^{t-1} - X^*) \cdot f_i \tag{2}$$

$$X_i^t = X_i^{t-1} + V_i^t \tag{3}$$

3)如果 rand1$>r_i$,选择目前最优位置的蝙蝠,按照式(4),以随机飞行的方法产生区域新解。

$$X_{new} = X_{\text{old}} + \varepsilon \cdot A^t \tag{4}$$

4) 如果 Fit(X_{new})>Fit(X^*)且 rand2$<A_i$,则保留 X_{new}的位置,按照式(5)、(6)更新 r_i和 A_i。

$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma t)] \tag{5}$$

$$A_i^{t+1} = \alpha \cdot A_i^t \tag{6}$$

式中,t 为目前迭代数,A_i^t 和 ${A_i}^{t+1}$代表蝙蝠 i 在第 t 代和 $t+1$ 代的声波强度;α、γ 都为常数;r_i^0 为蝙蝠 i 初始发射率;随着迭代次数增加,蝙蝠个体逐渐靠近最优解,此时有,$t\to\infty$, $r_i^t\to r_i^0$, $A_i^t\to 0$。

5)对种群 Fitness 进行排序,寻找新一代的最优解 X^*。

6)重复以上步骤 2)—5)种群开始迭代,直到达到提前规定的最大迭代数,输出最优蝙蝠个体对应的位置。

2　多策略改进的蝙蝠优化算法

传统蝙蝠种群会以当代的最优位置为导向,如果该位置只是局部最优点,则蝙蝠群不会飞往全局最优,而逐渐停止飞行,陷入局部,因此本文主要运用以下两种方式来改进 BA 的全局搜索能力。

2.1　高斯扰动过程

传统的蝙蝠算法易陷入局部最优,所以对于(4)式,加入高斯区域扰动[7],改进个体局部搜索,帮助个体逃离局部最优,使蝙蝠个体在区域最优值个体周围进行搜索。且它产生的响度 A_i 大于随机数,则该高斯扰动的个体被加入其中,扰动的目的是使 X_i^t 得到进一步搜索[8],如下式:

$$X_i^t(\text{new}) = X_i^t(\text{old}) + \alpha \cdot N(0,1) + \varepsilon A^t$$

高斯分布是一种概率分布,记为 $N(\mu,\sigma)$,σ 是标准差,α 为扰动的系数,$\varepsilon \in [-1,1]$之间的随机数。

2.2　权重策略分析与改进

2.2.1　指数权重策略分析　在传统的 BA 中,权重 W 一直为 1,这就导致算法迭代前期和后期,变换速率一直不变,易前期形成局部最优,收敛速度变慢,寻优结果不佳。为此许多研究者做了大量工作,相关研究表明:指数递减策略在算法的优化上有较好的效果,指数递减惯性权重

$$w = w_{\min} \cdot (w_{\max}/w_{\min})^{1/(1+ct/t_{\max})}$$

取 $c=12$, $W_{\max}=0.9$, $W_{\min}=0.4$

2.2.2　自适应调节权重　对于指数线性递减策略 W_1来说,在迭代前期随着迭代次数增大,权重值会以很快的速率减小,在迭代中期,W_1下降速度十分缓慢[9]。因此,本文在指数线性递减策略上加以改进,提出了自适应性指数调节权重。在公式中加入负指数项,在搜索初期 iter 值较小,权重 ω 值较大,个体可以在全部的飞行空间更新速度和位置;而搜索后期 iter 值较大,权重 ω 值较小,个体在小范围内更新速度和位置,使 ω 依据当前自身的适应度来调节大小,提高了算法寻优性能

$$w = (w_{\max} - w_{\min})\exp\left(-\left(\tau \cdot \frac{\text{iter}}{\text{iter}_{\max}}\right)^2\right) + w_{\min}$$

τ 是控制参数,本文取值为[2,5]之间,指数递减策略权重 W_1和自适应调节策略权重 W_2随迭代次数的变化见图 2。

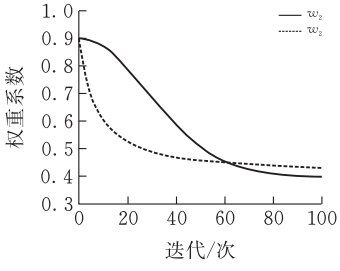


图 2　权重迭代对比

根据蝙蝠位置更新公式,将公式(2)改为下式:

$$V_i^t = wV_i^{t-1} + (X_i^{t-1} - X^*) \cdot f_i$$

基于多策略改进蝙蝠算法的文本特征选择算法 MS-BA 流程图见图 3。

3　实验与结果分析

3.1　实验环境及条件

本文实验使用 Python3.6 编写算法,使用的数据集为复旦大学语料库,共有 20 个类别。

分类算法是机器学习的重要组成部分。目前常用的分类算法包括朴素贝叶斯分类器[10]、KNN 分

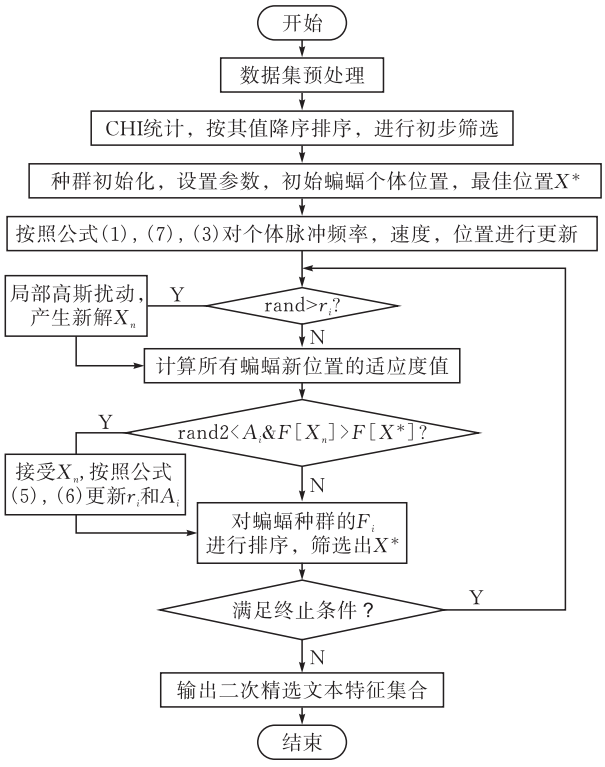


图 3 MS-BA 流程

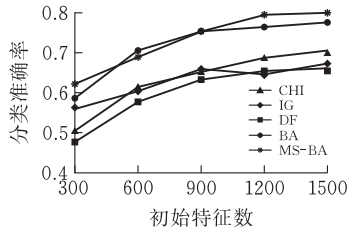
类器等，本实验运用上述两种算法来对文本进行分类。

在评价方面，使用文本检测准确率来测试分类性能，如下式：

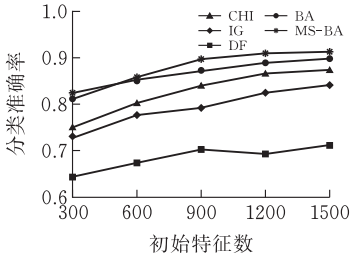
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \qquad (7)$$

3.2 实验结果

初始化 N 为 30，$\text{iter}_{\max}$ 为 100，使用上述两种分类器进行实验，其结果显示在表 1 中，并绘制趋势图(图 4)。



(a)使用 KNN 分类器



(b)使用 Bayes 分类器

图 4 分类准确率趋势

4 结论

本文研究了文本特征对于分类优劣的影响，介绍了一种改进蝙蝠优化的新的文本特征选择方法，使用高斯局部扰动和自适应调节权重机制改进传统蝙蝠群算法，实现了 BA 算法在文本特征选择中的应用。优化的文本功能集不受噪声影响，可以更好地对文本进行分类。实验表明，采用 MS-BA 文本特征选择方法进行特征优选后，其分类准确率得到有效提高。所提出算法优于普通蝙蝠算法和 CHI、IG、DF 传统方法。

[参 考 文 献]

[1] Pang Guansong, Jiang Shengyi. Text automatic classification technology research[J]. Information Studies: Theory & Application, 2012, 35(2):123-128.

[2] 周茜，赵明生，扈旻.中文文本分类中的特征选择研究[J].中文信息学报，2004，18(3)：18-24.

[3] 代六玲，黄河燕，陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报，2004，18(1)：27-33.

[4] 庞观松，蒋盛益.文本自动分类技术研究综述[J].情报理论与实践，2012，35(2)：123-128.

[5] Mirjalili S, Mirjalili S M, Yang X S. Binary bat algorithm[J]. Neural Computing and Applications, 2014, 25(3-4): 663-681.

[6] 李文慧，张英俊，潘理虎.多因素影响特征选择的短文本分类方法[J].计算机系统应用，2018，27(12)：216-221.

[7] 李煜，裴宇航，刘景森.融合均匀变异与高斯变异的蝙蝠优化算法[J].控制与决策，2017，32(10)：1775-1781.

[8] 朱德刚，孙辉，赵嘉，等.基于高斯扰动的粒子群优化算法[J].计算机应用，2014，34(3)：754-759.

[9] 吕石磊，黄永霖，陈海强，等.基于自适应步长的改进蝙蝠算法[J].控制与决策，2018，33(3)：57-564.

[10] Feng G, Guo J, Jing B Y, et al. Feature subset selection using naive Bayes for text classification[J]. Pattern Recognition Letters, 2015, 65: 109-115.

(下转第 71 页)