基于关联规则的垃圾邮件分类模型

邓 慧
(北医学院 四川 南充 637000)

摘 要 为了提高垃圾邮件分类精确,提出一种基于关联规则的垃圾邮件分类模型。首先通过改进的 FP-grow 算法挖掘垃圾邮件关联规则集,以关联规则集为基础构建垃圾邮件分类器模型,然后考虑垃圾邮件特征词权重对邮件进行分类,最后采用仿真实验测试模型的性能。结果表明,该方法提高了垃圾邮件分类精度,可以较好地对垃圾邮件进行分类。

关键词 邮件分类 关联规则 垃圾邮件变异 特征提取

中图分类号 TP391.3 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2015.08.075

SPAM EMAIL CLASSIFICATION MODEL BASED ON ASSOCIATION RULES

Deng Hui
(*North Medical University, Nanchong 637000, Sichuan, China*)

Abstract In order to improve spam email classification precision, we proposed a novel association rules-based spam email classification model. First, we used improved FP-grow algorithm to mine the association rules set of span emails, and built the spam email classifier model based on association rules set; then we classified the span emails by considering their feature words weights. Finally, we carried out simulation experiments to test the performance of the model. Results showed that the proposed model improved the classification precision on spam emails, and could better classify the spam emails.

Keywords Email classification Association rules Spam email variation Feature extraction

0 引 言

随着互联网技术迅速发展,网民越来越多,电子邮件成为网民通信的一种工具,为人们生活、交流、通信提供了方便,然而,由于网络的不设防性,垃圾邮件随之产生[1]。垃圾邮件就是一种不请自来、携带有不良信息的邮件,主要包括中奖诈骗、冒充银行扣款、违法出售票证等,扰乱人们的正常生活,因此提高垃圾邮件分类准确性具有十分重要的意义[2]。

为了应对大量垃圾邮件出现,国内许多学者对其进行了大量、广泛的研究,提出了许多有效的垃圾邮件分类模型[3]。垃圾邮件分类实质上一个邮件分类过程,因此是一种模式识别中的二分类问题,就是将邮件分为合法和垃圾邮件两类,因此如何提取垃圾邮件分类特征和设计邮件分类器至关重要。邮件是一种特殊文本,邮件特征维数相当高,当前邮件特征提取方法主要有主成分分析、非负矩阵分解、线性判别、核主分分析[3-5]。邮件分类器的构建主要有支持向量机、逻辑回归、神经网络、贝叶斯网络方法等[6]。神经网络基于“大数定理”进行建模,当样本数量有限,泛化能力比较差,易出现“过拟合”缺陷;逻辑回归是一种线性建模方法,无法正确描述垃圾邮件类别与特征之间的非线性、动态变化关系,分类结果不太理想;贝叶斯网络算法可以获得较好的邮件分类精度,但是该建模复杂、计算时间复杂度高;支持向量机算法具有较优的泛化能力,相对于其他算法,垃圾分类的精度相当高,但是由于每一个样本要与分类平面进行比较,训练时间相当长,尤其对于大规模垃圾邮件分类问题,效率相当低,应用范围受限[7]。近年来,随着数据挖掘技术发展,一些学者提出基于关联规则的文本分类算法,研究结果表明,关联规则可以获得较理想的文本分类结果,因此为垃圾邮件提供了一种新的分类工具[8,9]。

为了提高垃圾邮件分类精度,准确分类掉垃圾邮件,保证网络安全,提出一种基于关联规则的垃圾邮件分类模型。仿真结果表明,本文模型具有垃圾邮件分类精度高,错分率低等优点,是一种有效的垃圾邮件分类模型。

1 相关定义

设垃圾邮件的空间为: $S=\{<sm_1,c_1>,<sm_2,c_2>,\cdots,<sm_n,c_n>\}$, $<sm_i,c_i>$, n 表示包括的总邮件数, $<sm_i,c_i>$ 表示一条样本, C 为垃圾邮件, $\bar{C}$ 为正常邮件[10]。

定义 1 所有邮件特征词集合采用 $T=\{w_1,w_2,\cdots,w_k\}$ 表示 $W(sm)$ 为邮件 sm 经过分词后特征的特证词集合,如果 $T\subseteq W(sm)$,那就可以认为 sm 蕴含模式 T,即 $T\subseteq sm$,就可以采用如下函数关系表示:

收稿日期: 2014-01-13。邓慧,讲师,主研领域: 计算机网络,数据挖掘。

$$t(T,sm) = \begin{cases} 1 & T \subseteq sm \\ 0 & \exists w_i \in T, w_i \notin W(sm) \end{cases} \quad (1)$$

定义 2　关联规则 r 支持数是指匹配规则 r 的垃圾邮件数量，可以采用 $s(r)$ 描述。

定义 3　T 的发生数是指蕴含模式 T 的样本邮件数量，可以采用如下形式描述:

$$o(T) = \sum_{sm \in S} t(T,sm) \quad (2)$$

定义 4　垃圾邮件关联规则 r 支持度可以表示为:

$$d(r) = \frac{s(r)}{|S|} \quad (3)$$

定义 5　给定了阈值 θ，如果规则 r 的支持度大于等于 θ，则称 r 为频繁关联规则。

定义 6　垃圾邮件关联规则 r 置信度定义为:

$$\sigma(r) = \frac{s(r)}{o(r.T)} \quad (4)$$

定义 7　对规则 r_i 和 r_j，如果 $r_i.T \subset r_j.T$，那么规则 r_i 是规则 r_j 的子规则，r_j 是 r_i 的超规则。

定义 8　特征词描述垃圾邮件类别的能力可以采用特征词权重表示，$weight(w_i)$ 表示 w_i 的权重值。

2　垃圾邮件分类模型

2.1　FP-Growth 算法

2000 年，韩家炜提出了 FP-Growth(频繁模式增长)算法，首先将提供频繁项集的数据库压缩到一棵频繁模式树(FP-Tree)，但仍保留项集关联信息; 与 Apriori 算不同: 首先其不产生候选集，第二只需要两次遍历数据库，因此提高了算法的效率。

1) FP-树构造步骤

(a) 扫描事务数据库 D 一次。收集频繁项的集合 F 和它们的支持度，对 F 按支持度降序排序，结果为频繁项表 L。

(b) 创建 FP-树的根结点，以“null”标记它，对于 D 中每个事务 Trans，执行: 选择 Trans 中的频繁项，并按 L 中的次序排序。设排序后的频繁项表为 $[p|P]$，其中，p 是第一个元素，而 P 是剩余元素的表。调用 $insert\_tree([p|P],T)$，该过程执行情况如下:

如果 T 有子女 N 使得 $N.item\text{-}name = p.item\text{-}name$，则 N 的计数增加 1;

否则创建一个新结点 N，将其计数设置为 1，链接到它的父结点 T，并且通过结点链结构将其链接到具有相同 $item\text{-}name$ 的结点，如果 P 非空，递归地调用 $insert\_tree(P,N)$。

2) FP-树的挖掘过程伪代码

```
if Tree 含单个路径 P then
    for 路径 P 中结点的每个组合(记作 β)
        产生模式 β ∪ α，其支持度 support = β 中结点的最小支持度;
else
    for each ai 在 Tree 的头部{
        产生一个模式 β = ai ∪ α，其支持度 support = ai.support;
        构造 β 的条件模式基，然后构造 β 的条件 FP-树 Treeβ;
    if Treeβ ≠ ∅ then
调用 FP_growth(Treeβ,); }
```

2.2　提取特征词及计算权重

在垃圾邮件检测过程，一些常用高频词会对关联规则的挖掘进行不利影响，为了防止该种现象的出现，本文首先根据文献[11]中改进的 χ^2 方法计算垃圾邮件特征词权重，然后从中选择一些权重比较大的特征词作为垃圾邮件的特征词集合。

w_i 的权重计算公式为:

$$weight(w_i) = \rho \frac{[P(w_i,C)P(\overline{w_i},\overline{C}) - P(w_i,\overline{C})P(\overline{w_i},C)}{P(w_i)P(\overline{w_i})P(C)P(\overline{C})} \quad (5)$$

其中:

$$\rho = \frac{|w_i|}{2} \quad (6)$$

式中，$|w_i|$ 表示词 w_i 的字长; $\overline{w_i}$ 表示不含 w_i 的邮件; $P(\overline{w_i},\overline{C})$ 为 C 中不包 w_i 邮件的概率;

式(5)和式(6)综合考虑了词频、词在样本空间的分布情况和词的字长，字长越大的词反映垃圾邮件主题的能力越强，因此权重较高，定义阈值为:

$$\eta = (1-k)\min weight(w_i,c) + k\max weight(w_i,c) \quad (7)$$

式中，k 为特征提取系数。

2.3　FP-tree 的构造

针对垃圾邮件分类问题，本文采用改进的 FP-growth 算法来进行关联规则的挖掘。挖掘的过程分为两个阶段: 构建 FP-tree 和挖掘关联规则。FP-tree 是一种由一个头表和一个树所组成的数据结构，如图 1 所示。

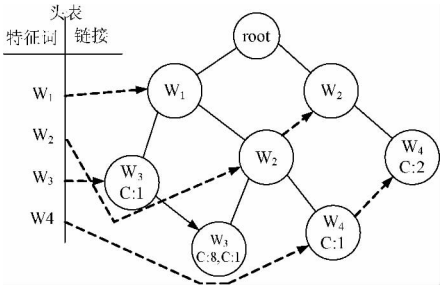

图 1　FP-tree 结构

2.4　考虑变异行为的 FP-tree

随机选择一些垃圾邮件进行分析，结果表明，有部分邮件已经发生了变异行为，因此在垃圾邮件分类建模过程中，需要考虑垃圾变异的变异行为，才能准确对垃圾进行分类。3 种常见变异行为如表 1 所示。从表 1 可知，在垃圾邮件分词过程中，需要将变异词替换为正常词，如: 代ˆ办”替换为“代办”; 发嫖”替换为“发票”等。

表 1　垃圾邮件常见变异行为

变异行为	示例
同音字变异	堤供发建档嫖……
敏感词分隔	代办张三丰田购买ˆ……
繁体字变异	……垃圾郵件檢測……

由于垃圾邮件会发生变异行为，而上述的 FP-tree 忽略了垃圾邮件变异行为，与实际垃圾不相符，存在一定的缺陷，为此解决该不足，本文将垃圾邮件变异行为作为一种特殊特征，采用 $\{a_i\}$ 表示垃圾邮件特殊特征词，i 为异常行为的类型，那么改进 FP-tree 如图 2 所示。

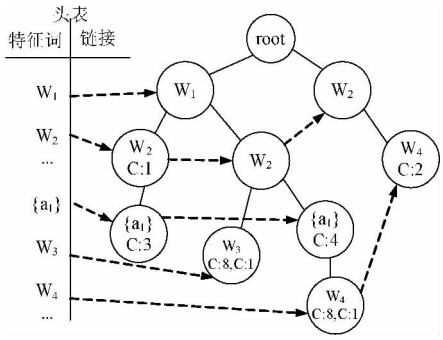

图 2 改进 FP-tree

2.5 关联规则

1) 关联规则挖掘

输入: 支持度阈值 θ,置信度阈值 β,FP-tree 实例 *tree*

输出: 垃圾邮件关联规则集 R

```
R =
Do While( count ( tree. HeadTable) >1) Begin
lastItem = ExtractLastTableItem ( tree)
Paths =
For each treeNode in lastItem. links do Begin
path = PathToRoot ( treeNode)
If( hasClassLabel( Path,C) Begin
Paths = Paths{ path}
End
MergeClassCountToParent ( treeNode)
RemoveFromTree ( treeNode)
End
R = RGenerateRules ( Paths,lastItem,θ,β)
End
```

2) 产生关联规则

输入: 一个路径集合 *Paths*,集合中的每一个路径都以头表项 *lastItem* 中的特征词 w_k 为终结点; 支持度阈值 θ,置信度阈值 β

输出: 含有特征词 w_k 最终关联规则集 $R(w_k)$

```
R( wk) =
CAR1 = EveryWordIn ( Paths)
For i =1 to 3 do Begin
Fi =
For each pattern_j in CAR_i do Begin
sumc = countC ( pattern_j,Paths)
sumnc = countNC ( pattern_j,Paths)
s( pattern_j) = sumc/ | S |
If s( pattern_j) ≥ θ then
F_i = F_i{ pattern_j}
( patternj) = sumc/( sumc + sumnc)
If ( patternj) ≥ β then
R( w_k) = R( w_k) { pattern_j w_k = > C}
End
End
End
CAR_{i+1} = F_i F_i
End
```

算法的输出 $R(w_k)$ 每一个规则都包含有特征词 w_k,因此在产生 $R(w_k)$ 中 i 阶关联规则时,只需要考虑另外的 $i-1$ 个特征词,将这 $i-1$ 个特征词和 w_k 组合即形成了 i 阶规则。

2.6 垃圾邮件分类过程

1) 如果 $|r_i| > |r_j|$,那么就表示 r_i 的优先级高于 r_j,可以采用 $r_i < r_j$ 表示,不然,需要比较规则的置信度,若 $\sigma(r_i) > \sigma(r_j)$,则 $r_i < r_j$。

2) 当 r_i、r_j 的阶数和置信度均相同,那么就规则的前件模式权重和进行比较,如果 $\sum_{w \in r_i.T} weight(w) > \sum_{w \in r_j.T} weight(w)$,那么表示 $r_i < r_j$。

首先根据关联规则优先级对规则 R 中所有的关联规则进行行排序,组成了一个关联规则优先级队列,然后采用规则对待检测邮件进行匹配,若匹配成功,那么就可以表示该是一种垃圾邮件,否则是正常邮件。

3 仿真实验

3.1 仿真环境

在迅驰酷睿 2 处理器、2.53 GHz 主频、4 GB RAM,Windows 7 操作系统,QL Server 2008 数据库,采用 VC ++ 进行编程实现仿真实验。

3.2 评价指标

为了能够对垃圾邮件分类模型的分类效果进行很好的评价,实验采用下列常见的性能评价指标: 查准率(Precision) 、查全率(Recall) [13] 。为了方便地对它们进行定义,假设待测试的邮件总数为 n,将它们用于构建好了的垃圾邮件分类模型,可得邮件判定结果如表 1 所示。

表 2 垃圾邮件分类模型的判定情况分布

$n = a + b + c + d$	实际为垃圾邮件	实际为正常邮件
判定为垃圾邮件	a	b
判定为正常邮件	c	d

那么查准率和的查全率计算公式分别为:

$$P = \frac{a}{a+b} \tag{8}$$

$$R = \frac{a}{a+c} \tag{9}$$

查准率体现了模型检对垃圾邮件的能力,即查准率越大,正常邮件被误判为垃圾邮件的概率越小; 查全率体现了模型识别垃圾邮件的能力,即查全率越大,被误判的垃圾邮件越少。可见,查准率 P 或查全率 R 越大,分类模型就越好。

然而,在某些模型中它们之间会相互影响(即一个大,而另一个小) ,因此实验将把 $F1$ 作为主要的性能评价指标。$F1$ 是查准率 P 和查全率 R 的调和平均,是它们的综合体现,公式为:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{10}$$

3.3 阈值和置信度阈值的确定

阈值 θ、置信度阈值 β 对邮件分类结果具有重要的影响,选择 30 000 条邮件,其中垃圾邮件 10 000 条,随机选择 5000 条邮件作为测试集,对阈值选择进行了仿真实验。阈值 θ、置信度阈值 β 的取值见表 3 所示。

表 3　阈值和置信度阈值组合方式

θ(%)	β(%)
0.25	80
0.30	82
0.35	84
0.40	86
0.45	88
0.50	90

采用错误损失率对分类效果进行评价,其定义如下:

$$\eta = \frac{fc + 2 \times fnc}{s} \times 100\% \quad (11)$$

式中,fc 表示垃圾邮件被错分为正常邮件数,fnc 表示正常邮件被错分为垃圾邮件的数。

实验结果表明,结果表明 $\theta = 0.30$、$\beta = 86$ 时,损失率 η 取最小值 12.25,那么后续实验采用 $\theta = 0.30$、$\beta = 86$ 进行。

3.4　实验结果

数据集来源于 UCI 机器学习数据库中的垃圾邮件数据库[13],包含 4601 个实例,每个实例分别由 58 个特征属性描述,最后一列表示邮件被认定为垃圾邮件(用 1 表示)或非垃圾邮件(用 0 表示),选择 4000 条邮件测试集进行实验,共采用了 4 种方法,将支持向量机、贝叶斯分类、决策树、本文模型。4 种模型的实验结果如表 4 所示,从表 4 可以看出,本文模型在两个指标上均优于其他垃圾分类模型,对比结果表明,本文模型是一种有效的、分类精度高、误分率较好的垃圾邮件分类模型,可以较好地满足垃圾邮件分类要求。

表 4　不同模型的 UCI 数据库垃圾邮件分类性能对比

模型	R%	P%	$F1$
支持向量机	87.87	82.65	85.18
贝叶斯分类	86.56	81.92	84.18
决策树	85.03	77.28	80.97
本文模型	95.80	90.95	93.31

为了进一步测试本文算法的有效性,邮件来自于中国教育和科研计算机网紧急响应组(Data Sets of Chinese Emails, CCERT 2005-Jun)[14],该样本集包含合法邮件 9272 封,垃圾邮件 25 088 封,从样本集中选择 1000 封合法邮件和 2700 封垃圾邮件作为邮件训练样本集,其中 100 封合法邮件和 270 封垃圾邮件用于 IWB 训练,其他用于测试。4 种模型的实验结果如表 5 所示,从表 5 可以知道,相对于对比模型,本文模型的垃圾分类性能更优,对比结果表明再一次证明了本文模型的有效性和优越性。

表 5　不同模型对 CCERT 数据库的垃圾邮件分类性能对比

模型	R%	P%	F1
支持向量机	89.34	87.39	88.35
贝叶斯分类	83.89	82.13	83.00
决策树	80.55	79.41	79.98
本文模型	95.17	94.51	94.84

4　结　语

垃圾邮件是一种高维、复杂的特殊文本,具有维数、规模大等特点,针对当前提出的垃圾邮件分类模型的不足,提出一种基于关联规则的垃圾邮件分类模型。实验结果表明,本文模型有效提高了垃圾分类正确率,降低漏报率和误报率,从而显著地提高对垃圾邮件的分类能力,具有良好的实用价值。

参　考　文　献

[1] 王斌,潘文峰. 基于内容的垃圾邮件分类技术综述[J]. 中文信息学报,2005,19(5):1-10.

[2] 马莉,柴乔林. 基于 Postfix 的垃圾邮件分类技术的实现[J]. 计算机工程与设计,2005,26(4):999-1001.

[3] 程卫华,尤晋元. 基于内容分类的反垃圾邮件系统的设计与实现[J]. 安徽大学学报,2007,31(3):30-33.

[4] 郑炜,沈文,张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件分类器的研究[J]. 西北工业大学学报,2010,28(4):622-627.

[5] Youn S,Mcleod D. A Comparative Study for Email Classification[J]. Advances and Innovations in Systems,Computing Sciences and Software Engineering,2007,12(9):387-391.

[6] Janecek A,Gansterer W. E-Mail Classification Based on NMF[C]. Applied Mathematics Sparks,Nevada,USA,2009,13:1345-1354.

[7] Liu Z,Zhang X,Zheng S. Lasso-based Spam Filtering with Chinese Emails[J]. Journal of Computational Information Systems,2012,8(8):3315-3322.

[8] Jiang L,Zhang H,Cai Z. A Novel Bayes Model Hidden Naive Bayes[J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(10):1361-1371.

[9] Provost F J,Domingos P. Tree induction for probability based ranking[J]. Machine Learning,2003,52(3):199-215.

[10] Orhan U,Adem K,Comert O. Least Squares Approach to Locally Weighted Naive Bayes Method[J]. Journal of New Results in Science,2012,12(5):71-80.

[11] Chang M,Yih W,Meek C. Partitioned Logistic Regression for Spam Filtering[C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining,2008,10:97-105.

[12] Amayri O,Bouguila N. A study of spam filtering using support vector machines[J]. Artificial Intelligence Review,2010,34(1):73-108.

[13] Mhamdi B,Grayaa K,Aguili T. Hybrid of particle swarm optimization, simulated annealing and tabu search for the reconstruction of two-dimensional targets from laboratory-controlled data[J]. Progress In Electromagnetic Research B,2011,28(1):1-18.

[14] Kennedy J,Mendes R. Population structure and particle swarm performance[C]. Honolulu,HI:2002,10:1671-1676.

(上接第 307 页)

[10] Chen G, Chen Y, Liao X. An extended method for obtaining S-boxes based on three-dimensional chaotic Baker maps[J]. Chaos Solitons Fractals,2007,31(35):571-577.

[11] Webster A F,Tavares S E. On the design of S-boxes[C]//Advances in Cryptology-Gypto'85,Lecture Notes in Computer Science 218. New York,Springer Verlag New York,Inc. 1985:523-534.

[12] Zhen W, Huang X, Li Y X, et al. A novel image encryption algorithm based on the fractional-order hyperchaotic Lorenz system[J]. Chinese Physics B,2013,22(1):010504-1-010504-7.