改进NB算法在垃圾邮件过滤技术中的研究

刘月峰，苑江浩，张晓琳
（内蒙古科技大学 信息工程学院，内蒙古 包头 014010）

摘 要：朴素贝叶斯(NB)是一种简单高效的分类算法，且在垃圾邮件过滤中得到广泛应用，但是其属性间独立性的假设在一定程度上影响了分类效果.针对这一问题，提出一种改进的NB算法——FOA-NB算法.该算法将NB算法与果蝇优化算法(FOA)相结合，根据不同特征属性对分类的影响程度赋予不同的权值，通过FOA对权值进行优化，得到全局最优特征权向量，该算法在保留NB算法的简洁高效的优点的同时，通过权值优化获取更加具有决策性的特征属性，从而提高垃圾邮件过滤的正确率和召回率.通过仿真实验与NB算法、加权贝叶斯(WB)进行对比，结果表明FOA-NB算法使得垃圾邮件过滤效果得到明显改善，正确率和召回率均有所提高，且提高幅度约为5%.

关键词：垃圾邮件；朴素贝叶斯；特征权重优化；果蝇优化算法
中图分类号：TP393.098 **文献标识码**：A **文章编号**：1000－7180(2017)04－0115－06

Improved NB Algorithm Research in Spam Filtering Technology

LIU Yue-feng, YUAN Jiang-hao, ZHANG Xiao-lin
(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: Naive Bayes (NB) is a simple and efficient classification algorithm, and it is widely used in spam filtering, but because of the independence between the attributes of the hypothesis which has to some extent affected in classification effect. To solve this problem, the FOA-NB algorithm is proposed which is an improved NB algorithm. The algorithm of NB algorithm and Fruit fly optimization algorithm (FOA) combination, according to the different feature attributes of the influence degree of the classification given different weights, to optimize the weights by FOA and get the global optimal feature weight vector, the algorithm in the NB algorithm retains the advantage of simple and efficient at the same time, by optimization of the weights to obtain attributes which have more decision-making, so as to improve the spam filtering correct rate and recall rate. Through the simulation experiment with NB algorithm, Weighted Bayesian (WB), the results show that the FOA-NB algorithm makes the spam filtering effect has been improved significantly, and the correct rate and recall rate are improved, and the increase of about 5%.

Key words: spam email; naive bayes; optimization of feature weight; fruit fly optimization algorithm

1 引言

电子邮件为人们的生活、工作和学习等方面带来了便利，但电子邮件也是一把双刃剑，有部分邮件携带不良信息，甚至带有病毒，这样的邮件严重浪费了互联网资源和用户的时间.垃圾邮件过滤技术的研究有着深远的社会意义与巨大的经济价值[1].

贝叶斯方法在效率和适应垃圾邮件变化等方面优于K近邻方法、支持向量机(SVM)方法，依靠其自身的准确性和稳定性，使其在垃圾邮件过滤中得到了广泛的应用[2].朴素贝叶斯算法(Naive Bayes, NB)由于其存在条件独立性假设，即属性间是相互独立，互不影响，从而影响了垃圾邮件过滤的性能.为了解决这一问题，众多学者提出不同的方法对朴素贝叶斯改进，其中特征权重法为一个研究热点.

Zhang等[3]提出了加权朴素贝叶斯分类，其采用的方式是赋予每个属性不同的权值，以削弱条件

收稿日期：2016－06－29；**修回日期**：2016－08－11
基金项目：国家自然科学基金项目(51565046)

独立性假设，使分类更加符合实际. 实验证明，加权朴素贝叶斯分类的准确率优于朴素贝叶斯分类. Lee[4]提出了一种对特征值分配不同权重的方法，同时采用梯度下降的方法对权值进行最优化处理，实验证明其在分类领域得到了较好的效果. Jiang 等[5]提出了一种基于局部加权的朴素贝叶斯文本分类算法，通过加权的方法弱化了特征项间强独立性假设，得到了良好的分类效果. Wu 等[6]提出了一种基于人工免疫系统的方法对特征权重做自适应调整，以提高分类效果. 学者们采用不同的方法对特征权值进行处理优化，但目前在垃圾邮件过滤领域内采用群智能优化算法对权值进行优化的研究尚少，且群智能优化算法具有较好的寻优性能. 计宏[7]通过改进粒子群算法对概率权值进行优化，寻求全局最优权值从而提高分类效果，但是粒子群算法参数较多，不易调节.

因此，本文提出一种新的算法——FOA-NB 算法，该算法将朴素贝叶斯算法(NB)与果蝇优化算法(Fruit Fly Optimization Algorithm, FOA)算法相结合，FOA 算法具有较强的全局寻优能力，且参数较少，容易调节，因此该算法既保留了 NB 算法的简洁高效性，同时利用 FOA 算法对权值进行全局寻优，得到更加符合实际的特征权值，提高垃圾邮件的过滤效果.

2 相关技术

2.1 朴素贝叶斯的相关模型

贝叶斯方法是一种基于概率统计的学习方法，其可以根据事件的训练数据计算得到该事件发生的概率，以计算下一次该事件发生时的概率大小. 在垃圾邮件过滤的应用中，根据贝叶斯法则，通过对邮件集的训练数据得到的先验概率来估计其后验概率的大小.

贝叶斯公式提供了根据先验概率和条件概率计算后验概率的方法，其公式可写为

$$P(C|X)=\frac{P(X|C)P(C)}{P(X)} \tag{1}$$

为叙述方便，现对符号做如下规定：C 表示邮件类别，其值域为$\{C_{\text{spam}},C_{\text{ham}}\}$；假设邮件 e_i 有 m 个特征属性，将邮件 e_i 表示为特征向量，即为 $X_{e_i}=(X_1,X_2,\cdots,X_m)$，其中 $X_1,X_2,\cdots,X_m$ 是邮件 e_i 的特征属性. 由贝叶斯公式可知，邮件 e_i 属于类别 C_{spam} 的概率可表示为

$$P(C_{\text{spam}}|X_{e_i})=\frac{P(X_{e_i}|C_{\text{spam}})P(C_{\text{spam}})}{P(X_{e_i})} \tag{2}$$

由于在朴素贝叶斯模型中，存在条件独立性假设，即各属性间是相互独立的，因此

$$P(X_{e_i}|C_{\text{spam}})=P(X_1,X_2,\cdots,X_i,\cdots,X_m|C_{\text{spam}})=\prod_{i=1}^{m}P(X_i|C_{\text{spam}}) \tag{3}$$

在计算过程中，$P(X_{e_i})$ 可看作是常数，结合式(3)，则在朴素贝叶斯模型中，邮件 e_i 属于类别 C_{spam} 的概率式(2) 可转化为

$$P(C_{\text{spam}}|X_{e_i})=P(C_{\text{spam}})\prod_{i=1}^{m}P(X_i|C_{\text{spam}}) \tag{4}$$

由于朴素贝叶斯属性间独立性的假设在实际应用中很难达到满足，其不能发挥真正具有决策功能的特征属性的作用，同时增加了其他无关特征项的影响. 因此可根据不同特征项对分类的影响大小赋予不同的权值，将朴素贝叶斯扩展成为加权朴素贝叶斯. 对于加权朴素贝叶斯模型(Weighted Native Bayes, WNB) 来说，贝叶斯公式可表示为

$$P(C_j|X_{e_i})=\frac{P(C_j)\prod_{i=1}^{m}P^{\omega_i}(X_{e_i}|C_j)}{P(x_{e_i})} \tag{5}$$

式中，ω_i 是对应每一个特征值的权重. 在理想情况下，当其结果为 1 时，判定为垃圾邮件，当其结果为 0 时，则判定为正常邮件. 但在实际应用中，理想状况无法达到，但是当 $P(C_j|X_{e_i})$ 的值越接近于 0 或 1 时，分类结果越具有可信度，在实验过程中可以采用对邮件样本的训练使其接近理想情况.

加权朴素贝叶斯(WNB) 的大致过程[8] 如下.

输入：数据集

输出：分类结果

(1) 邮件预处理，对邮件进行向量化处理，经过分词与停用词处理，获得特征向量.

(2) 任务判断，若是训练任务，则执行步骤(3)，若为分类任务，则执行步骤(5).

(3) 训练所有样本，通过训练阶段，计算条件概率 $P(X_{e_i}|C_j)$ 以及类先验概率 $P(C_j)$，同时计算初始权值 ω_i.

(4)构造加权朴素贝叶斯分类器.

(5)调用训练所得的概率值以及权值，输出分类结果.

2.2 果蝇优化算法

特征权重法有效地创造了加权贝叶斯模型，对于权重的选取直接影响到分类的效果，为了提高分类的准确性，本文引入果蝇优化算法对权值进行全局寻优，获得最优权值.

根据果蝇优化算法的思想，可以把算法优化的过程分为以下步骤[9-11].

(1) 设定参数，即确定果蝇种群的规模大小 *sizepop*，果蝇算法的最大迭代次数 *maxgen*，初始化果蝇位置：随机产生果蝇的初始位置 (*X_axis*，*Y_axis*).

(2) 利用果蝇的嗅觉，赋予每个果蝇随机的方向和距离，如式(6)

$$\begin{cases} X_i = X\_axis + RandomValue \\ Y_i = Y\_axis + RandomValue \end{cases} \tag{6}$$

式中，*RandomValue* 为搜索距离，搜索方向为随机数.

(3) 计算果蝇当前位置距原点的位置 $Dist_i$ 以及味道浓度的判定值 S_i；

$$Dist_i = \sqrt{(X_i^2 + Y_i^2)} \tag{7}$$

$$S_i = 1/Dist_i \tag{8}$$

(4) 将计算所得味道浓度的判定值 S_i 代入味道浓度判断函数 *function*，计算出当前每个果蝇所在位置的味道浓度 $smell_i$；

$$smell_i = function(S_i) \tag{9}$$

式中，味道浓度判断函数 *function* 需要根据实际情况进行确定.

(5) 比较每个果蝇所在位置的味道浓度 $smell_i$，找到当前果蝇种群中味道浓度最大的果蝇个体.

(6) 确定味道浓度最大的果蝇个体所在位置 (*X*，*Y*) 以及当前的味道浓度值，此时的位置记为最优个体位置，味道浓度记为最佳浓度值，利用果蝇的视觉，所有果蝇向最优个体位置飞去，更新种群的位置.

(7) 迭代寻优，迭代次数若小于最大迭代次数 *maxgen*，重复(2) ～ (6)，再判断当前最佳浓度值是否优于前一最佳浓度值，若优于，执行步骤(6)，否则重复执行(2) ～ (6)，直至迭代次数达到最大迭代次数 *maxgen*，找到目标所在位置.

在文献[12] 中，吴小文等人将 FOA 算法与其他群智能优化算法进行比较，得到结论 FOA 算法简单，参数相对较少易调节，全局寻优能力强并且寻优精度较高.

3　FOA-NB 算法

FOA-NB 算法融合 NB 算法和 FOA 算法，一方面保留了 NB 算法的高效简洁性，即特征属性间相互独立，使得计算复杂度降低；另一方面根据不同特征属性对分类的影响大小赋予不同的权值，利用 FOA 算法进行全局寻优，提高分类准确性. 该算法步骤如下.

输入：数据集

输出：分类结果

(1)邮件预处理，对邮件进行向量化处理，经过分词与停用词处理，获得初步特征向量.

(2)特征提取，采用信息增益的方法对特征项提取，获得具有决策性的特征属性.

信息增益定义如下：

$$\begin{aligned} IG(t) = &-\sum_{j=1}^{|c|} P(c_j)\log P(c_j) + \\ &P(t)\sum_{j=1}^{|c|} P(c_j|t)\log P(c_j|t) + \\ &P(\bar{t})\sum_{j=1}^{|c|} P(c_j|\bar{t})\log P(c_j|\bar{t}) \end{aligned} \tag{10}$$

(3) 任务判断，若是训练任务，则执行步骤(4)，若为分类任务，则执行步骤(7)；

(4) 训练所有样本，通过训练阶段，计算条件概率 $P(X_{e_i} \mid C_j)$ 以及类先验概率 $P(C_j)$，同时计算初始权值 ω_i；

(5) 确定目标函数，即味道浓度判定函数，利用 FOA 算法优化目标函数，获取全局最优权值.

当在理想情况下，则如式(11) 所示.

$$\varepsilon = \begin{cases} 1, x_j \in c_s \\ 0, x_j \in c_h \end{cases} \tag{11}$$

令式(2) 为 ε_s，本文通过最小二乘法确定关于权值 ω 的目标函数如下：

$$f(\omega) = \min\sum_{j=1}^{n} (\varepsilon - \varepsilon_s)^2 \tag{12}$$

(6)构造加权朴素贝叶斯分类器；

(7) 调用所得的概率值以及权值，输出分类结果.

算法流程图如图 1 所示.

4　实验及结果分析

为验证算法的性能，对算法进行了实验测评. 采用交叉验证方法，把邮件样本集随机地分为 5 个大小相等但互不相交的子集，对邮件样本学习和测试分别进行 5 次，计算出每次分类结果的正确率和召回率，为了使结果更具科学性，避免实验的随机性和偶然性，采用 5 次结果的平均值作为最终衡量标准.

4.1　实验环境与测试数据

本文选取的实验环境如下所示.

硬件配置：CPU 为 I5-3470 双核 3.20 GHz，内存 8.00 GB，硬盘 1 TB.

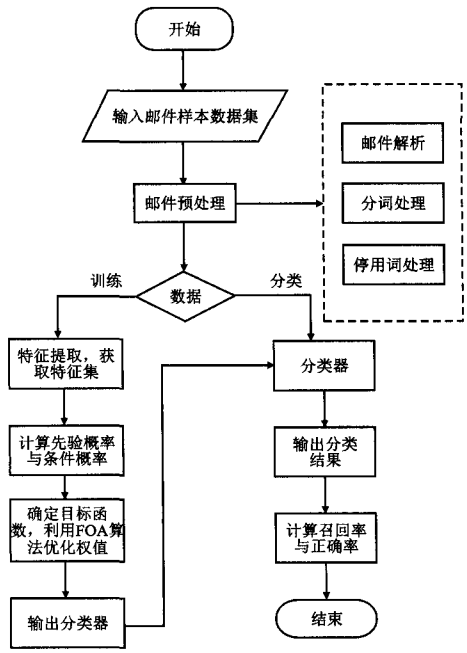


图 1 FOA-NB 算法流程图

操作系统：Windows 7.0；

测试平台：MyEclipse10.0，Matlab R2011a；

实验数据：中国教育和科研计算机网紧急响应组(Data Sets of Chinese Emails，CCERT2005-jun)，其包含合法邮件 9 272 封，垃圾邮件 25 088 封. 从中选取正常邮件 1 000 封，垃圾邮件 1 500 封作为邮件样本集.

4.2 实验评价指标

为了有效评价算法的分类效果，实验中采用以下三个评价指标.

(1)R(召回率)：垃圾邮件被检测出来的概率，当其值越高时，我们认为分类效果越好.

$$R=\frac{n_{\text{spam}\to\text{spam}}}{N_{\text{spam}}}\times 100\% \tag{13}$$

式中，$n_{\text{spam}\to\text{spam}}$ 为判断正确的垃圾邮件数目，N_{spam} 为样本集中垃圾邮件的总数目.

(2)P(正确率)：垃圾邮件被判断正确的概率大小. 当正确率越大，则说明误判率越低，分类效果越好.

$$P=\frac{n_{\text{spam}\to\text{spam}}}{n_{\text{spam}\to\text{spam}}+n_{\text{ham}\to\text{spam}}}\times 100\% \tag{14}$$

式中，$n_{\text{ham}\to\text{spam}}$ 为将正常邮件判定为垃圾邮件的数目.

(3)F 值：一个综合考虑指标，其综合考虑了召回率和正确率两个因素.

$$F=\frac{2\times R\times P}{R+P} \tag{15}$$

4.3 测试与分析

4.3.1 FOA 算法与 PSO 算法的对比

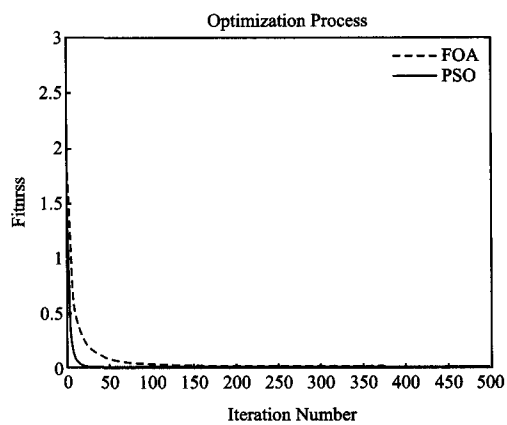
本文首先对 FOA 算法与粒子群优化算法(PSO)进行实验对比，采用三个测试函数进行分析，分别为 Ackley 函数、Rastrigin 函数以及 Girewank 函数，如表 1 所示.

表 1 测试函数

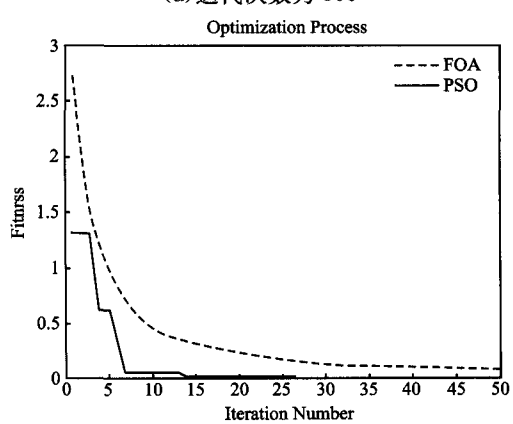
函数名	维数	公式
Ackley	30	$f(x)=-20\exp\left(-0.2\sqrt{\frac{1}{30}\sum_{i=1}^{n}x_i^2}\right)-\exp\left(\frac{1}{30}\sum_{i=1}^{n}\cos(2\pi x_i)\right)+20+e$
Rastrigin	30	$f(x)=\sum_{i=1}^{n}(x_i^2-10\cos(2\pi x_i)+10)$
Girewank	30	$f(x)=\frac{1}{4\,000}\sum_{i=1}^{n}(x_i)^2-\prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$

算法的参数设置如下所示.

种群大小 100，迭代次数分别为 500 和 50. 在 Matlab 中进行实现，本文以 Ackley 函数为例进行分析. 两种算法迭代寻优过程如图 2 所示.



(a)迭代次数为 500



(b)迭代次数为 50

图 2 Ackley 函数迭代寻优过程

由图 2(a)可看出，将迭代次数设置为 500 次，FOA 算法与 PSO 算法均能寻得最优值，且寻得最优值所迭代的次数基本相同，区别在于迭代过程中，PSO 算法极易陷入局部极值，为更方便看出这一点，将迭代次数设为 50 次，如图 2(b)所示，可以看出 PSO 算法在迭代过程中极易陷入局部极值，这将会影响其寻优效果. 除此之外，在实验过程中，记录两种算法的时间，其中 PSO 算法的时间约为 0.097 s，而 FOA 算法的时间约为 0.077 s，可证明 FOA 算法的效率高于 PSO 算法. 因此本文选择 FOA 算法进行权值寻优，而非 PSO 算法.

4.3.2　FOA-NB 算法验证

为了验证提出的 FOA-NB 算法的效果，仿真实验分别测试 NB、WB、FOA-NB 这 3 种不同的算法，采用交叉验证的方法进行 5 次，取 5 次结果的平均值作为结果，获得 3 种算法下 R(召回率)、P(正确率)、F 值的值，从而分析其垃圾邮件辨别能力. 结果如表 2 所示.

由表 2 可以看出，在采用 WB 算法时，准确率以及召回率要优于 NB 算法；当采用 FOA-NB 算法进行权值寻优操作后，召回率和正确率有了进一步的提升，相对于 NB 算法，其提升幅度约为 5%. 为便于直观的观察结果，本文对 5 次交叉实验的平均值进行了描述，如图 3 所示.

在实验过程中，针对误判的邮件(即正常邮件被判断为垃圾邮件或垃圾邮件被判断为正常邮件的情况)进行了统计，发现虽然 FOA-NB 依旧误判该邮件，但所得到的概率发生了变化. 本文将以正常邮件 606，垃圾邮件 24085 为例说明. 如表 3 所示(注：由于在计算过程中条件概率连乘造成数值极小，表中所列数据为乘以扩大因子 Zoomfactor 的值，取值为 16.5).

表 2　仿真实验结果对比

实验次数	算法	R/%	P/%	F/%
1	NB	86.67	88.68	87.66
	WB	88.30	89.89	89.09
	FOA-NB	91.41	92.85	92.12
2	NB	82.27	87.02	84.58
	WB	84.33	87.74	86.00
	FOA-NB	88.02	90.03	89.01
3	NB	85.20	87.90	86.53
	WB	86.56	88.64	87.59
	FOA-NB	89.30	90.92	90.10
4	NB	87.87	86.65	87.26
	WB	89.34	87.39	88.35
	FOA-NB	92.10	90.11	91.09
5	NB	83.89	84.70	84.29
	WB	85.43	85.67	85.55
	FOA-NB	89.66	89.37	89.51
Average	NB	85.18	86.99	86.06
	WB	86.79	87.87	87.32
	FOA-NB	**90.10**	**90.66**	**90.37**

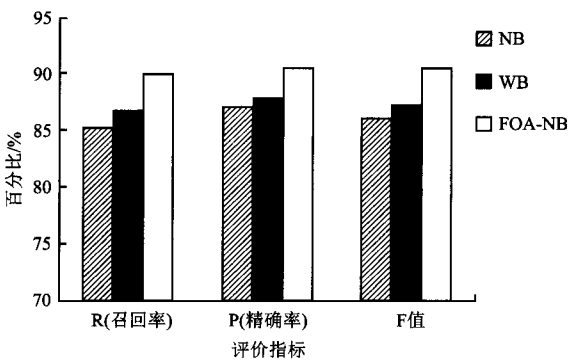


图 3　三种算法实验结果对比图

表 3　误判邮件概率对比

测试邮件	算法	判为正常邮件的概率(P_h)	判为垃圾邮件的概率(P_s)	P_s/P_h
606	NB	6.474918016341828E-296	4.991157495454274E-216	2.09539592E+80
	FOA-NB	4.863376662485665E-71	6.562824968883778E-28	1.34945633E+41
24085	NB	1.0445114055873754E-214	8.576083738391078E-302	8.21062179E-88
	FOA-NB	4.600898972386032E-76	1.352959201071769E-141	2.17285613E-67

由表 3 可以看出，当为正常邮件时，虽然两种算法最终均误判为垃圾邮件，但是 FOA-NB 算法相对于 NB 算法，比值(P_s/P_h)却减小了，说明其误判为垃圾邮件的概率降低了；当为垃圾邮件时，虽然两种算法最终均误判为正常邮件，但是 FOA-NB 算法相对于 NB 算法，比值(P_s/P_h)却增大了，说明其误判为正常邮件的概率降低了. 这说明 FOA-NB 算法中的权值寻优过程在垃圾邮件过滤中起到了作用，即本文所提出的 FOA-NB 算法在垃圾邮件过滤应用中是有效的.

5　结束语

本文基于贝叶斯相关理论研究了垃圾邮件过滤技术，提出了一种 FOA-NB 算法，该算法在保留朴素贝叶斯的简洁高效的优点的同时，通过加权的方法获取更加具有决策性的特征属性，利用最小二乘

算法确定权值的目标函数，利用 FOA 算法优化权值，以提高垃圾邮件过滤的效果. 仿真实验结果表明，FOA-NB 算法在召回率和正确率上都有提高，提高了垃圾邮件的过滤性能.

参考文献：

[1] 中国互联网协会. 2014 年第一季度反垃圾邮件状况调查报告[EB/OL]. [2016-05-06]. http://www.12321.cn/pdf/2014chinadbcg.pdf, 2014.

[2] 翟军昌，秦玉平，王春立. 改进的朴素贝叶斯垃圾邮件过滤算法[J]. 计算机工程与应用，2009，45(14)：145-148.

[3] Zhang H, Sheng S. Learning weighted native bayes with accurate ranking[C] //Proceedings of the Fourth IEEE International Conference on Data Ming. IEEE Computer Society, 2004: 567-570.

[4] Chang-Hwan Lee. A gradient approach for value weighted classification learning in native Bayes[J]. Knowledge-Based Systems, 2015, 85(3): 71-79.

[5] Jiang L, Cai Z, Zhang H, et al. Naive Bayes text classifiers: a locally weighted learning approach[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25(2): 273-286.

[6] Jia Wu, Pan Shirui, Zhu Xingquan, et al. Self-adaptive attribute weighting for Native Bayes classification[J]. Expert Systems with Application. 2015, 42(5): 1487-1502.

[7] 计宏. 改进贝叶斯垃圾邮件过滤技术的研究[J]. 计算机测量与控制，2013，21(8)：2181-2184.

[8] 邓维斌，王国胤，王燕. 基于 Rough Set 的加权朴素贝叶斯分类算法[J]. 计算机科学，2007，34(2)：204-219.

[9] Pan W T. A new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example[J]. Knowledge-Based Systems, 2012, 26(2): 69-74.

[10] 潘文超. 果蝇最佳化演算法[M]. 中国台北：沧海书局，2011：10-12.

[11] 韩俊英，刘成忠. 自适应混沌果蝇优化算法[J]. 计算机应用，2013，33(5)：1313-1316.

[12] 吴小文，李擎. 果蝇算法和 5 种群智能算法的寻优性能研究[J]. 火力与指挥控制，2013，38(4)：17-25.

作者简介：

刘月峰 男，(1977-)，博士研究生，副教授. 研究方向为机器学习、文本分类.

苑江浩(通讯作者) 男，(1992-)，硕士研究生. 研究方向为机器学习、数据挖掘. E-mail：qiji_2014@163.com.

张晓琳 女，(1966-)，博士，教授. 研究方向为数据库、大数据隐私保护.