doi:10.3969/j.issn.1673-9833.2020.01.012

# 基于随机森林的异常邮件检测方法研究与实现

彭 成[1, 2]，展万里[1]，周晓红[1]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；2. 中南大学 自动化学院，湖南 长沙 410083）

**摘 要**：目前现有技术在中文异常邮件过滤方面，存在误判、效率不高等缺陷。为了缓解此问题，结合随机森林算法的优点，采用了中文分词方法进行特征提取，并对词频进行权重计算，通过奇异值降解，更好地填充算法以完成对中文异常邮件的检测。多种算法的对比分析检测效果表明，提出的基于随机森林异常邮件检测器在精准度、召回率的性能均优于其他算法，而在时间效能上也处于较好水平。

**关键词**：异常邮件；随机森林；特征提取；奇异值降解；邮件过滤

**中图分类号**：TP391 **文献标志码**：A **文章编号**：1673-9833(2020)01-0070-07

**引文格式**：彭 成，展万里，周晓红 . 基于随机森林的异常邮件检测方法研究与实现 [J]. 湖南工业大学学报，2020，34(1)：70-76.

# Research and Implementation of Abnormal Mail Detection Method Based on Random Forest Algorithm

PENG Cheng[1, 2]，ZHAN Wanli[1]，ZHOU Xiaohong[1]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. School of Automation，Central South University，Changsha 410083，China）

**Abstract**：Currently, such deflects as misjudgment and inefficiency can be found in the technology of filtering Chinese abnormal mails. In order to efficiently solve this problem, this paper combines the advantages of random forest algorithm, adopts Chinese word segmentation method to extract features, and calculates the weight of word frequency. Based on a singular value degradation, this new approach performs better in filling in the algorithm to complete the detection of Chinese abnormal mail. Compared with the detection results of various algorithms, the experimental results show that the performance of the proposed random forest anomaly email detector is superior to other algorithms in accuracy, recall rate and time efficiency.

**Keywords**：abnormal mail；random forest algorithm；feature extraction；singular value degradation；mail filtering

## 0 引言

近年来，随着网络通信技术飞速发展，电子邮件成为人们日常生活和工作的主要交流方式之一，但异常邮件问题也随之而来。异常邮件占用了大量的网络资源，对互联网中的用户造成了巨大影响和威胁，

**收稿日期**：2019-05-29

**基金项目**：国家自然科学基金资助面上项目（61871432），国家自然科学基金资助青年项目（661702178），湖南省自然科学基金资助青年项目（20173065），湖南省教育厅高等教育教学改革研究基金资助项目（2017[452]-289）。

**作者简介**：彭 成（1982-），男，湖南长沙人，湖南工业大学副教授，博士，硕士生导师，主要研究方向为工业大数据分析，E-mail：10822060@qq.com

**通信作者**：周晓红（1982-），女，湖南衡阳人，湖南工业大学教师，主要研究方向为工业大数据分析，E-mail：10822060@qq.com.

甚至导致用户损失数据和金钱。异常邮件破坏性强、传播速度快、危害范围广，如何有效阻断异常邮件的传播，提高对异常邮件的判别能力是当前研究的迫切要求。为了保护用户的权益、减少网络带宽和资源的消耗，异常邮件的鉴别与过滤技术也逐渐受到研究者的重视。本文结合随机森林算法的优点，突破邮件特征提取、分类及异常邮件检测等关键技术难点，并与典型的算法进行实验对比分析，实验结果表明该方法在准确率等方面具有明显优势。

## 1　国内外研究现状与分析

异常邮件概念自 1978 年提出以来，全世界的专家学者对异常邮件检测技术进行研究与实践，至今为止已取得了丰硕的研究成果。

邮件分类检测方法大体可以分为两类：基于 IP 地址的邮件检测技术和基于内容的邮件检测技术 [1]。在基于 IP 地址的邮件检测技术中主要包括黑白名单检测技术 [2]、实时黑名单检测技术以及主机名反向验证技术 [3] 等。实际应用中，黑名单检测技术和白名单检测技术通常结合起来应用于服务器。而基于内容的异常邮件检测技术是目前主流异常邮件检测过滤技术。为了提高过滤效果，反异常邮件产品往往结合使用多种过滤技术 [4-5]。

邮件的分类其实质是对文本信息进行处理，现有的 K- 近邻、贝叶斯、神经网络、支持向量机、决策树等经典机器学习算法 [6-9] 被广泛应用到专利文本分类领域。于是，研究者试图将对文本的处理方法引入邮件分类处理中，通过文本聚类或分类方法将邮件分为异常和正常两类。但是与普通文本相比，邮件具有不一样的特点，它是一种非结构化的文本，采用一般的文本分类算法和传统的机器学习方法不能很好地区分正常和异常邮件，错误率较高 [10]。

为了缓解此问题，在研究大量参考文献的基础上，课题组发现随机森林（random forest，RF）算法是机器学习中一个可尝试的精确分类算法 [11-12]，该算法由 Leo Breiman 等在 21 世纪初提出 [13]。它是一种利用多棵决策树对样本进行训练并预测的分类器，与其他算法相比具有以下几个方面的优点：1）具有通用性，适合多种环境，可用于聚类分析，引导无监督聚类、异常检测和数据透视等；2）不需要剪枝，相比单一决策树算法不易产生过拟合；3）对异常值、噪声数据不敏感，能保持良好的精确度；4）能提取高维数据的主要特征，可用于数据降维。本文在异常邮件中的过滤技术基础上，结合随机森林算法，设计并实现了异常邮件检测方法。实验结果表明，该算法获得了较高判别率。

## 2　整体思路

本研究采用的方法在机器学习领域被称作有监督学习 [14-15]（supervised learning）方法，因此实现的流程也按照有监督学习的基本步骤完成。有监督学习是指用已知某种特性的样本作为训练集，以建立一个数学模型再用已建立的模型来预测未知样本，其流程如图 1 所示。

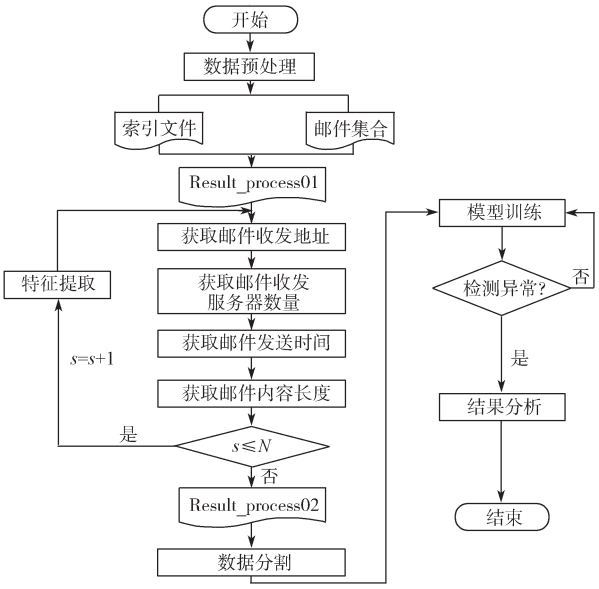


**图 1　整体流程图**

Fig. 1　Overall flow chart

如图 1 所示，本文具体思路如下。

1）数据清洗。数据收集完毕后，由于数据集中可能存在无关或冗余信息，将影响邮件分类的精确度，有必要进行数据清洗。具体步骤：根据 indexFolder/indexFile 索引文件对邮件数据集合（dataSet/data/…）处理，得到处理后的数据文件 processxx_xxx 到 dataSet/processSet 文件夹下，以及 Result_process01 到 dataSet/firstResult 文件夹下。

2）特征提取 [16]。分别对获取的邮件地址、邮件服务器数量、邮件发送时间及邮件内容进行特征提取，具体是：对 Result_process01 文件中数据进行特征提取，生成数据文件 Result_process02 到 dataSet/secondResult 文件夹下，本文为了判断邮件长度对异常邮件信息量的影响，得到了在不同邮件长度下异常邮件占比，以及在不同邮件长度大小下邮件信息量的大小。

3）数据分割。将收集的邮件集按照比例分为测试集和训练集，并输出到对应的文件夹，具体是：对 Result_process02 文件进行分割（train_test_split）得到 *x_train*、*x_test*、*y_test* 3 个集合，分别对应输出到

testSet 文件夹下和 trainSet 文件夹下。

4）模型训练 [17]。首先对训练集进行词频权重计算（term frequency-inverse document frequency，tf−idf）并做奇异值降解（singular value decomposition，SVD），构建对应的数据矩阵用来填充。

5）结果分析。经过训练后，对分割好的测试集进行预测得到结果并进行对比，输出结果图以及结果表到 result 文件夹下。

算法的详细流程如下。

### 2.1 邮件数据集合处理

本研究收集了近 10 000 封邮件，其中有异常邮件和非异常邮件，已通过索引文件对各个邮件分类，并且按照（spam ../data/000/000 或者 ham ../data/000/001，前者标记为 data/000/000 是异常邮件，后者标记为 data/000/001 是非异常邮件）格式存放，之后的数据处理利用索引文件中存放的信息定位到各个邮件，并获取各个邮件数据。对于单一的文本信息类型邮件，每一封邮件都有着固定的格式（From 为发送方，To 为接收方，Date 为日期，Content 为具体内容）。为了方便后续特征提取，此处按照邮件固定格式将所有邮件合并，每一封邮件内所有信息按照固定格式排成一行（将一封邮件按照 From、To、Date、Content 的格式放在一行上），制作成二维表的形式合并到一个文本文件中。即从 10 000 封邮件文本中，将各个邮件文本按格式提取，之后压缩到同一个文本文件中方便处理。

### 2.2 特征提取

异常邮件的建模与过滤过程中，无法直接对异常邮件进行过滤操作，首先需要对异常邮件进行分析，找出一些关键元素，如词、字或短词等，从而提取邮件特征 [18]。为了提高过滤效果，使用正则表达式对分词后的邮件进行二次处理 [19]。对邮件数据集合处理完毕后，得到一个由二维表 [20] 填充的文本文档。具体方法如下：

1）对邮件地址的提取。采用正则表达式 re.findall(r"@([A-Za-z0-9]*\.[A-Za-z0-9\.]+)", str(str1)) 根据邮件格式获取邮件地址。

2）对邮件服务器数量提取。str(df.xx_address.unique().shape) 将获取的邮件地址进行归一化处理，得到邮件收发服务器类别的数量。

3）对时间的提取。采用 rex=r"([A-Za-z]+\d?[A-Za-z]*) .*?(\d{2}):\d{2}:\d{2}.*" 提取时间。

同样利用正则表达式根据格式对时间进行提取，获取的结果少数为 none，另外一部分则根据时间段划分（由于某一封邮件是否是异常邮件并不能仅根据一个准确的时间来判断，因此划分不同时间段作为特征提取出来）。

4）对内容长度提取。根据数据清洗完成后的文件，通过二维表格形式读取，并获取内容列中不同的长度，然后对不同长度段分不同类型（由于某一封邮件是否是异常邮件并不能仅根据一个准确的内容长度来判断，因此划分不同内容长度类型作为特征提取出来），此处将内容长度不大于 10 的划分为 0，不大于 100 的划分为 1，不大于 500 的划分为 2，不大于 1 000 的划分为 3，…，不大于 50 000 的划分为 13，否则为 14。图 2 为邮件长度对异常邮件所占比例的影响。

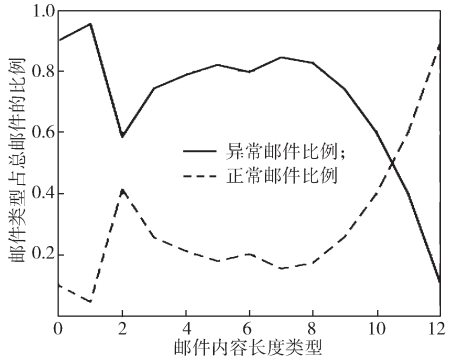


**图 2 邮件长度对异常邮件所占比例的影响**

Fig. 2 Effect of mail length on the proportion of abnormal mail

从图 2 的实验结果可以看出，邮件内容长度类型不大于 1 时，异常邮件占比高，不小于 2 时占比逐渐下降。而邮件内容长度类型在 2 到 10 之间时，异常邮件占比随内容长度呈现凸增长，在 7 的位置达到极大值，而后趋减。

### 2.3 数据集分割以及模型训练

在经过了上述的邮件集合处理以及特征提取之后，读取得到的文件并进行分割。利用 sklearn.model_selection 中的 train_test_split 随机将 10 000 封邮件集按照比例分为测试集和训练集，并输出到对应的文件夹下。再将训练集合中已经分词好的内容部分进行类型转换，从文本数据转换为数值型数据以进行特征提取，也就是 tf−idf 权重计算部分，即词频以及逆文本频率指数的计算，再将数据进行模型转换得到数据模型。

### 2.4 模型填充

由于在 sklearn 库中已有对各个算法的详细实现，本文只需按参数要求，向各个算法实现的函数填充数据参数即可获得对应的算法模型。另因本文主要研究随机森林算法，而随机森林算法又基于决策树，所以此处仅列出决策树算法和随机森林算法在模型

填充时候的参数选择。本文经过多次调参，力求得到最精确的结果。下面提供关于决策树分类器以及随机森林分类器主要参数。

decision_tree 算法：

在构建 decision_tree 模型时，采用 sklearn.tree 下的 DecisionTreeClassisfier 的决策树分类器模型，设置参数如下。

1）criterion 为切分质量的评价准则。默认为 'mse'(mean squared error)。

2）splitter 为在每个节点切分的策略。

3）max_depth 为指定树的最大深度。如果为 None，则表示树的深度不限，直到每个叶子都是纯净的，即叶节点中所有样本都属于同一个类别，或者叶子节点中包含小于 min_samples_split 个样本。

4）random_state。该参数如果为整数，则它指定了随机数生成器的种子；如果为 RandomState 实例，则指定了随机数生成器；如果为 None，则使用默认的随机数生成器。

5）max_leaf_nodes。如果为 None，则叶子节点数量不限。如果不为 None，则 max_depth 被忽略。

random_forest 算法：

random_forest 本身是建立在 decision_tree 的基础上，在构建 random_forest 模型时，采用 sklearn.svm 下的随机森林分类器模型，设置参数如下：

1）n_estimators。该参数为弱学习器的最大迭代次数，或者是最大弱学习器的个数。一般来说参数越小，越容易欠拟合；越大，越容易过拟合。默认为 10，实际参数和 learning_rate 一起考虑。

2）criterion。对树做划分时，对特征的评价标准。分类模型和回归模型的损失函数不同。分类 RF 对应的有基尼指数 gini，另一个标准是信息增益，回归 RF 默认是均方差 mse，另一个可选择的标准是绝对值差 mae，本文采用信息增益作为划分标准，下文将进行讨论。

3）max_depth。该参数为树的最大深度，默认为 None，直到使每一个叶节点只有一个类别，或是达到 min_samples_split。

4）random_state。如果给定相同的参数和训练数据，random_state 的确定值将始终产生相同的结果。一个具有不同随机状态的多个模型的集合，并且所有最优参数有时比单个随机状态更好。

# 3　算法描述

## 3.1　决策树算法

决策树是数据挖掘领域应用最广泛的方法之一，在很多实际应用中都被采用。它是一种非线性监督学习模型，能将数据分成不同的类别并对未知数据进行预测。决策模型将结果分解为 if-then-else 规则，并以树型结构展示。这种树形模型的高可读性使得人机更易于理解发现的知识。推断决策树的过程主要由以下几个方面决定：

1）分割标准，即用于选择要插入节点和分支属性的方法；

2）停止分支的标准；

3）在叶节点上分配类标签或概率分布的方法；

4）用于简化树结构的后修剪过程。

目前有两种分割标准：传统的分割标准和基于不精确概率的分割标准。区分它们的一个基本点是如何从数据中获得概率。通常，传统标准使用香农准则作为信息的基本测度。而基于不精确概率的准则使用最大熵测度，这种测量方法基于最大不确定度原理，在经典信息理论中被广泛使用，称为最大信息增益（information gain，IG）原理，本文在构建决策树时也是采用这种方法。

设属性 $X$ 为一般特征，其值属于 $\{x_1, x_2, \cdots, x_t\}$，信息增益 IG 解释如下：

1）数据集 $D$ 的熵 $C$ 定义为

$$H^D(C)=\sum_i p(c_i)\log_2\left(1/p(c_i)\right), \quad (1)$$

式中 $p(c_i)$ 为 $D$ 中 $i$ 类的概率。

2）属性 $X$ 生成的平均熵为

$$H^D(C\mid X)=\sum_i P^D(X=x_i)H^{D_i}(C\mid X=x_i), \quad (2)$$

式中：$P^D(X=x_i)$ 表示 $D$ 中 $X=x_i$ 的概率；$D_i$ 为 $D$ 的子集。

3）最后可得信息增益（IG）为

$$IG(C,X)^D=H^D(C)-H^D(C\mid X)。 \quad (3)$$

## 3.2　随机森林算法

随机森林是由多颗决策树构成的。如果必须对一个新实例进行分类，那么这个实例的特性将呈现给森林中的每颗决策树，每颗决策树返回一个分类值，投票给该类。最后，由随机森林给出的分类值是与类变量的最优投票相关联的值，超过了森林中的所有决策树。每颗决策树具有以下特征：

1）若 $N$ 是一个数据集中的实例数，那么随机森林从原始数据中选择一个随机样本，替换 $N$ 个实例，此样本将作为构建决策树的训练集。

2）若 $M$ 是数据集中的特征数，则指定一个 $m<<M$ 的数，在森林构建期间，$m$ 的值保持不变。

3）对于树中的每一个节点，

①从 $M$ 个原始特征中随机选择 $m$ 个特征；

②根据这 $m$ 个特征计算分割标准，具有最佳值的特征用于拆分节点。

4）在构建完每颗决策树之后没有修剪。

## 4 实验及结果分析

### 4.1 实验环境及数据说明

实验平台包括：

操作系统为 Windows10；

IDE 为 Pycharm 2019.1.1，Python 3.7.3；

实验数据为 10 000 封邮件，其中有一定数量异常邮件和一定数量非异常邮件，均由 IndexFile 的索引文件指明（spam 代表异常邮件，ham 代表非异常邮件）。

### 4.2 实验结果及分析

本文从 3 个指标对算法性能进行对比分析，具体定义如下：

*FN*（false negative），被判定为负样本、事实上是正样本的数目。

*FP*（false positive），被判定为正样本、事实上是负样本的数目。

*TN*（true negative），被判定为负样本、事实上也是负样本的数目。

*TP*（true positive），被判定为正样本、事实上也是正样本的数目。

准确率 = 所有预测正确的样本 / 总的样本，即，（*TP*+*TN*）/ 总样本数目；在本文中，准确率 = 对异常邮件测试集中预测的样本数目 / 所有测试集中的样本数目；

召回率 = 将正类预测为正类 / 所有正真的正类，即，*TP*/（*TP*+*TN*）；在本文中召回率 = 对异常邮件测试集中预测的样本数目 / 所有测试集中的异常邮件样本数目。

F1 值 = 正确率 * 召回率 *2/( 正确率 + 召回率 )；F1 值是精确率和召回率的调和平均数。

本文采用了准确率、召回率、F1 值 3 个主要的评判标准，并对 6 种算法，包括随机森林、K 最近邻（k-NearestNeighbor，KNN）、梯度提升树（gradient Boosting decison tree，gbdt）、贝叶斯、决策树、支持向量机（support vector machine，SVM），在上述 3 个标准和模型构建时间上进行对比。测试邮件集合的大小分别为 500, 1 000, 1 500, 2 000, 2 500，对比结果分别如图 3~6 所示。

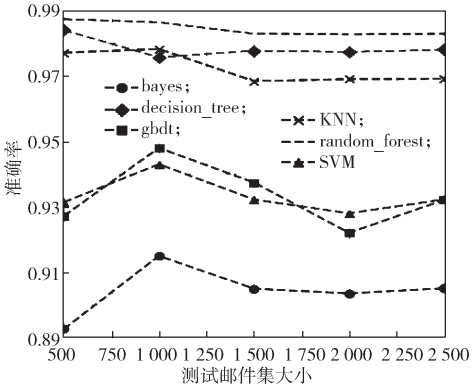


图 3 准确率对比图

Fig. 3 Accuracy comparison chart

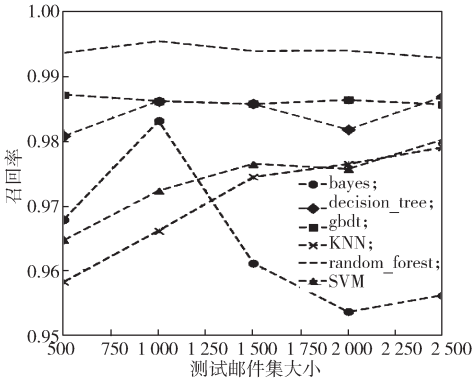


图 4 召回率对比图

Fig. 4 Recall rate comparison chart

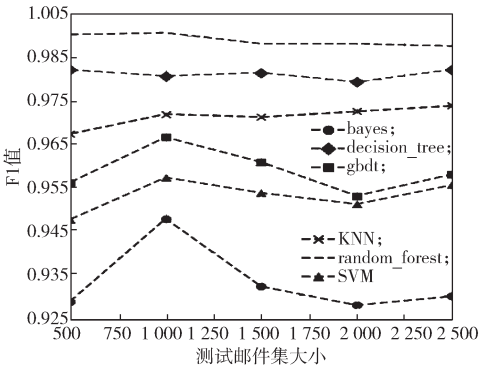


图 5 F1 值对比图

Fig. 5 Comparison chart of F1 values

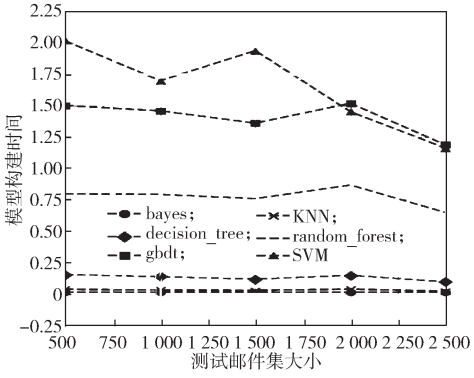


图 6 模型构建时间对比图

Fig. 6 Model construction time contrast diagram

从控制台输出结果以及对比图中不难发现，在同组训练集与测试集的情况下，随机森林算法的准确率为 0.985 89，召回率为 0.993 68，F1 值为 0.989 77，均优于其它算法。但在模型构建时间上随机森林算法慢于贝叶斯算法、决策树算法、KNN。不过，由于计算机性能不断增强，并且出现了云计算以及并行计算等计算模式，在模型构建时间上，并不是一个严重的问题。

## 5　总结与展望

异常邮件检测是一个概率性问题，准确率不高或者误判都会给用户带来困扰。通过实验分析表明本文采用的随机森林算法比其他几种算法有明显优势。但仍存在以下几个方面可以进一步研究：

1）一个足够大的邮件集合数据库对异常邮件的检测非常重要，样本量越大也越能高精准的预判未知邮件。因此合理构建一个共享的邮件集合仓库是有必要的。

2）异常邮件类型件在不断变化，利用单一的异常邮件检测机制是不合理的，可以考虑在不同的算法之间取长补短，将各种算法进行整合，以达到更高的准确率。

3）也同样由于本文只是针对于纯文本的邮件格式，格式单一，而异常邮件在当今社会不只是文本类型，比如音频、视频、压缩文件等异常邮件。近年来，研究人员对图像异常邮件的识别和过滤技术的研究较为关注，但当前研究出的过滤系统都不能很好地实现异常邮件图像的识别和分类，难以满足图像型异常邮件过滤的准确性、实时性及高效性要求。

**参考文献：**


[1] 李艳涛，冯伟森 . 堆叠去噪自编码器在垃圾邮件过滤中的应用 [J]. 计算机应用，2015，35(11)：3256-3260，3292.
LI Yantao，FENG Weisen. Application of Stacked Denoising Autoencoder in Spamming Filtering[J]. Journal of Computer Applications，2015，35(11)：3256-3260，3292.

[2] 杨　雷，曹翠玲，孙建国，等 . 改进的朴素贝叶斯算法在垃圾邮件过滤中的研究 [J]. 通信学报，2017，38(4)：140-148.
YANG Lei，CAO Cuiling，SUN Jianguo，et al. Study on an Improved Naive Bayes Algorithm in Spam Filtering[J]. Journal on Communications，2017，38(4)：140-148.

[3] 杨艳燕，郭红转，路新华 . 基于粗糙集的带决策规则边界的邮件过滤算法 [J]. 计算机应用研究，2015，32(1)：258-261.
YANG Yanyan，GUO Hongzhuan，LU Xinhua. E-mail Filtering Algorithm with Boundary Decision Rules Based on Rough Set[J]. Application Research of Computers，2015，32(1)：258-261.

[4] 沈元辅，沈跃伍 . 基于多层 grams 的在线支持向量机的中文垃圾邮件过滤 [J]. 中文信息学报，2015，29(1)：126-132.
SHEN Yuanfu，SHEN Yuewu. Typed N-gram for Online SVM Based Chinese Spam Filtering[J]. Journal of Chinese Information Processing，2015，29(1)：126-132.

[5] 孙　雪，韩　蕾，李昆仑 . 基于类别特征选择与反馈学习随机森林算法的邮件过滤系统研究 [J]. 计算机应用与软件，2015，32(4)：67-71.
SUN Xue，HAN Lei，LI Kunlun. On Email Filtering System Based on Category Feature Selection and Feedback Learning Random Forest Algorithm[J]. Computer Applications and Software，2015，32(4)：67-71.

[6] 王　辉，黄自威，刘淑芬 . 基于特征项区分度的加权朴素贝叶斯邮件过滤方法 [J]. 计算机应用与软件，2015，32(10)：67-71，81.
WANG Hui，HUANG Ziwei，LIU Shufen. Weighted Naive Bayes Spam Filtering Method Based on Feature Term Discrimination[J]. Computer Applications and Software，2015，32(10)：67-71，81.

[7] 宋智洋 . 一种基于规则的垃圾邮件过滤算法实现 [J]. 南方农机，2018，49(2)：137.
SONG Zhiyang. Implementation of a Rules-Based Spam Filtering Algorithm[J]. China Southern Agricultural Machinery，2018，49(2)：137.

[8] 袁国鑫，于　洪 . 一种基于邮件头信息的三支决策邮件过滤方法 [J]. 计算机科学，2017，44(9)：74-77，114.
YUAN Guoxin，YU Hong. Method of Three-Way Decision Spam Filtering Based on Head Information of E-Mail[J]. Computer Science，2017，44(9)：74-77，114.

[9] 齐浩亮，程晓龙，杨沐昀，等 . 高性能中文垃圾邮件过滤器 [J]. 中文信息学报，2010，24(2)：76-83.
QI Haoliang，CHENG Xiaolong，YANG Muyun，et al. High Performance Chinese Spam Filter[J]. Journal of Chinese Information Processing，2010，24(2)：76-83.

[10] 关晓蔷，庞继芳，梁吉业 . 基于类别随机化的随机森林算法 [J]. 计算机科学，2019，46(2)：196-201.
GUAN Xiaoqiang，PANG Jifang，LIANG Jiye. Randomization of Classes Based Random Forest Algorithm[J]. Computer Science，2019，46(2)：196-

201.

[11] 刘庆雄 . 基于数据驱动的垃圾邮件检测技术研究 [D]. 南昌：华东交通大学，2016.
LIU Qingxiong. The Detection Method of SPAM Based on Data Driven[D]. Nanchang：East China Jiaotong University，2016.

[12] 宋洪正 . 基于用户行为关系和内容的邮件分类算法的研究与实现 [D]. 成都：电子科技大学，2016.
SONG Hongzheng. Research and Implementation of Classification Algorithm Based on Message Content and User Behavior Relationship[D]. Chengdu：University of Electronic Science and Technology of China，2016.

[13] 蒋亚平，田月霞，梅 骁 . 基于免疫 Agent 的垃圾邮件过滤模型 [J]. 计算机应用与软件，2016，33(3)：294−298，313.
JIANG Yaping，TIAN Yuexia，MEI Xiao. A Spam Filtering Model Based on Immune-Agent[J]. Computer Applications and Software，2016，33(3)：294−298，313.

[14] RUSLAND N F，WAHID N，KASIM S，et al. Analysis of Naïve Bayes Algorithm for Email Spam Filtering Across Multiple Datasets[J]. IOP Conference Series：Materials Science and Engineering，2017，226：012091.

[15] PAUL A，MUKHERJEE D P，DAS P，et al. Improved Random Forest for Classification[J]. IEEE Transactions on Image Processing，2018，27(8)：4012−4024.

[16] TSAGKRASOULIS D，MONTANA G. Random Forest Regression for Manifold-Valued Responses[J]. Pattern Recognition Letters，2018，101：6−13.

[17] LIU M，LI Z R，ZHANG H T，et al. Feature Selection Algorithm Application in Near-Infrared Spectroscopy Classification Based on Binary Search Combined with Random Forest Pruning[J]. Laser & Optoelectronics Progress，2017，54(10)：103001.

[18] CAO W H，XU J P，LIU Z T. Speaker-Independent Speech Emotion Recognition Based on Random Forest Feature Selection Algorithm[C]//2017 36th Chinese Control Conference (CCC). Dalian：IEEE，2017：1072−1075.

[19] KOPRINSKA I，POON J，CLARK J，et al. Learning to Classify E-Mail[J]. Information Sciences，2007，177(10)：2167−2187.

[20] HU Y，GUO C，NGAI E W T，et al. A Scalable Intelligent Non-Content-Based Spam-Filtering Framework[J]. Expert Systems with Applications，2010，37(12)：8557−8565.

（责任编辑：申 剑）

（上接第 69 页）

[15] 汤亚南，魏 玲 . 注意力缺陷多动障碍儿童脑部核磁共振成像技术应用进展 [J]. 临床儿科杂志，2013，31(3)：277−279，282.
TANG Yanan，WEI Ling. The Progress in Magnetic Resonance Imaging of the Brain in Children with ADHD[J]. Journal of Clinical Pediatrics，2013，31(3)：277−279，282.

[16] DURSTON S. A Review of the Biological Bases of ADHD：What Have we Learned from Imaging Studies?[J]. Mental Retardation and Developmental Disabilities Research Reviews，2003，9(3)：184−195.

[17] 王 剑 . 注意缺陷 / 多动障碍儿童灰质体积的形态学研究及 3D-ASL 在 ADHD 的初步应用 [D]. 武汉：华中科技大学，2016.
WANG Jian. A Voxel Based Morphometry MRI Study of Altered Gray Matter Volume and the Preliminary Application of 3D-ASL in Attention Deficit Hyperactivity Disorder Children[D]. Wuhan：Huazhong University of Science and Technology，2016.

（责任编辑：申 剑）