基于 RSSI 的贝叶斯垃圾邮件过滤算法

陈铁军，靖丰年，段谊海

（郑州大学 电气工程学院，河南 郑州 450001）

摘 要：针对现有贝叶斯算法应用于垃圾邮件过滤时，贝努利模型精度低、不能区分文本特征重要性、多项式模型计算量大、无关特征项浪费计算时间、对出现次数少的特征项反应敏感等缺点，提出 RSSI（remove similar and sensitive items）特征模型。通过计算并比较特征项出现的频率，去除无关和敏感特征项，减小运算量，增加正确率，减少过拟合。Matlab 仿真结果表明，与现有的朴素贝叶斯算法（naïve Bayes）和支持向量机（support vector machine，SVM）等算法相比，RSSI 算法能显著减少分类时间，降低合法邮件被误判的概率。

关键词：邮件分类；贝叶斯分类器；特征提取；多项式事件模型；过拟合

中图法分类号：TP391.9 **文献标识号**：A **文章编号**：1000-7024（2015）07-1790-04

doi：10.16208/j.issn1000-7024.2015.07.022

RSSI-based Bayesian anti-spam filtering algorithm

CHEN Tie-jun, JING Feng-nian, DUAN Yi-hai

(College of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: When Bayesian algorithm is applied in spam filtering, Bernoulli model's accuracy is low and can not distinguish the importance of text features, and the multinomial model has larger computation. In addition, it is a waste of time in calculating unrelated feature elements and this model is sensitive to low frequency elements. For these shortcomings, an improved feature extraction algorithm named RSSI was proposed, which not only reduced the amount of computation, but also improved the classification performance by calculating and comparing the occurrence frequency of feature items, so that overfitting phenomenon was reduced. Experimental results show that compared with early naïve Bayes algorithm and SVM algorithm, the RSSI algorithm can significantly reduce the classification time and the probability of misjudging legitimate emails.

Key words: mail classification; Bayesian classifier; feature extraction; multinomial event model; overfitting

0 引 言

基于内容的垃圾邮件过滤法比一般白名单与黑名单技术、规则过滤以及基于关键词匹配的内容扫描等智能化程度高，可采用属于有监督学习的朴素贝叶斯分类器，实践结果表明分类效果佳。其中，贝叶斯过滤器是基于文本的过滤技术，准确率较高，但是，现有朴素贝叶斯分类器基于一个假设：从邮件中提取的文本特征是相互独立的。这是一个很强的假设，因为文本特征是相互关联的，所以现有贝叶斯算法过多的简化文本特征的相关性，导致判别垃圾邮件的召回率减低。在早期基于贝叶斯分类器的算法中，特征值被简化为 0 和 1，没有体现特征出现的概率，为更多利用文本特征的相关性，提出基于多项式模型的贝叶斯分类器[2]。与伯努利模型相对比，多项式模型更精确地描述了特征的重要性，然而，算法时间代价却激增由 $O(n)$ 上升到 $O(n^2)$。另外，对于那些出现次数较少的对判断会造成较大的误差。针对以上情况，本文提出基于 RSSI 特征选择器的贝叶斯垃圾邮件过滤算法，剔除无关特征和不稳定特征，有效减少过拟合，提高算法效率。与现有朴素贝叶斯算法（naïve Bayes）和支持向量机（support vector machine，SVM）等算法相比，RSSI 算法能显著减少分类时间，降低合法邮件被误判的概率。

收稿日期：2014-07-23；**修订日期**：2014-09-22

基金项目：教育部高等学校博士学科点专项科研基金项目（20114101110005）

作者简介：陈铁军（1954-），男，河南信阳人，教授，博士生导师，研究方向为复杂工业过程控制技术及控制系统；靖丰年（1993-），男，河南安阳人，硕士研究生，研究方向为模式识别与人工智能；段谊海（1988-），河南周口人，硕士研究生，研究方向为模式识别。E-mail：970523052@qq.com

1 朴素贝叶斯分类器的构建

朴素贝叶斯分类器是有监督学习的一种，分类器对邮件进行分类时，考虑到时间开支本文选择文档频数[3]（document frequency，DF）作为特征来进行建模。通过邮件解析和中文分词[4]预处理，对出现的词语生成一个词典，设邮件的特征向量为 $X=(x_1,x_2,\cdots,x_i,\cdots,x_n)$[5]，$x_i$ 表示每一封邮件的特征项。邮件共有两类：正常的邮件集合 G 和垃圾邮件集合 B，其中 $G=(G_1,G_2,\cdots,G_i,\cdots,G_n)$，$B=(B_1,B_2,\cdots,B_i,\cdots,B_n)$。设邮件 E 的特征向量为 $X^E=(x_1^E,x_2^E,\cdots,x_i^E,\cdots,x_n^E)$，根据贝叶斯公式，则邮件 E 属于垃圾邮件的概率为

$$P(B\mid X^E)=\frac{P(X^E\mid B)P(B)}{P(X^E)} \tag{1}$$

$P(X^E)$ 可由全概率公式 $P(X^E)=P(X^E\mid B)+P(X^E\mid G)$ 算得，$P(B)$ 是训练样本的先验概率，根据朴素贝叶斯假设，每个特征项 x_i 条件独立，则 $P(X^E\mid B)=\prod_{i=1}^{n}P(x_i^E\mid B)$。

就伯努利模型而言，x_i 取 1 或 0，设参数 $\phi_{i|B}=P(x_i=1\mid B)$，$\phi_{i|G}=P(x_i=1\mid G)$，$\phi_y=P(B)$，给定一个训练集合 $\{(x^{(i)},y^{(i)});i=1\cdots m\}$，可以得出参数 $\phi_{i|B}$、$\phi_{i|G}$、ϕ_y 的似然函数

$$\ell(\phi_y,\phi_{i|B},\phi_{i|G})=\prod_{i=1}^{m}P(x^{(i)},y^{(i)}) \tag{2}$$

分别求出 ϕ_y，$\phi_{i|B}$，$\phi_{i|G}$ 的极大似然估计

$$\phi_{j|B}=\frac{\sum_{i=1}^{m}1\{x_j^{(i)}=1\wedge y^{(i)}=1\}}{\sum_{i=1}^{m}1\{y^{(i)}=1\}} \tag{3}$$

$$\phi_{j|G}=\frac{\sum_{i=1}^{m}1\{x_j^{(i)}=1\wedge y^{(i)}=0\}}{\sum_{i=1}^{m}1\{y^{(i)}=0\}} \tag{4}$$

$$\phi_y=\frac{\sum_{i=1}^{m}1\{y^{(i)}=1\}}{m} \tag{5}$$

$y^{(i)}=1$ 表示第 i 个训练样本属于 B 集合，$1\{\}$ 表示指示符号，规定 $1\{true\}=1$，$1\{fault\}=0$，进而可以得出给定测试邮件属于垃圾邮件的概率

$$P(B\mid X)=\frac{(\prod_{i=1}^{n}P(x_i\mid B))P(B)}{(\prod_{i=1}^{n}P(x_i\mid B))P(B)+(\prod_{i=1}^{n}P(x_i\mid G))P(G)}$$
$$=\frac{(\prod_{i=1}^{n}\phi_{i|B}{}^{x_i}(1-\phi_{i|B})^{(1-x_i)})\phi_y}{(\prod_{i=1}^{n}\phi_{i|B}^{x_i}(1-\phi_{i|B})^{(1-x_i)})\phi_y+(\prod_{i=1}^{n}\phi_{i|G}^{x_i}(1-\phi_{i|G})^{(1-x_i)})(1-\varphi_y)} \tag{6}$$

2 贝叶斯多项式模型

伯努利模型贝叶斯分类器可根据特征项的出现与否计算给定测试邮件与正常邮件和垃圾邮件的匹配程度，对邮件分类。由于伯努利模型 x_i 只能取两个值 1 和 0，为表达更多的特征信息，提出了多项式模型[6]，此时对 $X=(X_1,X_2,\cdots,X_i,\cdots,X_n)$，$X_i$ 表示邮件经中文分词后第 i 个字符的标识，n 表示邮件的长度。给定参数 $\phi_{i|B}=P(x_j=i\mid B)$，$\phi_{i|G}=P(x_j=i\mid G)$，$\phi_y=P(B)$，给定训练集合 $\{(x^{(i)},y^{(i)});i=1,\cdots,m\}$，这里 $x^{(i)}=(x_1^{(i)},x_2^{(i)},\cdots,x_{n_i}^{(i)})$，$n_i$ 是第 i 个训练样本的邮件长度，可以得到参数 $\phi_{i|B}$，$\phi_{i|G}$，ϕ_y 的似然函数

$$\ell(\phi_{i|B},\phi_{i|G},\phi_y)=\prod_{i=1}^{m}P(x^{(i)},y^{(i)})=$$
$$\prod_{i=1}^{m}(\prod_{j=1}^{n_i}P(x_j^{(i)}\mid y;\phi_{i|B},\phi_{i|G}))P(y^{(i)};\phi_y) \tag{7}$$

求得参数的极大似然估计

$$\phi_{k|B}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}1\{x_j^{(i)}=k\wedge y^{(i)}=1\}}{\sum_{i=1}^{m}1\{y^{(i)}=1\}n_i} \tag{8}$$

$$\phi_{k|G}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}1\{x_j^{(i)}=k\wedge y^{(i)}=0\}}{\sum_{i=1}^{m}1\{y^{(i)}=0\}n_i} \tag{9}$$

$$\phi_y=\frac{\sum_{i=1}^{m}1\{y^{(i)}=1\}}{m} \tag{10}$$

$x_j^{(i)}=k$ 表示第 i 个训练样本中第 j 个字符为字典中的标号 k。给定测试样本 X，则 X 属于垃圾邮件的概率为

$$P(B\mid X)=\frac{(\prod_{i=1}^{n}P(x_i\mid B))P(B)}{(\prod_{i=1}^{n}P(x_i\mid B))P(B)+(\prod_{i=1}^{n}P(x_i\mid G))P(G)}$$
$$=\frac{(\prod_{i=1}^{n}\phi_{i|B}{}^{t_i}(1-\phi_{i|B})^{(1-t_i)})\phi_y}{(\prod_{i=1}^{n}\phi_{i|B}^{t_i}(1-\phi_{i|B}^{(1-t_i)}))\phi_y+(\prod_{i=1}^{n}\phi_{i|G}{}^{t_i}(1-\phi_{i|G})^{(1-t_i)})(1-\phi_y)} \tag{11}$$

t_i 取 1 或 0，以上就是多项式模型贝叶斯分类器的数学模型。比较式（6）和式（11），发现在计算判别 X 属于垃圾邮件的概率时，计算量差别在 $\phi_{i|B}$ 上（注意两式中 $\phi i\mid B$ 不同），比较式（3）和式（8）发现多项式模型 $\phi_{k|B}$ 的计算量是 $O(n^2)$，而伯努利模型中 $\phi_{i|B}$ 的计算量仅为 $O(n)$。对计算机来说 $O(n^2)$ 的运算量尚可接受，精简算法结构将大幅减少运算时间，下面将对生成的特征向量 X 简化。

3 RSSI 算法过程

对式（11）剖析发现计算了大量的无关特征项[9]，如

字典中收纳的量词、语气词等，对正常邮件和垃圾邮件建模过程中发现，这些词的在正常邮件和垃圾邮件出现概率 $\phi_{k|G}$ 和 $\phi_{k|B}$ 大致相同，固可以利用这个特征来对生成的字典进行合理的“瘦身”。

设阈值为 $T=5\%$，如果超过了这个范围，认为此为相关特征项，如果在这个范围内，则认为是无关特征项，在生成字典中删除此特征项

$$\begin{aligned} &if \quad \phi_{k|B}\in[(1-T)\phi_{k|G},(1+T)\phi_{k|G}] \\ &then \qquad\qquad delete[\ ] \end{aligned} \tag{12}$$

在对一些错误分类的邮件研究发现，有些出现频次很小的特征项是导致分类错误的主因，比如特征项 x_i 在集合 G 中出现一次，而没有在集合 B 中出现，显然这存在很大的偶然因素，然而贝叶斯分类器通过对特征项 x_i 的计算后将有极大的倾向把邮件 X 分给集合 G。为了克服这个缺点，我们设阈值 $T_f=5$，有关计算公式如下

$$\begin{aligned} &if\ \phi_{k|B}\in\left[0,\frac{T_f}{\sum_{i=1}^{m}1\{y^{(i)}=1\}n_i}\right]\wedge\ \phi_{k|G}\in\left[0,\frac{T_f}{\sum_{i=1}^{m}1\{y^{(i)}=1\}n_i}\right] \\ &then \qquad\qquad\qquad\qquad delete[\ ] \end{aligned} \tag{13}$$

到此，我们降低了 X 维数，降低了算法的运算量，减少了内存空间的消耗，并提高了分类正确率。

4 实验分析

4.1 实验环境

在上述基于多项式模型的贝叶斯垃圾邮件分类算法中，采用 RSSI 方法降低特征维数，本文仿真基于 matlab 平台，WEKA Java API 和 Eclipse 开发环境，选用 PU 系列英文语料库和 ZH1 中文语料库，实验方法采用“十字交叉验证法”[10]，将每一个语料库中平均分 10 份，9 份作为训练集，1 份作为测试集，采用原始处理训练集和测试集 txt 文本集为一个 $m\times n$ 维的矩阵[11]，m 为集合中元素的个数，n 为几何元素的权值，1 表示对应特征项出现，0 表示不出现。

4.2 评价指标

为了评价算法的好坏，引入正确率（$R_{Precision}$）和召回率（R_{Recall}）两个概念

$$R_{\text{Precision}}=\frac{A}{A+B}\times 100\% \tag{14}$$

式中：A ——垃圾邮件被正确分类的数量，B ——被错误判定为垃圾邮件的数量。正确率越高，正确分类垃圾邮件和正常邮件的数量越多

$$R_{\text{Recall}}=\frac{A}{A+C}\times 100\% \tag{15}$$

式中：A ——正确区分垃圾邮件的数量，C ——漏掉的垃圾邮件的数量。召回率越高，检测到的垃圾邮件越多，漏掉的垃圾邮件越少。

4.3 实验结果

实验中，首先对传统贝叶斯算法和本文改进算法对邮件的分类正确率进行测试，采用“十字交叉验证法”，首先选取集合中的 70% 作为样本集，利用 Weka 软件对其进行训练，建立特征集。然后对集合中剩下的 30% 提取特征，进行分类测试，随着特征数目的不算增加，对分类正确率的影响如图 1 所示。

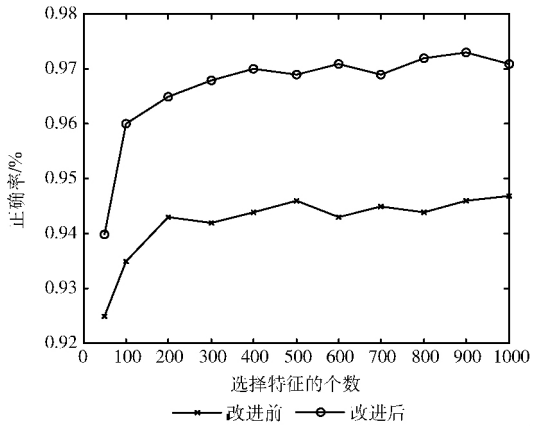

图 1 两种算法下选择特征个数对邮件正确率的影响

从图 1 可以看出，当特征个数较少时，分类准确率较低，是因为特征项不能完全反映文本特征，导致文本区分度不高，当特征数量大于等于 200 时，分类准确率趋于稳定，此时出现了大量冗余的特征项。观察到改进后的方法在特征个数较少时仍有很高的准确率，趋于稳定时特征个数比改进前少，这是因为改进算法剔除了大量的无关特征性，使算法在仅有少量特征个数下就能充分反映文本特征。

研究特征数目对召回率的影响，实验结果如图 2 所示。

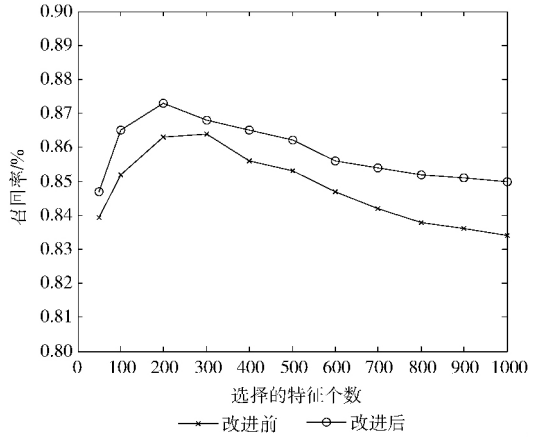

图 2 两种算法下选择特征个数对邮件召回率的影响

从上图中可以看出，在特征项个数较少时，邮件召回率较低，这是因为特征个数不能充分描述垃圾邮件文本，导致漏掉的垃圾邮件过多，随着特征数量的增加，召回率出现极大值，其原因是特征项正好反映文本特征，不同类

型的邮件区分度高。随着特征数量的增多，算法召回率下降趋于平稳，这是因为掺杂与分类无关的特征项，产生分类干扰，使分类效果变坏。两条曲线比较发现，改进前的极值出现在 300 左右，而改进后的极值在 200 左右，这是因为剔除了无关特征项使极值提前，因为剔除了一部分分类干扰，所以提高了召回率。

测试改进算法的时间代价，对分类器进行训练后，随机选取测试集中 100 份邮件进行测试，得出结果见表 1。

表 1 传统贝叶斯算法和本文算法的分类时间

参数	改进前	改进后
分类时间/ μs	583	376

分类时，对测试邮件的特征向量 X 分别与垃圾邮件模型和正常邮件模型相匹配，看哪种模型匹配度高。改进后算法分类时间短的原因是去除了无关特征项，在最后计算 $P(B \mid X)$ 时特征项数比改进前少，固节省了计算时间。

4.4 与其它垃圾邮件算法的比较

对于时下流行的 K 最邻近算法、支持向量机算法用于垃圾邮件分类，实验选取了共 1089 封邮件的实验集，对比了 KNN 算法、SVM 算法和 RSSI 算法，实验结果见表 2。

表 2 RSSI 算法、KNN 算法、SVM 算法性能对比

算法	正确率/%	召回率/%	分类时间
KNN	96.3	85.9	5582
SVM	96.9	88.3	6447
RSSI	97.2	87.4	3973

从表 2 中可以看出：①从邮件过滤性能上看，基于 RSSI 的贝叶斯垃圾邮件过滤算法的正确率和召回率与 SVM 算法相当，但比 KNN 算法要好；②从邮件过滤速度上看，基于 RSSI 的贝叶斯垃圾邮件过滤算法要比 KNN 算法和 SVM 算法快一倍以上，这是因 RSSI 算法有效减少特征项，降低了计算机的工作量。

5 结束语

贝叶斯垃圾邮件分类模型是广泛使用的一种垃圾邮件分类模型，但是需要使用大量的训练集合训练，占用大量网络资源和系统资源见下方[12]。本文提出了基于 RSSI 的贝叶斯垃圾邮件过滤算法，与传统贝叶斯垃圾邮件分类机制相比：①本文算法能去除无关特征，使召回率极大值提前，准确率在取较小值即达到平稳，改进了算法性能；②本文算法由于去除了无关特征的干扰，提高了准确率和召回率，减少了计算时间，提高了效率。

与 KNN 算法和 SVM 算法相比，性能相当，但由于简化了算法，执行效率得到了大幅提升。

参考文献：

[1] ZHENG Dongdong，SONG Shunlin. Survey of image spam filtering technology [J]. Computer Engineering and Design，2010，31 (1)：41-44 (in Chinese). [郑冬冬，宋顺林. 图片垃圾邮件过滤技术综述 [J]. 计算机工程与设计，2010，31 (1)：41-44.]

[2] EP Sanz JGHJ. Email spam filtering [J]. Advances in Computers，2008，74：45-114.

[3] LI Xiao，LUO Junyong，YIN Meijuan. Email filtering based on structural feature analysis and text classification [J]. Computer Engineering and Design，2010，31 (21)：4555-4558 (in Chinese). [李潇，罗军勇，尹美娟. 基于结构特征分析与文本分类的邮件筛选 [J]. 计算机工程与设计，2010，31 (21)：4555-4558.]

[4] YANG Kaifeng，ZHANG Yikun，LI Yan. Feature selection method based on document frequency [J]. Computer Engineering，2010，36 (17)：33-35 (in Chinese). [杨凯峰，张毅坤，李燕. 基于文档频率的特征选择方法 [J]. 计算机工程，2010，36 (17)：33-35.]

[5] LIU Hongzhi. Research on Chinese word segmentation techniques [J]. Computer Development & Applications，2010，23 (3)：1-3 (in Chinese). [刘红芝. 中文分词技术的研究 [J]. 电脑开发与应用，2010，23 (3)：1-3.]

[6] LIANG Zhiwen，YANG Jinmin，LI Yuanqi. A Bayesian spam filtering algorithm based on polynomial model and low risk [J]. Journal of Central South University (Science and Technology)，2013，44 (7)：2787-2792 (in Chinese). [梁志文，杨金民，李元旗. 基于多项式模型和低风险的贝叶斯垃圾邮件过滤算法 [J]. 中南大学学报 (自然科学版)，2013，44 (7)：2787-2792.]

[7] ZHAO Jing. Research on spam filtering technologies based on content characteristics analysis [D]. Shandon：Shandong Normal University，2012：7-15 (in Chinese). [赵静. 基于内容特征分析的垃圾邮件过滤关键技术研究 [D]. 山东：山东师范大学，2012：7-15.]

[8] ZHENG Wei，SHEN Wenzhang，YING Peng. Implementing spam filter by improving naive Bayesian algorithm [J]. Journal of Northwestern Polytechnical University，2010，28 (4)：623-627 (in Chinese). [郑炜，沈文张，英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究 [J]. 西北工业大学学报，2010，28 (4)：623-627.]

[9] FU Huitao，Kamil Moydin. Study and design of an improved text feature selection method [J]. Computer Applications and Software，2011，28 (4)：238-241 (in Chinese). [符会涛，木衣丁·卡米力. 一种改进的文本特征选择方法的研究与设计 [J]. 计算机应用与软件，2011，28 (4)：238-241.]

[10] Blanzieri E，Bryl A. A survey of learning-based techniques of email spam filtering [J]. Artif Intell Rev，2008，29：63-92.

[11] LUO Qin，LIU Bing，YAN Junhua，et al. Research of a spam filtering algorithm based on naive Bayes and AIS [C] // International Conference on Computational and Information-Sciences. Washington：IEEE，2010：152-155.

[12] Kosmopoulos A，Paliouras G，Androutsopoulos I. Adaptive spam filtering using only naive bayes text classifiers [C] // Fifth Conference on Email and Anti-Spam，2008.