第 46 卷 第 12 期 计 算 机 工 程 2020 年 12 月
Vol. 46 No. 12 Computer Engineering December 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)12-0299-06 文献标志码：A 中图分类号：TP391

基于 IMI-WNB 算法的垃圾邮件过滤技术研究

刘 洁，王 铮，王 辉

（河南理工大学 计算机科学与技术学院，河南 焦作 454000）

摘 要： 互信息和朴素贝叶斯算法应用于垃圾邮件过滤时，存在特征冗余和独立性假设不成立的问题。为此，提出一种改进互信息的加权朴素贝叶斯算法。针对互信息效率较低的问题，通过引入词频因子与类间差异因子，提出一种改进的互信息特征选择算法，从而实现更高效的特征降维。针对朴素贝叶斯分类算法的独立性假设问题，在朴素贝叶斯分类时使用改进互信息值进行特征加权，消除部分朴素贝叶斯条件独立性假设对邮件分类的不利影响。实验结果表明，相比传统朴素贝叶斯算法，该算法提高了垃圾邮件过滤的精确度、召回率与稳定性。

关键词： 互信息；垃圾邮件过滤；加权朴素贝叶斯算法；特征选择；词频

开放科学（资源服务）标志码（OSID）：

中文引用格式： 刘洁，王铮，王辉. 基于 IMI-WNB 算法的垃圾邮件过滤技术研究[J]. 计算机工程，2020，46(12)：299-304，312.

英文引用格式： LIU Jie，WANG Zheng，WANG Hui. Research on spam filtering technology based on IMI-WNB algorithm[J]. Computer Engineering，2020，46(12)：299-304，312.

Research on Spam Filtering Technology Based on IMI-WNB Algorithm

LIU Jie，WANG Zheng，WANG Hui

（School of Computer Science and Technology，Henan Polytechnic University，Jiaozuo，Henan 454000，China）

【Abstract】 The application of Mutual Information（MI） and Naive Bayes（NB） algorithm to spam filtering is faced with feature redundancy and invalid independence assumption. To address the problem，this paper proposes an Improved Mutual Information-Weighted Naive Bayes（IMI-WNB） algorithm. As for the low efficiency of mutual information，an improved feature selection algorithm based on MI is proposed by introducing the word frequency factor and inter-class difference factor in order to achieve more efficient feature dimensionality reduction. To solve the problem of independence assumption of NB classification algorithm，the Improved Mutual Information（IMI） value is used for feature weighting in NB classification，which eliminates the adverse effect of part of the NB conditional independence assumption on mail classification. The experimental results show that compared with the traditional NB algorithm，the proposed algorithm improves the accuracy，recall rate and stability of spam filtering.

【Key words】 Mutual Information（MI）；spam filtering；Weighted Naive Bayes（WNB） algorithm；feature selection；word frequency

DOI： 10.19678/j.issn.1000-3428.0056577

0 概述

电子邮件能够为用户间的通信提供便捷，但垃圾邮件也随之产生。根据卡巴斯基信息安全网站 Securelist 发布的 2018 年全球垃圾邮件数据显示，中国成为全球第一垃圾邮件来源地，占全球垃圾邮件来源的 11.69%[1]。垃圾邮件不仅占据大量的网络带宽和邮箱空间，容易造成网络拥堵，而且其包含一些恶意软件和钓鱼网站，可能会给用户带来巨大的经济损失。因此，对垃圾邮件进行过滤研究具有重要意义。

目前，关于垃圾邮件的过滤技术主要有基于黑白名单过滤技术、基于行为模式识别技术以及基于内容的过滤技术。其中，基于内容的过滤技术可行

基金项目： 国家自然科学基金（61300216）。

作者简介： 刘 洁（1979—），女，副教授、硕士，主研方向为网络安全、数据库技术、软件技术；王 铮（通信作者），硕士研究生；王 辉，副教授、博士。

收稿日期： 2019-11-13 **修回日期：** 2019-12-25 **E-mail：** wz960123@qq.com

性较高、耗费较少,已经成为当前研究垃圾邮件过滤技术的主流方向[2-3],主要包括支持向量机(Support Vector Machine,SVM)、K 邻近(K-Nearest Neighbor,KNN)、朴素贝叶斯(Naive Bayes,NB)等[4]。朴素贝叶斯分类器实现较为简单、且准确率高,已成为对垃圾邮件进行过滤的广泛应用分类方法[5-6],但该方法基于条件独立性假设,即假设条件之间完全独立,在一定程度上影响了分类结果的精确度。

在对垃圾邮件分类前,特征选择算法的优劣性对分类效果会造成影响,常见的几种特征选择方法有文档频率(DF)、信息增益(IG)、TF-IDF、开方拟合检验(χ^2test)和互信息(Mutual Information,MI)等。其中,互信息效果虽然较差,但是该方法复杂度低、容易理解,是普遍使用的一种特征选择方法[7-8]。传统的互信息方法没有计算特征词的频度,可能会出现低频词汇的互信息值较高的情况,导致分类精确度受到影响[9-10]。

针对特征冗余和独立性假设的问题,研究人员对特征选择和分类算法进行改进,以提高邮件的分类精度。文献[11]将朴素贝叶斯、随机树和随机森林 3 种机器学习算法应用于垃圾邮件数据集,其分类精度高于仅基于贝叶斯分类器的算法。文献[12]提出一种支持向量机算法与 K-均值聚类算法相结合的邮件分类算法,以提高分类精度、减少训练时间。文献[13]将互信息应用于加权朴素贝叶斯,通过加权部分消除朴素贝叶斯条件独立性假设对分类效果的影响,从而提高了朴素贝叶斯的文本分类效果,但该方法存在没有对传统的互信息算法进行改进的问题。文献[14]提出一种 TSVM-NB 算法,该算法利用朴素贝叶斯算法进行初次训练,并使用支持向量机算法构造最优分类超平面以降低特征项维度。同时,再次利用朴素贝叶斯算法生成分类模型,提高垃圾邮件过滤的速度和正确率,但该算法适用于属性向量重叠较大的语料集,对混叠性较弱的语料集的效率提升有限。文献[15]引入熵的思想,并结合 MapReduce 技术提出一种基于 MapReduce 的改进互信息文本特征选择机制,提高文本分类的精度。文献[16]提出一种基于 MapReduce 的并行特征选择方法,利用最大互信息理论选择信息丰富的特征变量组合。上述方法仅改进分类过程中的特征选择算法,并未联同分类算法对分类进行综合改进。

在以上研究基础上,本文提出一种基于改进互信息的加权朴素贝叶斯(Improved Mutual Information-Weighted Naive Bayes,IMI-WNB)算法。在特征选择阶段,引入词频因子以及类间差异因子对传统的互信息算法进行改进,实现特征降维。在分类阶段引入改进的互信息(IMI)值对朴素贝叶斯算法进行属性加权,实现对垃圾邮件的精确分类。

1 改进的互信息算法

1.1 互信息算法

垃圾邮件在经过文本预处理后引入大量特征项,然而大量的特征项对于分类没有意义,属于噪音特征,不对其进行降维处理将会影响垃圾邮件过滤的分类效果[17]。互信息算法是特征选择算法的一种,互信息值表示出特征项与类别之间的相关程度,且互信息值越大,则该特征项与类别的关联性越紧密。互信息值的计算方法为:

$$\mathrm{MI}(w,C)=\mathrm{lb}\frac{P(w,C)}{P(w)P(C)}=\mathrm{lb}\frac{P(w\mid C)}{P(w)} \tag{1}$$

其中,w 表示特征项,C 表示类别,$P(w,C)$表示特征项 w 与类别 C 共同出现的概率,$P(w)$表示特征项在整个训练文本中出现的概率,$P(C)$表示训练文本中该类别在训练文本中出现的概率,$P(w\mid C)$表示特征项 w 在类别 C 中出现的概率。

m 个类别训练文本的互信息值计算方法为:

$$\mathrm{MI}(w_i)=\sum_1^m P(C_j)\times\mathrm{MI}(w_i,C_j)=\sum_1^m P(C_j)\times\mathrm{lb}\frac{P(w_i\mid C_j)}{P(w_i)} \tag{2}$$

通过式(2)计算出互信息值,并选取合适的阈值,可针对分类不重要的特征项进行过滤,从而实现特征的选择。

1.2 基于词频因子与类间差异因子的 IMI 算法

1.2.1 词频因子

互信息算法的计算方式只考虑到特征词的文本频率而没有考虑到词频,这在一定程度上会影响其分类精度。例如,2 个特征项 w_j 和 w_q 的文本频率相同,且特征项 w_j 的词频是特征项 w_q 词频的数倍,即 $\mathrm{tf}(w_j)>>\mathrm{tf}(w_q)$,一般认为词频更大的特征项 w_j 与该类别的相关程度更高。然而按照传统互信息的计算方式,这 2 个特征项的互信息值是相同的,这显然与实际情况不符。因此,引入词频因子 α 对不同特征项间的词频差异进行描述,词频因子 α 可定义为:

$$\alpha_{ij}=\frac{\mathrm{tf}_{C_j}(w_i)}{\mathrm{df}_{C_j}(w_i)} \tag{3}$$

$$\alpha_i=\frac{\mathrm{tf}_{C_{\mathrm{spam}}}(w_i)}{\mathrm{df}_{C_{\mathrm{spam}}}(w_i)}+\frac{\mathrm{tf}_{C_{\mathrm{ham}}}(w_i)}{\mathrm{df}_{C_{\mathrm{ham}}}(w_i)} \tag{4}$$

其中,$\mathrm{tf}_{C_{\mathrm{spam}}}(w_i)$与 $\mathrm{tf}_{C_{\mathrm{ham}}}(w_i)$分别为特征项 w_i 的垃圾邮件与非垃圾邮件类词频,$\mathrm{df}_{C_{\mathrm{spam}}}(w_i)$表示特征项 w_i 的垃圾邮件类文本频率,$\mathrm{df}_{C_{\mathrm{ham}}}(w_i)$表示特征项 w_i 的非垃圾邮件类文本频率。

引入词频因子 α 后,改进的互信息值计算方法为:

$$\mathrm{IMI}(w_i)=\alpha_i\times\sum_{1}^{m}P(C_j)\times\mathrm{lb}\frac{P(w_i|C_j)}{P(w_i)} \tag{5}$$

特征项的词频高于文本频率时,词频因子的权重越大,说明该特征项对邮件分类的能力越强。

1.2.2　类间差异因子

如果特征项在 2 个类别中都平均分布时,则不利于类别的判定,在某一类别出现较多而在另一类别中极少出现,一般认为该特征项对于邮件类别的判别作用较大。在概率统计中标准差反映了数据集的离散程度,标准差较大的特征项更利于邮件类别的判定。通过计算垃圾邮件类 C_{spam} 与非垃圾邮件类 C_{ham} 之间特征项 w_i 频数的标准差对互信息模型进行改进。假设特征项 w_i 在垃圾邮件 C_{spam} 类中的频数为 $\mathrm{tf}_{C_{\mathrm{spam}}}(w_i)$,在非垃圾邮件 C_{ham} 类中的频数为 $\mathrm{tf}_{C_{\mathrm{ham}}}(w_i)$,频数平均值为 $\mathrm{tf}_{\mathrm{avg}}(w_i)$,则有:

$$\mathrm{tf}_{\mathrm{avg}}(w_i)=\frac{1}{2}(\mathrm{tf}_{C_{\mathrm{sapm}}}(w_i)+\mathrm{tf}_{C_{\mathrm{ham}}}(w_i)) \tag{6}$$

引入类间差异因子 σ 对类间词频差异进行描述,类间差异因子 σ 定义为:

$$\sigma_{ij}=\sqrt{(\mathrm{tf}_{C_j}(w_i)-\mathrm{tf}_{\mathrm{avg}}(w_i))^2} \tag{7}$$

$$\sigma_i=\sqrt{\frac{1}{2}[(\mathrm{tf}_{C_{\mathrm{spam}}}-\mathrm{tf}_{\mathrm{avg}}(w_i))^2+(\mathrm{tf}_{C_{\mathrm{ham}}}-\mathrm{tf}_{\mathrm{avg}}(w_i))^2]} \tag{8}$$

引入类间差异因子 σ 后,改进的互信息值计算方法为:

$$\mathrm{IMI}(w_i)=\alpha_i\times\sigma_i\times\sum_{1}^{m}P(C_j)\times\mathrm{lb}\frac{P(w_i|C_j)}{P(w_i)} \tag{9}$$

式(9)在式(5)的基础上增加了类间频数差异权重因子,体现出类间频数差异对邮件分类的影响,提高互信息算法对有效特征项的选择效率。

1.2.3　IMI 算法描述

算法 1　IMI 算法

输入　邮件特征向量集 $\boldsymbol{T}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$,特征子集维度 k

输出　特征子集 $\boldsymbol{F}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_k\}$

1. 计算 $P(C_{ham})$ 和 $P(C_{spam})$
2. for i = 1 to n
3. 统计词频 $tf_{C_{spam}}(w_i)$ 和 $tf_{C_{ham}}(w_i)$
4. 统计文档频率 $df_{C_{spam}}(w_i)$ 和 $df_{C_{ham}}(w_i)$
5. 计算 $P(w_i|C_{spam})$ 和 $P(w_i|C_{ham})$
6. 计算 $P(w_i)$
7. 式(2)计算互信息值 $MI(w_i)$
8. 式(4)计算词频因子 α_i
9. 式(8)计算类间差异因子 σ_i
10. 将式(2)、式(4)、式(8)结果代入式(9),计算 IMI 值
11. end
12. Sort(T) //将特征向量按 IMI 值降序排列
13. for i = 1 to k
14. 将特征项 w_i 加入特征子集 F 中
15. end

算法 1 是 IMI-WNB 算法中特征选择阶段的算法,IMI 算法改进了传统互信息算法中只考虑到文本频率而未考虑到词频的问题,定义并引入词频因子与类间差异因子,体现词频与类间词频差异对分类的贡献度,在完成特征降维的同时,还增强了特征项的表达能力。

2　基于 IMI 的朴素贝叶斯分类算法

2.1　朴素贝叶斯分类模型

朴素贝叶斯分类是基于贝叶斯定理与特征条件独立假设的分类方法,其通过计算已有的事件训练集得到事件概率,并对事件发生的概率进行预测。给定类别 C_j 与文本对象 d 时,贝叶斯公式可表示为:

$$P(C_j|d)=\frac{P(d|C_j)P(C_j)}{P(d)} \tag{10}$$

其中,$P(C_j)$ 表示 C_j 类发生的先验概率,对于垃圾邮件分类,类别 C 可被分为垃圾邮件与非垃圾邮件,即 $C=\{C_{\mathrm{spam}},C_{\mathrm{ham}}\}$。$P(C_j|d)$ 表示在给定输入文本对象为 d 时,该对象属于类别 C_j 的后验概率。假设文本 d 的特征项为 $\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$,根据朴素贝叶斯条件独立性假设,则有:

$$P(d|C_j)=P(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n|C_j)=P(\boldsymbol{w}_1|C_j)\times P(\boldsymbol{w}_2|C_j)\times\cdots\times P(\boldsymbol{w}_n|C_j)=\prod_{i=1}^{n}P(\boldsymbol{w}_i|C_j) \tag{11}$$

将式(11)代入式(10)可得:

$$P(C_j|d)=\frac{P(C_j)\prod_{i=1}^{n}P(\boldsymbol{w}_i|C_j)}{P(d)} \tag{12}$$

先验概率 $P(d)$ 为标准化常量,是一个常数。因此,朴素贝叶斯计算的最大后验概率类别 C_{map} 如下所示:

$$C_{\mathrm{map}}=\arg\max_{C_j\in C}P(C_j)\prod_{i=1}^{n}P(\boldsymbol{w}_i|C_j) \tag{13}$$

为了避免大量较小数相乘造成下溢出问题,对式(13)乘积取对数可得:

$$C_{\mathrm{map}}=\arg\max_{C_j\in C}\left[\mathrm{lb}\,P(C_j)+\sum_{i=1}^{n}\mathrm{lb}\,P(\boldsymbol{w}_i|C_j)\right] \tag{14}$$

2.2　基于 IMI 的加权朴素贝叶斯分类器

朴素贝叶斯分类算法是基于条件独立性假设的分类方法,然而在实际应用中,该独立性假设通常不成立。为了消除部分条件独立性假设对分类造成的不利影响,可通过在朴素贝叶斯公式中加入属性权重值以区分不同特征项对分类的贡献度。

IMI 值可以作为属性权重应用于贝叶斯分类中，当 IMI 值计算结果较大时，特征项与类别的相关性较高，当 IMI 值较低甚至为负值时，表示该特征项对分类的作用较小。互信息值可以在一定程度上表示特征项与类别之间的相关性，消除部分条件独立性假设对分类的不利影响。将式(13)中的后验概率赋予互信息权值可得：

$$C_{\text{map}} = \underset{C_j \in C}{\arg\max}\, P(C_j) \prod_{i=1}^{n} P(\boldsymbol{w}_i \mid C_j)^{\text{IMI}(\boldsymbol{w}_i, C_j)} \tag{15}$$

特征项 $\text{IMI}(\boldsymbol{w}_i, C_j) = \alpha_{ij} \times \sigma_{ij} \times \text{lb}\dfrac{P(\boldsymbol{w}_i \mid C_j)}{P(\boldsymbol{w}_i)}$ 对类别 $\text{IMI}(\boldsymbol{w}_i, C_j) = \alpha_{ij} \times \sigma_{ij} \times \text{lb}\dfrac{P(\boldsymbol{w}_i \mid C_j)}{P(\boldsymbol{w}_i)}$ 的权重属性值计算方法为：

$$\text{IMI}(\boldsymbol{w}_i, C_j) = \alpha_{ij} \times \sigma_{ij} \times \text{lb}\frac{P(\boldsymbol{w}_i \mid C_j)}{P(\boldsymbol{w}_i)} \tag{16}$$

将属性权值代入上式并取对数可得：

$$C_{\text{map}} = \underset{C_j \in C}{\arg\max}\left[\text{lb}\, P(C_j) + \sum_{i=1}^{n}\left(\text{lb}\, P(\boldsymbol{w}_i \mid C_j) \times \alpha_{ij} \times \sigma_{ij} \times \text{lb}\frac{P(\boldsymbol{w}_i \mid C_j)}{P(\boldsymbol{w}_i)}\right)\right] \tag{17}$$

为避免出现概率为 0 的情况，本文对互信息公式中的 $P(\boldsymbol{w}_i)$ 和 $P(\boldsymbol{w}_i \mid C_j)$ 进行拉普拉斯平滑处理，具体如下式所示：

$$P(\boldsymbol{w}_i) = \frac{\text{df}(\boldsymbol{w}_i) + 1}{\text{df}_{\text{total}} + 2} \tag{18}$$

$$P(\boldsymbol{w}_i \mid C_j) = \frac{\text{df}_{C_j}(\boldsymbol{w}_i) + 1}{\text{df}_{C_j} + 2} \tag{19}$$

其中，$\text{df}(\boldsymbol{w}_i)$ 表示特征项 w_i 在整个训练集中的文本频率，df_{total} 表示整个训练集的文本频率，$\text{df}_{C_j}(\boldsymbol{w}_i)$ 表示特征项 $\boldsymbol{w}_i$ 在类 C_j 训练集中的文本频率，df_{C_j} 表示类 C_j 训练集中的文本频率。

IMI-WNB 算法的实现过程如下所示：

算法 2 IMI-WNB 算法

输入 特征子集 $\boldsymbol{F} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_k\}$，邮件样本集 $D = \{d_1, d_2, \cdots, d_l\}$

输出 样本集各样本所属类别 C

1. 计算 $P(C_{ham})$ 和 $P(C_{spam})$
2. for i = 1 to k
3. 统计词频 $tf_{C_{spam}}(w_i)$ 和 $tf_{C_{ham}}(w_i)$
4. 统计文档频率 $df_{C_{spam}}(w_i)$ 和 $df_{C_{ham}}(w_i)$
5. 式(19)计算 $P(w_i \mid C_{spam})$ 和 $P(w_i \mid C_{ham})$
6. 式(18)计算 $P(w_i)$
7. 式(1)计算互信息值 $MI(w_i \mid C_{spam})$ 与 $MI(w_i \mid C_{ham})$
8. 式(3)计算词频因子 α_{ij}
9. 式(7)计算类间差异因子 σ_{ij}
10. 将式(1)、式(3)与式(7)结果代入式(16)计算得到 $IMI(w_i, C_{spam})$ 与 $IMI(w_i, C_{ham})$ 值
11. end
12. for i = 1 to l
13. for each w_i in d_i
14. 计算 $P(w_i, C_{spam})$ 和 $P(w_i, C_{ham})$
15. 将 $IMI(w_i, C_{spam})$、$IMI(w_i, C_{ham})$、$P(C_{ham})$、$P(C_{spam})$ 与 $P(w_i \mid C_{spam})$ 代入式(17)中进行计算
16. end
17. $C_{(d,map)} = \max\{C_{(d,ham)}, C_{(d,spam)}\}$ //判别类型
18. end

算法 2 是垃圾邮件过滤分类阶段算法，在通过特征选择阶段算法 1 获得特征子集后，算法 2 将 IMI 值作为属性权重值应用于朴素贝叶斯分类中，体现出不同特征项对分类决策贡献的差异，消除部分朴素贝叶斯条件独立性假设对分类的不利影响，从而提高分类精度。

2.3 IMI-WNB 算法的垃圾邮件过滤流程

IMI-WNB 算法的垃圾邮件过滤流程如图 1 所示。首先，在邮件预处理阶段对文本进行去停用词处理，然后再对文本进行分词，采用 Python 中文分词组件 jieba 对文本进行自动分词。其次，在特征选择阶段使用本文所提 IMI 算法对文本中的特征项进行特征选择。通过 IMI 算法可以将对分类无关的特征项筛选出去。最后，在训练阶段统计出样本中的先验概率与条件概率，并在应用阶段使用 IMI-WNB 分类器分类时代入计算，根据计算出的最大后验概率对邮件文本进行判定，当垃圾邮件概率大于非垃圾邮件概率时，分类器判定该邮件文本为垃圾邮件。

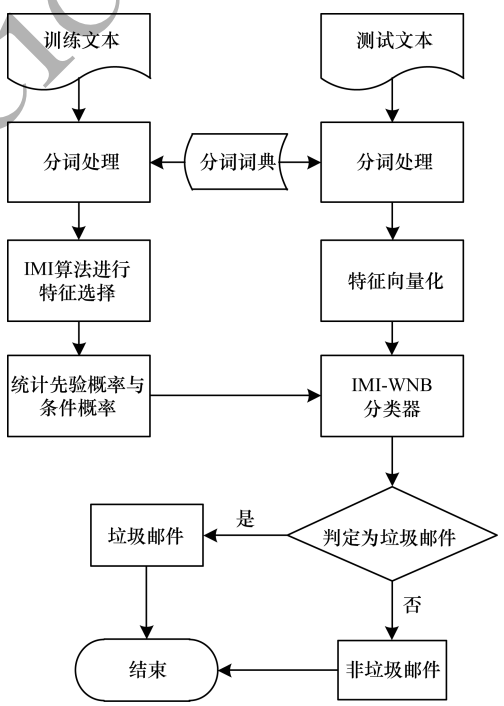

图 1 IMI-WNB 算法的垃圾邮件过滤流程

Fig. 1 Spam filtering procedure of IMI-WNB algorithm

3　实验结果与分析

本文使用 trec06c 邮件语料库对垃圾邮件进行过滤实验,并对 IMI-WNB 算法与传统的 NB 算法进行过滤效果对比。同时,为了更充分地体现本文算法的过滤效果与现实意义,实验将本文算法与其他改进算法进行过滤效果对比。

3.1　实验环境与数据

实验在 Windows 10 下进行,硬件配置为i5-7300HQ 2.50 GHz CPU,内存 8.00 GB,硬盘 500 GB。采用 Python 3.7 为实验环境。实验数据来自公开的垃圾邮件语料库 trec06c,从中随机抽取 14 000 封邮件作为样本集,其中 7 000 封为正常邮件,7 000 封为垃圾邮件。

3.2　评价标准

为了对垃圾邮件过滤效果进行评价,实验引入精确度 P、召回率 R 和 F 值 3 个评价指标。假设垃圾邮件被判定为垃圾邮件的总数为 T_{spam},垃圾邮件被判定为正常邮件的总数为 F_{spam},正常邮件被判定为垃圾邮件的总数为 F_{ham},则 3 个评价指标的计算方法分别为:

$$P=\frac{T_{\text{spam}}}{T_{\text{spam}}+F_{\text{ham}}} \tag{20}$$

$$R=\frac{T_{\text{spam}}}{T_{\text{spam}}+F_{\text{spam}}} \tag{21}$$

$$\text{F}=\frac{2\times P\times R}{P+R} \tag{22}$$

其中,精确度代表了垃圾邮件的检对率,正常邮件被误判为垃圾邮件会降低精确度,召回率代表了垃圾邮件的检出率,召回率低说明有大量垃圾邮件被漏检,F 值为综合精确度和召回率的评价标准,其表示邮件过滤的综合效果。

3.3　实验结果

本文实验过程步骤如下:

步骤 1　对 trec06c 语料库中选取的邮件样本进行分词处理,建立停用词表去除文本中的停用词,并对文本进行特征选择。使用 MI 算法得到的特征项集为 $\boldsymbol{T}_{\text{MI}}$,使用 IMI 算法得到的特征项集为 $\boldsymbol{T}_{\text{IMI}}$。

步骤 2　分别从特征项集 $\boldsymbol{T}_{\text{MI}}$ 与 $\boldsymbol{T}_{\text{IMI}}$ 中提取 n 个特征项 $t_1,t_2,\cdots,t_n$ 组建特征向量空间 $\boldsymbol{R}_{\text{MI}}$ 与 $\boldsymbol{R}_{\text{IMI}}$,在特征向量空间中分别利用 NB 算法与 IMI-WNB 算法进行分类。

步骤 3　将 14 000 封邮件样本平均分为 10 份,采用十折交叉法对样本进行计算,即每次选取其中9 份样本作为训练集,1 份样本作为测试集进行分类实验,每个样本均有一次作为训练集进行测试,每个维度总共进行 10 次测试,最后计算 10 次实验平均值作为该维度的数据结果。实验选取向量空间维度 n 从 10 到 700 对邮件进行分类,取平均值后绘制折线图。精确度、召回率、F 值的实验结果如图 2 ~ 图 4 所示。

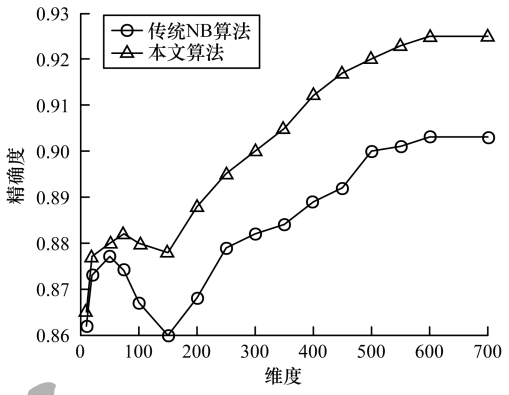

图 2　2 种算法的精确度实验结果

Fig. 2　Experimental results of precision of two algorithms

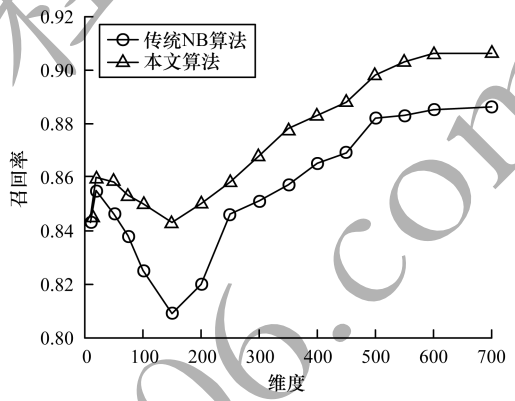

图 3　2 种算法的召回率实验结果

Fig. 3　Experimental results of recall rate of two algorithms

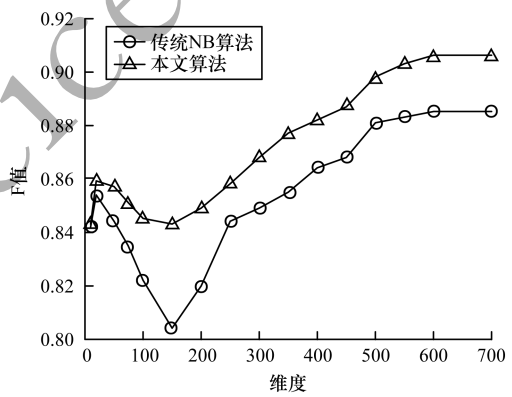

图 4　2 种算法的 F 值实验结果

Fig. 4　Experimental results of F-value of two algorithms

从图 2 可以看出:传统 NB 算法的精确度随着特征项维度的增大呈现先增大后降低再增大的趋势,且当维度为 150 时精确度最低,当特征维度达到 500 后,精确度逐渐趋于平缓;然而本文算法在低维度时的精确度与传统 NB 算法相差不大,当特征项的维度达到 75 后,其精确度开始下降,且在维度为 150 时达到最低,但其精确度受到的影响明显小于传统 NB 算法,且本文算法的精确度整体上高于传统 NB 算法。

从图 3 可以看出:传统 NB 算法的召回率在特征维度达到 20 后开始下降,当特征维度达到 150 时召

回率降至最低,接下来随着特征项维度的增加召回率逐渐增加,当维度达到 500 时召回率趋于稳定;本文算法的召回率在特征维度为 20 时开始下降,但下降速度相对传统 NB 算法更加趋缓,且整体召回率高于传统 NB 算法。类似地,从图 4 可以看出,相比传统 NB 算法,本文算法的 F 值有明显提高,且波动更加平缓。

在使用 trec06c 语料库作为邮件样本进行邮件过滤时,本文算法与 PTw2v 算法[18]、C4.5 算法[19]、GWO_GA 算法[20]的性能对比如表 1 所示。从表 1 可以看出:PTw2v 算法的精确度与召回率相差不大,且有较好的分类效果;本文算法相较 C4.5 算法召回率更高,说明 C4.5 算法存在较多的垃圾邮件被漏检,本文算法在 F 值上也高于该算法,说明本文算法具有更好的分类效果;GWO_GA 算法的召回率较高,但在精确度上远低于本文算法,说明该算法存在大量的正常邮件被误判为垃圾邮件,且该算法的 F 值也略低于本文算法。

表 1 4 种算法的性能对比

Table 1 Performance comparison of four algorithms

算法	精确度	召回率	F 值
PTw2v 算法	0.957	0.960	0.969
C4.5 算法	0.914	0.869	0.891
GWO_GA 算法	0.828	0.979	0.897
本文算法	0.925	0.906	0.908

综合分析上述实验结果可知,相比传统 NB 算法,本文算法的精确度、召回率与 F 值明显提高,且变化趋势更加稳定。

4 结束语

由于传统互信息算法在特征选择中对于词频以及类间频数差异考虑不足,本文提出一种改进的互信息算法,并针对特征项在文本中的词频数以及类间频数差异对分类的影响进行分析与改进,有效利用训练集中的频数信息,弥补了传统互信息算法仅考虑到文本频率的缺陷。同时,本文对朴素贝叶斯算法进行属性加权并提出一种 IMI-WNB 算法,部分消除了朴素贝叶斯算法独立性假设对分类的不利影响。仿真实验结果表明,该算法明显提高了邮件分类的精确度、召回率、F 值及稳定性,且取得良好的过滤效果。本文的邮件过滤技术是基于邮件的文本内容进行分类,然而除了邮件文本内容外,邮件还有题目、发送时间、收件人与发件人等邮件头信息可供分类判定。因此,下一步将利用加权朴素贝叶斯算法对邮件的文本内容与邮件头信息进行综合分析,以提高邮件过滤分类效果。

参考文献

[1] MARIA V, TATYANA S, TATYANA S. Kaspersky 2018 annual report on spam and phishing attacks[EB/OL]. [2019-08-01]. https://securelist.com/spam-and-phishing-in-2018/89701/.

[2] DELANY S J, BUCKLEY M, GREENE D. SMS spam filtering: methods and data[J]. Expert Systems with Applications, 2012, 39(10): 9899-9908.

[3] LIU Peng, ZHAO Huihan, TENG Jiayu, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark[J]. Journal of Central South University, 2019, 26(1): 1-12.

[4] BENNASAR M, HICKS Y, SETCHI R. Feature selection using joint mutual information maximisation[J]. Expert Systems with Applications, 2015, 42(22): 8520-8532.

[5] YANG Y M, PEDERSEN J O. A comparative study on feature selection in text categorization[EB/OL]. [2019-08-01]. https://www.ixueshu.com/document/ab514a35b74a74b2318947a18e7f9386.html.

[6] LIANG Ting. Research on content based spam filtering technology[D]. Shanghai: East China Normal University, 2013. (in Chinese)
梁婷. 基于内容的垃圾邮件过滤技术研究[D]. 上海: 华东师范大学, 2013.

[7] WANG J J Y, WANG Y, ZHAO S G, et al. Maximum mutual information regularized classification[J]. Engineering Applications of Artificial Intelligence, 2015, 37: 1-8.

[8] WANG Wenmin, ZHOU Dan. A multi-level approach to highly efficient recognition of Chinese spam short messages[J]. Frontiers of Computer Science, 2018, 12(1): 135-145.

[9] MADISETTY S, DESARKAR M S. A neural network-based ensemble approach for spam detection in twitter[J]. IEEE Transactions on Computational Social Systems, 2018, 5(4): 973-984.

[10] LIU Haifeng, YAO Zeqing, SU Zhan. Optimization mutual information text feature selection method based on word frequency[J]. Computer Engineering, 2014, 40(7): 179-182. (in Chinese)
刘海峰, 姚泽清, 苏展. 基于词频的优化互信息文本特征选择方法[J]. 计算机工程, 2014, 40(7): 179-182.

[11] MISHRA R, THAKUR R S. Analysis of random forest and naïve Bayes for spam mail using feature selection catagorization[J]. International Journal of Computer Applications, 2013, 80(3): 42-47.

[12] ELSSIED N O F, IBRAHIM O, OSMAN A H. Enhancement of spam detection mechanism based on hybrid\varvec {k}-mean clustering and support vector machine[J]. Soft Computing, 2015, 19(11): 3237-3248.

[13] WU Jianjun, LI Changbing. Mutual information-based weighted naive Bayes text classification algorithm[J]. Computer Systems & Applications, 2017, 26(7): 178-182. (in Chinese)
武建军, 李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法[J]. 计算机系统应用, 2017, 26(7): 178-182.

(下转第 312 页)

(上接第 304 页)

[14] 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. (in Chinese)
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.

[15] TAO Yongcai, ZHAO Guohua, SHI Lei, et al. Improved MapReduce mutual information text feature selection mechanism[J]. Journal of Chinese Computer Systems, 2018, 39(3): 433-438. (in Chinese)
陶永才,赵国桦,石磊,等. 一种改进的 MapReduce 互信息文本特征选择机制[J]. 小型微型计算机系统,2018,39(3):433-438.

[16] LI Zhao, LU Wei, SUN Zhanquan, et al. A parallel feature selection method study for text classification [J]. Neural Computing and Applications, 2017, 28(S1): 513-524.

[17] EL-ALFY E S M, ALHASAN A A. Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm [J]. Future Generation Computer Systems, 2016, 64: 98-107.

[18] TANG Xianlun, WAN Yali, LIU Yuwei, et al. Chinese spam classification based on weighted distributed characteristic[C]// Proceedings of 2017 Chinese Automation Congress. Washington D.C., USA: IEEE Press, 2017: 6618-6622.

[19] HU Wei, DU Jinglong, XING Yongkang. Spam filtering by semantics-based text classification[C]//Proceedings of the 8th International Conference on Advanced Computational Intelligence. Washington D.C., USA: IEEE Press, 2016: 89-94.

[20] LIU Haoran, DING Pan, GUO Changjiang, et al. Study on Chinese spam filtering system based on Bayes algorithm [J]. Journal on Communications, 2018, 39(12): 151-159. (in Chinese)
刘浩然,丁攀,郭长江,等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究[J]. 通信学报,2018,39(12):151-159.

编辑　刘继娟