

[文章编号] 1003-4684(2021)01-0061-04

# 基于改进樽海鞘群算法的垃圾邮件分类

刘芳瑞, 陈宏伟

(湖北工业大学计算机学院, 湖北 武汉 430068)

**[摘要]** 为了规避电子邮件中的垃圾信息, 提出一种基于改进樽海鞘群算法的垃圾邮件分类。因樽海鞘群算法缺少惯性参数和找到全局搜索潜在解决方案的能力, 故利用 Tent 映射对初始种群施加混沌扰动, 并在位置更新中加入了惯性权重策略。采用增强的算法优化分类器的参数, 使得分类效果愈加显著。基于不同分类器和算法的实验表明, 优化后的算法明显提高了垃圾邮件的分类精确度和最佳识别准确度。

**[关键词]** 樽海鞘群算法; 混沌映射; 惯性策略; 垃圾邮件

**[中图分类号]** TP399 **[文献标识码]** A

电子邮件给网民提供了便利的方式来传递数据与别人交流, 但是, 垃圾邮件的泛滥<sup>[1]</sup>不仅影响用户体验, 也带来日益增多的安全风险。

一些邮件文本分类问题已被研究, 文献[2]提出支持向量机作为分类器开发了两种策略, 以最大可信度分类和统一可信度分类的未标记评论识别垃圾邮件; 文献[3]基于词向量间余弦相似度的朴素贝叶斯分类算法广泛应用于机器学习分类问题; 文献[4]提出基于统计特征, 通过极端梯度提升方法(XG-Boost)和广义提升回归模型(GBM)检测英语数据集中的垃圾意见。现在一些智能优化算法开始被应用于分类问题, 文献[5]提出了遗传算子和粒子群算法结合(H-FSPSOTC)来提高聚类算法性能, 用于大量文本文档的分类问题。

本文利用樽海鞘群算法优化支持向量机分类器, 深入研究了分类垃圾邮件的算法策略, 并采用 Wrapper 方法<sup>[6]</sup>对函数评估, 用评估指标对所提出的算法进行实验对比, 能够最大程度地将垃圾邮件正确过滤。

## 1 樽海鞘群优化算法

樽海鞘通常以群集的方式形成相当大的鱼群进行活动, 以此为灵感, Mirjalili 等人<sup>[7]</sup>提出了樽海鞘群算法(Salp Swarm Algorithm, SSA), 并将整个种群划分成领导者和跟随者。其中, 领导者位于前端并在多维搜索空间中引导一些群体(跟随者)搜寻最

佳解决方案(搜索目标)。在该算法中, 整个群体位于  $n \times d$  维搜索空间中, 其中  $n$  是问题变量的数量,  $d$  是空间维数。种群的位置用  $X_i$  表示成二维矩阵如下:

$$X_i = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & \cdots & x_d^n \end{bmatrix}$$

其中  $i = 1, 2, \dots, n$ 。算法的大致过程如下:

Step 1: 随机初始化群体公式如下:

$$X_{n \times d} = lb + \text{rand}(n, d)(ub - lb)$$

给定了搜索空间的搜寻范围是  $ub = [ub_1, ub_2, \dots, ub_d]$  和  $lb = [lb_1, lb_2, \dots, lb_d]$ , 来分别表示上下界。使其在搜索过程中不得超出边界, 否则将它拉回到规定的范围内。这个群体在搜索空间中的搜寻目标可定义为  $F = [F_1, F_2, \dots, F_d]$ 。

Step 2: 更新领导者的位置公式如下:

$$x_j^i = \begin{cases} F_j + c1((ub_j - lb_j)c2 + lb_j) & c3 \geq 0.5 \\ F_j - c1((ub_j - lb_j)c2 + lb_j) & c3 < 0.5 \end{cases} \quad (1)$$

其中  $j = 1, 2, \dots, d$ ,  $x_j^i$  和  $F_j$  分别是第  $j$  维度中的领导者和第  $i$  个食物来源的位置。 $c_1$  是一个非线性逐渐减小的过程, 并按如下公式计算:

$$c_1 = 2 \exp(-(4l/L))^2$$

其中  $l$  是当前迭代,  $L$  是最大迭代次数。

Step 3: 跟随者的位置更新公式如下:

$$x_j^i = \frac{1}{2}(x_j^i + x_j^{i-1})$$

[收稿日期] 2020-07-20

[基金项目] 国家自然科学基金项目(61772180)

[第一作者] 刘芳瑞(1996-), 女, 河南许昌人, 湖北工业大学硕士研究生, 研究方向为大数据

[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, 研究方向为大数据

式中  $i \geq 2$ , 表示第  $i$  个跟随者种群在第  $j$  维的位置。判断条件是否满足约束的阈值, 如果是, 则停止更新并输出优化结果。否则, 继续迭代。

## 2 改进的 SSA 算法应用研究

### 2.1 改进的樽海鞘群算法

正如刘建新等人提出的混沌策略<sup>[8]</sup>, 本文使用改进的 Tent 映射, 让初始种群广泛而又均匀地探索位置。初始化公式如下:

$$x_{n+1} = \begin{cases} 2x_n + \mu \sin(\theta x_n) & 0 \leq x_n < \frac{1}{2} \\ 2 - 2x_n - \mu \sin(\theta x_n) & \frac{1}{2} \leq x_n \leq 1 \end{cases}$$

其中取  $\mu = 1/32$ ,  $\theta = 4\pi$  时初始的种群个体分布均匀(图 1)。

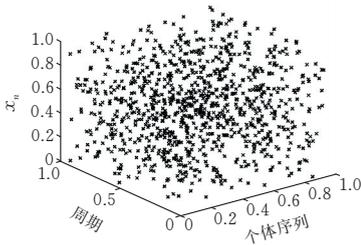


图 1 均匀分布的初始位置序列

为了提高收敛精度, 将权重因子  $\omega$  引入到跟随者的位置更新公式中<sup>[9]</sup>, 使群体在早期能够加快寻优能力, 在后期搜索相对准确的结果, 降低了陷入局部最优的风险。  $\omega$  的非线性递减的函数公式如下:

$$\omega(t) = \omega_{\min} + (\omega_{\max} - \omega_{\min}) \times \exp(-10t/T)$$

其中  $\omega(t)$  表示第  $i$  个个体在第  $t$  次迭代时的权重取值。经多次实验取  $\omega_{\min} = 0.5$ ,  $\omega_{\max} = 0.9$ ,  $T$  是最大迭代次数。那么, 新的跟随者更新公式如下:

$$x_j^i = \frac{1}{2} \times \omega(t) \times (x_j^i + x_j^{i-1}) \quad (2)$$

### 2.2 SVM 分类器

在线性 SVM 分类器中, 决策超平面能够正确的分离训练集中的数据点, 引入惩罚参数  $C$  控制了 SVM 的泛化能力, 防止过拟合。给定训练集  $\{(x_1, y_1), \dots, (x_m, y_m)\}$ ,  $m$  是训练集份数,  $x_i \in R$ ,  $y_i \in \{-1, 1\}$ , 确定权向量  $\omega$  和偏项  $b$ , 数据点用下式进行分类。

$$f(x_i) = \text{sign}(\omega^T x_i + b)$$

其中  $\xi_i$  表示松弛变量。在特征空间的特征具有非线性依赖性时, 将最小化函数公式(3)引入 SVM, 采用 SVM 和高斯核函数公式(4)集成。

$$\min \frac{1}{2} \omega^T \omega + C \sum_i \xi_i \quad \xi_i \geq 0 \quad (3)$$

$$k(x_i, x_j) = \exp(-\lambda \|x_i - x_j\|^2) \quad (\lambda > 0) \quad (4)$$

### 2.3 基于 SVM 改进算法的垃圾邮件分类

本文提出改进后的樽海鞘群算法(Improved salp swarm algorithm, ISSA)如图 1 所示, 算法过程描述如下:

Step1: 采用混沌映射初始化种群位置, 设置参数  $N=80$ , 最大迭代次数 100;

Step2: 对邮件的文本数据进行预处理, 将文本内容分词后, 创建词典作为邮件的原始特征  $P = [p_1, w_1, p_2, w_2, \dots, p_n, w_n]$ ,  $p_i$  表示特征词,  $w_i$  表示权重;

Step3: 采用空间向量表示, 用词频-逆向文件频率(TF-IDF)方法提取权重赋值更高的特征子集  $P = [p_1, p_2, \dots, p_d]$ , 划分数据集;

Step4: 计算个体的初始适应度值  $f(x)$ , 利用 ISSA 算法优化 SVM 高斯核函的惩罚参数  $C$  和  $\lambda$ , 实验得出最佳参数;

Step5: 更新个体和食物源的位置, 评估分类器模型, 获得全局最优分类。

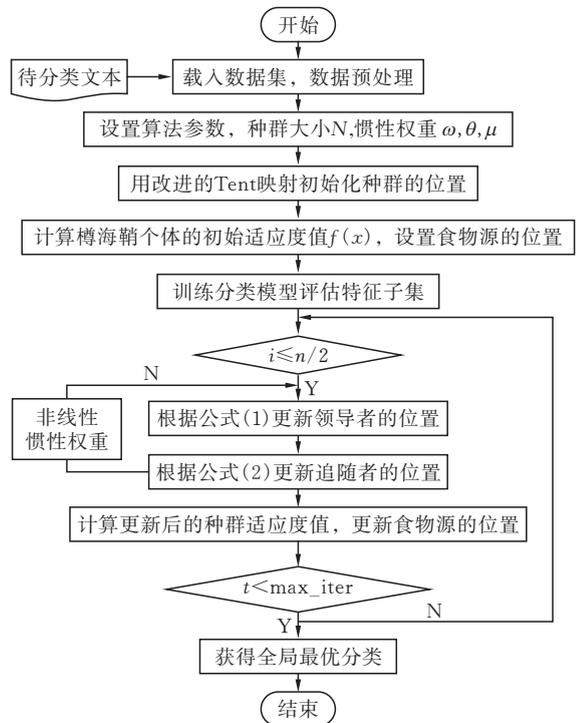


图 2 基于 ISSA 的垃圾邮件分类算法流程图

## 3 实验与结果分析

### 3.1 评价指标

精确率(ACC)是正确分类的邮件与总邮件数的比例, 公式如下:

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^K x(i, j)$$

其中  $x(i, j)$  表示第  $i$  个种群链在第  $j$  维的位置,  $N$  是所有个体的数量,  $K$  是种群的数量。

F-Measure(FM)是对完成分类的垃圾邮件占垃圾邮件的比例,以及对完成分类垃圾邮件占总邮件比例的整体评价,两者的值越高越有效。如

$$FM = \sum_j \frac{N_j}{N} \max_i \{x(i, j)\}$$

其中  $N_j$  表示樽海鞘群链的数量。

### 3.2 实验结果与分析

本文选取了 trec06c<sup>[11]</sup>的一个公开的中文邮件数据集,其中训练集有 45 360 份,测试有 19 440 份。实验选取特征数量 {500,1000,2000},惩罚因子 C 和  $\lambda$  取值范围是  $C = \{2^{-3}, 2^{-2}, \dots, 2^3\}$ ,  $\lambda = \{2^{-9}, 2^7, \dots, 2^3\}$ 。将 ISSA 与传统的 SSA 性能进行比较,分别采用 K 近邻(KNN)、逻辑回归(LR)和支持向量机(SVM)分别评估了三种不同的优化模型。

如图 3、4 所示,基于 SVM 的 ISSA 分类准确率和 F 值优于其他算法,在精确度上较其他算法也提高了 0.9%~6.2%。如表 1 所示,基于 KNN 的 ISSA 算法的执行时间最短,然而,数据集中的合法邮件和垃圾邮件不平衡比为 1.92,其他算法的稳定性不如 SVM-ISSA。当迭代次数达到 100 时,则特征数量为 2000,  $C = 2^5$ ,  $\gamma = 2$  时,SVM-ISSA 达到最佳效果,说明分类器对参数的取值较为敏感。

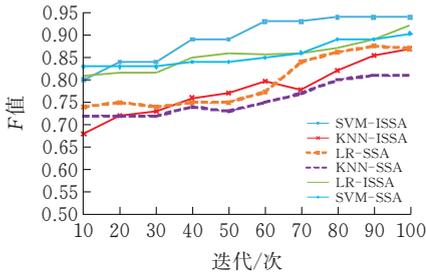


图 3 不同算法上的 F 值

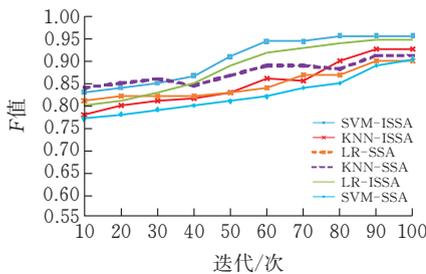


图 4 不同算法上的精确度

表 1 实验结果

测试指标	精确度	F 值	准确率	运行时间/s
SVM-SSA	0.904	0.903	0.952	404.4
SVM-ISSA	0.957	0.942	0.963	339.3
LR-SSA	0.901	0.878	0.903	420.2
LR-ISSA	0.948	0.908	0.922	342.9
KNN-SSA	0.913	0.812	0.926	360.6
KNN-ISSA	0.928	0.868	0.908	306.5

进制樽海鞘群算法(Binary SSA, BSSA)进行对比,ISSA 的收敛速度快于有明显波动的 BSSA,能够快速达到均衡的状态。实验证明,ISSA 在分类效果上更加稳健。

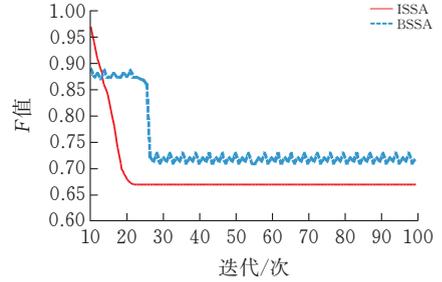


图 5 在 trec06c 上的适应度函数值

## 4 结论

将两种改进策略引入樽海鞘群算法,提高了算法的整体寻优性能,评估模型采取了三种机器学习算法用于训练分类器。其中 SVM 和 ISSA 算法模型表现出更好的性能,验证了基于樽海鞘群的 SVM 和 ISSA 分类算法在数据集上鲁棒性更好、更稳健。此外,该算法也可以用于解决大规模的工业问题。

### [ 参 考 文 献 ]

- [1] Thiago S. Guzella, Waldir M. Caminhas. A review of machine learning approaches to spam filtering[J], Expert Systems with Applications,2009,36 : 10206 - 10222.
- [2] Wen Zhang,Chaoqi Bu,CoSpa: A cotraining approach for spam review identification with support vector machine[J]. Information,2016,7(1):1-15.
- [3] 黄勇,罗文辉,张瑞舒.改进朴素贝叶斯算法在文本分类中的应用[J].科技创新与应用,2019,2019(5):24-27.
- [4] Hazim M, Anuar N B, Ab Razak M F, et al. Detecting opinion spams through supervised boosting approach [J].PLOS ONE 2018, 13(6):1-23.
- [5] Abualigah, Laith Mohammad,Khader, et al. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering [J]. Journal of Super computing, 2017,73(11):4773-4795.
- [6] 张俐,王枫.基于最大相关最小冗余联合互信息的多标签特征选择算法[J].通信学报,2018,039(005):111-122.
- [7] Seyedali.Mirjalili, Amir.H.Gandomi ,Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems[J].Advances in Engineering Software, 2017,114: 163-191.

图 5 绘制出平均适应度函数的变化情况,与二

- [8] 刘建新,李朝伟,张楷生.一种新的改进型 Tent 混沌映射及其性能分析[J].科学技术与工程,2013,13(8):2161-2165.
- [9] Harrison Kyle Robert, Engelbrecht Andries P, Beatrice. Inertia weight control strategies for particle swarm optimization : Too much momentum, not enough analysis[J]. Swarm Intelligence, 2016, 10(4): 267-305.
- [10] Freddy A. Lucay, Luis A. Cisternas. An LS-SVM classifier based methodology for avoiding unwanted responses in processes under uncertainties[J]. Computers and Chemical Engineering, 2019, 138: 1-33.
- [11] 张柳艳, 聂云峰, 段生月. 基于堆叠式降噪自编码器的中文垃圾邮件过滤[J]. 数学的实践与认识, 2020, 50(1): 105-114.

## An Improved Salp Swarm Algorithm for Spam Classification

LIU Fangrui, CHEN Hongwei

(School of Computer Science, Hubei Univ. of Tech., Wuhan 430068, China)

**Abstract:** In order to avoid spam in E-mail, an improved salp swarm algorithm for spam classification is proposed. However, the SSA lacks inertial parameters and ability to find global search potential solutions. The chaotic perturbation is applied to the initial population using Tent mapping, and the inertial weight strategy is added to the position update. Then, the parameters of the classifier are optimized by the enhanced algorithm, which makes the classification effect more significant. Finally, based on different classifiers and algorithm, experiments demonstrate that the optimized algorithm has an enormously increase on classification accuracy and the optimal recognition precision of spam.

**Keywords:** salp swarm algorithm; chaos mapping; inertial strategy; spam

[责任编辑: 张岩芳]

(上接第 25 页)

## Field Measurement of Slip Rate of Paddy Field Plant Protection Machine

CHEN Zheng, LIU Jiacheng, ZHOU Jingdong

(Agricultural Machinery Institute, Hubei Univ. of Tech., Wuhan 430068, China)

**Abstract:** In order to effectively monitor the walking condition of the plant protection machine in the complex paddy field environment and ensure that the slip rate of the plant protection machine is within the allowable range, this paper adopts a CM16-65P-1-24 Hall sensor and a low-speed radar to measure the paddy field. The wheel speed and the traveling speed of the plant protection machine were measured, and finally the slip rate was calculated. The results show that when the plant protection machine is in the low-speed gear, the average slip rate of the front wheels of the plant protection machine is 26.01%, and the average slip rate of the rear wheels is 23.33%. when the plant protection machine is in high-speed gear, the average slip rate of the front wheels of the plant protection machine is 19.39%, and the average slip rate of the rear wheels is 17.96%. It lays the foundation for the follow-up research on power distribution of plant protection machines, and has certain application value.

**Keywords:** plant protection machine; slip rate; power distribution

[责任编辑: 张 众]