一种基于邮件头信息的三支决策邮件过滤方法

袁国鑫 于 洪
(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

摘 要 提出一种基于邮件头信息的三支决策垃圾邮件过滤方法。该方法使用一种新的属性重要度度量方法,并用该度量方法将邮件头信息属性依据重要度大小进行排序,然后按属性重要度的大小顺序对邮件计算贝叶斯概率并进行三支决策。当信息较少以致不足以决策时,按属性重要度大小顺序增加新的属性信息以帮助进一步的决策,直到得到最后的邮件分类。对比实验结果表明,该方法是合理且有效的。

关键词 邮件头信息,属性重要性,三支决策,垃圾邮件过滤
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.015

Method of Three-way Decision Spam Filtering Based on Head Information of E-mail

YUAN Guo-xin YU Hong
(Chongqing Key Lab of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract A method of three-way decision spam filtering was proposed in this paper based on the head information of E-mail. The head information is sorted by a new measurement of attribute significance. Bayesian probability based on the most significant attributes is computed to do the actions of three-way decisions. When the information is not enough to make decisions, more attribute information is added to the computing of Bayesian probability until the final decisions are made. The results of comparative experiments show that the new method is reasonable and effective.

Keywords Head information of E-mail, Attribute significance, Three-way decisions, Spam filtering

电子邮件是最广泛使用的互联网产品之一,在人们日常工作和生活中发挥着重要作用。与此同时,垃圾邮件的出现破坏了和谐的网络文明,浪费了互联网资源和网民的时间。2014年中国互联网协会组织、中国互联网协会反垃圾信息工作委员会和12321网络不良与垃圾信息举报受理中心联合发布的《2014年第一季度中国反垃圾邮件状况调查报告》[1]显示,中国网民平均每周接收到的邮件数量为35.0封,其中垃圾邮件数量为14.3封,垃圾邮件占比高达41.0%。垃圾邮件不仅耗费网络带宽和计算机时空开销,而且会对企业的正常运作和用户的正常工作造成严重的干扰,因此垃圾邮件的过滤问题受到广大学者的关注。

根据垃圾邮件过滤技术来划分,目前主要有基于规则的过滤和基于统计的过滤两类[2],这两类过滤技术多数是将垃圾邮件过滤问题视为二分类问题,即判断邮件是正常邮件还是垃圾邮件。这种观点没有反映出邮件真实的状态或特性,每个人由于喜好不同,关注点也不同,比如有人认为广告邮件都是垃圾邮件,而有人则对广告邮件并不反感,不将其视为垃圾邮件。因此一些学者提出将垃圾邮件过滤问题看作是三元分类问题并用三支决策理论[3-5]进行解决,即先划分出确定属于正常的邮件和确定属于垃圾的邮件,而将不能决策类别的邮件划分为待决策邮件,在得到更多信息后再进行进一步的决策,这样可以提高正确分类邮件的概率。结合三支决策,众多学者提出了基于三支决策的垃圾邮件过滤方法,如Li等[6]提出了一种基于三支决策的多阶段垃圾邮件过滤模型,模型中引入信息粒度的概念,使用不同的信息粒度对邮件进行表示,运用序贯决策,在不同的决策阶段基于不同的信息粒度分别进行三支决策;Zhou等[7]提出了一种代价敏感三支决策垃圾邮件过滤方法,较好地解决了邮件分类概率和所需阈值的问题;Jia等[8]研究了三支决策和二支决策在垃圾邮件过滤中的应用,并用实验证明了三支决策在处理垃圾邮件过滤问题时有更低的错误率和误分率;Deng和Hong[9]将邮件头信息和邮件内容信息相结合,提出了一种基于粗糙集的两阶段邮件过滤方法。

现有的大多数垃圾邮件过滤技术是基于邮件内容信息来进行分类的,而少部分基于邮件头信息的工作也没有重视邮件头信息中属性度量的重要性,比如如果邮件头信息中的属性"发件人"是好友,则可以立刻判断该邮件为正常邮件,那么可以推断出属性"发件人"的重要性应当是较高的。因此,本文在已有的序贯三支决策垃圾邮件过滤模型[6]的基础上提出一种新的基于概率统计理论的属性重要度度量方法来计算属性重要度的大小并排序,然后按属性重要度的大小顺序对邮件计算贝叶斯概率并进行三支决策分类,当信息较少以致不

到稿日期:2016-08-01 返修日期:2016-09-13 本文受国家自然科学基金(61379114,61533020)资助。
袁国鑫(1989-),男,硕士生,主要研究方向为数据挖掘、三支决策;于 洪(1972-),女,博士,教授,主要研究方向为粗糙集、三支决策、智能信息处理、Web智能、数据挖掘等,E-mail:yuhong@cqupt.edu.cn(通信作者)。

足以决策时,按属性重要度大小顺序增加新的属性信息以帮助进一步的决策,直到得到最后的邮件分类。

1　相关基本理论

1.1　三支决策垃圾邮件过滤模型

三支决策理论是 Yao[10-11] 近几年来在粗糙集和决策粗糙集的基础上提出来的。在很大程度上,粗糙集的正域、负域和边界域为三支决策提供了理论基础。简单来说,正域对应表示接受;负域对应表示拒绝;边界域对应表示不做决策或者延迟决策。三支决策的基本理念被广泛应用于日常生活中许多科研领域[12],众多学者也给出了许多研究成果[13-14]。

在三支决策垃圾邮件过滤模型中[6-7],给定一个四元邮件决策表 $S=(U,Atr=C\cup D,V,f)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为邮件集合。在 U 中的每个邮件对象都用一个特征向量 $x=\{v_1,v_2,\cdots,v_m\}$ 来表示,$v_i(1<i<m)$ 是对象 x 的第 i 个特征(属性)的取值;Atr 为邮件属性;$C=\{c_1,c_2,\cdots,c_m\}$ 为邮件条件属性,也称为邮件特征;$D=\{d_1\}$ 为邮件决策属性,即邮件所属的类别(正常邮件或垃圾邮件);V 为邮件属性 Atr 的值域;f 为信息函数。表 1 列出一个邮件的决策表示列。

表 1　一个邮件决策表示例

E-mail	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	d_1
x_1	0	0	1	0	0	1	0	0	0	0	1	0	1
x_2	1	0	1	0	1	1	0	1	1	0	1	0	1
x_3	1	1	0	1	0	2	1	1	0	0	1	0	1
x_4	0	1	1	0	1	0	0	1	0	0	0	1	0
x_5	1	2	0	1	0	1	1	0	1	0	0	1	0

根据文献[6],假设有状态空间 $\Omega=\{Y,\neg Y\}$ 表示正常邮件和垃圾邮件两种状态,即每个邮件对象有两种状态:正常邮件状态(用 $x\in Y$ 表示)和垃圾邮件状态(用 $x\in\neg Y$ 表示)。而对象 x 的状态通常用评价函数来判定,评价函数一般表示为 $v(Y|x)$。引入一对阈值 α 和 $\beta(0<\beta<\alpha<1)$,则三支决策的正域、负域和边界域的定义如下:

$$POS_{(\alpha,\beta)}(Y)=\{x\in U|v(Y|x)\geqslant\alpha\}$$

$$NEG_{(\alpha,\beta)}(Y)=\{x\in U|v(Y|x)\leqslant\beta\}$$

$$BND_{(\alpha,\beta)}(Y)=\{x\in U|\beta<v(Y|x)<\alpha\}$$

上式表示当 $v(Y|x)\geqslant\alpha$ 时,判定 $x\in Y$ 为正常邮件,划分到正域;当 $v(Y|x)\leqslant\beta$ 时,判定 $x\in\neg Y$ 为垃圾邮件,划分到负域;当 $\beta<v(Y|x)<\alpha$ 时,不能判定邮件类别,划分到边界域。

本文方法所用的评价函数为贝叶斯概率公式,即 $v(Y|x)=P(Y|x)$,$P(Y|x)$ 表示邮件 x 为状态 Y(正常邮件)的贝叶斯概率。

1.2　朴素贝叶斯分类

贝叶斯分类是基于贝叶斯公式的。假设有状态空间 $\Omega=\{Y,\neg Y\}$ 表示正常邮件和垃圾邮件两种状态,则对于邮件 x,贝叶斯公式表示如下:

$$P(Y|x)=P(Y)P(x|Y)/P(x)\tag{1}$$

其中,

$$P(x)=P(Y)P(x|Y)+P(\neg Y)P(x|\neg Y)\tag{2}$$

$P(Y|x)$ 为条件 x 下 Y 的后验概率,表示邮件 x 为正常邮件的概率;$P(Y)$ 为 Y 的先验概率,表示在邮件数据集中为正常邮件的概率;$P(x|Y)$ 为条件 Y 下 x 的概率。贝叶斯公式的主要思想是将难以估算的后验概率 $P(Y|x)$ 转换为容易估算的概率 $P(x|Y)$。在朴素贝叶斯中,假设各属性相对于类别条件独立,则有:

$$P(x|Y)=P(v_1,v_2,\cdots,v_m|Y)=\prod_{i=1}^{m}P(v_i|Y)\tag{3}$$

$$P(x|\neg Y)=P(v_1,v_2,\cdots,v_m|\neg Y)=\prod_{i=1}^{m}P(v_i|\neg Y)\tag{4}$$

将式(2)—式(4)代入式(1)即可求得后验概率 $P(Y|x)$,给定阈值 α 和 $\beta(0<\beta<\alpha<1)$,那么三支决策的正域、负域和边界域可表示为:

$$POS_{(\alpha,\beta)}(Y)=\{x\in U|P(Y|x)\geqslant\alpha\}$$

$$NEG_{(\alpha,\beta)}(Y)=\{x\in U|P(Y|x)\leqslant\beta\}$$

$$BND_{(\alpha,\beta)}(Y)=\{x\in U|\beta<P(Y|x)<\alpha\}$$

$P(Y|x)$ 表示邮件 x 属于正常邮件的概率,则邮件为正常邮件或垃圾邮件的三支决策语义如下:

如果 $P(Y|x)\geqslant\alpha$,则邮件是正常邮件;

如果 $P(Y|x)\leqslant\beta$,则邮件是垃圾邮件;

如果 $\beta<P(Y|x)<\alpha$,则邮件为可疑邮件,将其置于边界域中等待进一步处理。

2　基于邮件头信息的垃圾邮件过滤方法

2.1　属性重要度度量方法

属性重要度是粗糙集理论中的重要概念之一,被广泛应用于属性约简算法中。在粗糙集理论中,典型的属性重要性定义方法主要有基于 Pawlak[15] 和基于条件熵[16] 两类。这两类属性重要度度量方法都是从条件属性集中去掉或向条件属性子集添加某个属性,再考查分类的变化情况,分类变化越大,说明该属性越重要,反之,属性重要度越低。Pawlak 的属性重要度定义侧重属性的定性分析,条件熵的属性重要度定义侧重属性的定量分析[17]。为了更准确地度量各属性在对象分类时所起的作用,本文从概率统计的角度提出一种新的属性重要度定义方法。

在信息表 $S=(U,Atr=C\cup D,V,f)$ 中,设属性 $c_i\in C(i=1,2,\cdots,m)$,根据属性 c_i 的取值划分的类为 $X_j^{c_i}(j=1,2,\cdots,u)$,决策属性 D 根据其取值划分的类为 $Y_k(k=1,2,\cdots,v)$。

比如在表 1 中,属性 c_1 的取值值域为 $\{0,1\}$,根据属性 c_1 的取值为 0 或 1 把邮件对象分别划分到两个类别:$\{X_1^{c_1}=\{x_1,x_4\},X_2^{c_1}=\{x_2,x_3,x_5\}\}$;属性 c_2 的取值值域为 $\{0,1,2\}$,根据属性 c_2 的取值把对象划分为 3 类:$\{X_1^{c_2}=\{x_1,x_2\},X_2^{c_2}=\{x_3,x_4\},X_3^{c_2}=\{x_5\}\}$;决策属性 d_1 的取值值域为 $\{0,1\}$,根据其取值把对象划分为两类:$\{Y_1=\{x_1,x_2,x_3\},Y_2=\{x_4,x_5\}\}$。

设条件属性划分类 $X_j^{c_i}$ 对决策属性划分类 Y_k 的条件概率为 $P(Y_k|X_j^{c_i})$,而决策属性划分类 Y_k 对条件属性划分类 $X_j^{c_i}$ 的条件概率为 $P(X_j^{c_i}|Y_k)$,其公式定义如下:

$$P(Y_k|X_j^{c_i})=|Y_k\cap X_j^{c_i}|/|X_j^{c_i}|\tag{5}$$

$$P(X_j^{c_i}|Y_k)=|X_j^{c_i}|/|Y_k|/|Y_k|\tag{6}$$

其中,$i=1,2,\cdots,m;j=1,2,\cdots,u;k=1,2,\cdots,v$。

$P(Y_k|X_j^{c_i})$ 和 $P(X_j^{c_i}|Y_k)$ 都是反映在属性 c_i 划分下的类对决策属性分类能力的大小,$P(Y_k|X_j^{c_i})$ 和 $P(X_j^{c_i}|Y_k)$ 同时取

值越大则表示属性 c_i 的分类能力越强，比如当 $P(Y_k|X_j^{c_i})=1$ 且 $P(X_j^{c_i}|Y_k)=1$ 时，可以判定属性 c_i 的分类能力等价于决策属性的分类能力。因此，对于 $j=1,2,\cdots,u$，每个属性对应的条件概率 $P(Y_k|X_j^{c_i})$ 和 $P(X_j^{c_i}|Y_k)$ 均取最大值，这样取值可确保其确定性程度最大，也更符合概率统计的实际意义，即：

$$P(Y_k|X^{c_i})=\max(P(Y_k|X_j^{c_i})) \tag{7}$$

$$P(X^{c_i}|Y_k)=\max(P(X_j^{c_i}|Y_k)) \tag{8}$$

比如，对于属性 c_1 有如下计算：

$$P(Y_1|X_1^{c_1})=\frac{|Y_1\bigcap X_1^{c_1}|}{|X_1^{c_1}|}=\frac{|\{x_1,x_2,x_3\}\bigcap\{x_1,x_4\}|}{|\{x_1,x_4\}|}=\frac{1}{2}$$

$$P(Y_1|X_2^{c_1})=\frac{|Y_1\bigcap X_2^{c_1}|}{|X_2^{c_1}|}=\frac{|\{x_1,x_2,x_3\}\bigcap\{\{x_2,x_3,x_5\}|}{|\{x_2,x_3,x_5\}|}=\frac{2}{3}$$

取最大值时 $P(Y_1|X^{c_1})=\frac{2}{3}$，同理可得：

$$P(Y_2|X^{c_1})=\frac{1}{2},P(X^{c_1}|Y_1)=\frac{2}{3},P(X^{c_1}|Y_2)=\frac{1}{2}$$

对于邮件数据集而言，邮件只分为两类：正常邮件和垃圾邮件，即 $k=P,N$。$k=P$ 表示正常邮件类，$k=N$ 表示垃圾邮件类，则有以下表达：

$$P(Y_P|X^{c_i})=\max(|Y_P\bigcap X_j^{c_i}|/|X_j^{c_i}|)$$

$$P(X^{c_i}|Y_P)=\max(|Y_P\bigcap X_j^{c_i}|/|Y_P|)$$

$$P(Y_N|X^{c_i})=\max(|Y_N\bigcap X_j^{c_i}|/|X_j^{c_i}|)$$

$$P(X^{c_i}|Y_N)=\max(|Y_N\bigcap X_j^{c_i}|/|Y_N|) \tag{9}$$

其中，$i=1,2,\cdots,m;j=1,2,\cdots,u$。

则属性 c_i 对决策属性 D 的正域划分能力和负域划分能力可分别定义为：

$$E_P^{c_i}=\sqrt{(P(Y_P|X^{c_i}))^2+(P(X^{c_i}|Y_P))^2} \tag{10}$$

$$E_N^{c_i}=\sqrt{(P(Y_N|X^{c_i}))^2+(P(X^{c_i}|Y_N))^2} \tag{11}$$

那么属性 c_i 的重要性则可定义为：

$$SGF^{c_i}=l_1E_P^{c_i}+l_2E_N^{c_i}(l_1,l_2\in(0,1)) \tag{12}$$

当 $l_1=1,l_2=0$ 时，属性重要性的定义侧重于正域（正常邮件判定）的划分；当 $l_1=0,l_2=1$ 时，属性重要度的定义侧重于负域（垃圾邮件的判定）的划分；当 $l_1=0.5,l_2=0.5$ 时，属性重要性的定义取其平均判定。

比如，对于属性 c_1：

$$E_P^{c_1}=\sqrt{(2/3)^2+(2/3)^2}=0.94$$

$$E_N^{c_1}=\sqrt{(1/2)^2+(1/2)^2}=0.70$$

若取 $l_1=0.5,l_2=0.5$，则属性 c_1 的重要度大小的计算为：$SGF^{c_1}=0.5\times E_P^{c_1}+0.5\times E_N^{c_1}=0.82$。

2.2 邮件头的特征提取

一个用户在判断一封邮件是否为垃圾邮件时，往往首先从其主要形象特征即可做出判断（比如根据邮件发送者的信息往往能判断其是否为垃圾邮件），而不会仔细阅读邮件本身的内容。因此，本文提出的垃圾邮件过滤方法是基于邮件头的主要形象特征来做出判断。考虑邮件格式、时间等特征，给出基于邮件头信息的 12 个属性的邮件主要形象特征集，如表 2 所列。

表 2 电子邮件的形象特征描述

特征	取值	特征说明
c_1	整数	收件人个数，大于 4 时取值为 4
c_2	0,1	发送时间：0:00—6:00 取值 0；6:00—00:00 取值 1
c_3	0,1	主题(Subject)：有主题取值 1，否则取值 0
c_4	0,1	抄送(Cc)：有抄送取值 1，否则取值 0
c_5	0,1	邮件格式(Text/HTML)：Text 取值 1，HTML 取值 0
c_6	0,1	邮件类型：回复类型取值 1，否则取值 0
c_7	0,1	“From”中的原始发送地址与“Received”中的原始发送地址一致则取值 1，否则取值 0
c_8	整数	邮件路由信息中断次数，大于 4 时取值为 4
c_9	0,1	“To”中的目的地址与“Received”中“for”的地址一致则取值 1，否则取值 0
c_{10}	0,1	“To”中的目的地址与“Received”中的实际收信人地址一致则取值 1，否则取值 0
c_{11}	0,1	“Message-ID”项与“From”项一致则取值 1，否则取值 0
c_{12}	0,1	“Delivered-To”项与“To”项一致则取值 1，否则取值 0

2.3 邮件过滤算法

假设一封邮件信息包含条件属性 C，根据属性重要性大小重新排序为 $C=\{c_1,c_2,\cdots,c_m\}$，将 C 划分为 m 个属性子集 $C_1,C_2,\cdots,C_m$，$C_1=\{c_1\}$，$C_2=C_1\cup\{c_2\}$，$\cdots$，$C_m=C_{m-1}\cup\{c_m\}$，则基于邮件头信息的三支决策垃圾邮件过滤方法如下。

(1)根据属性子集 $C_j(1\leqslant j\leqslant m)$（初始化 $j=1$）对邮件 x_i $(1\leqslant i\leqslant n)$ 计算贝叶斯概率 $P(Y|x_i(C_j))$，表示为在属性子集 C_j 信息下邮件 x_i 的贝叶斯概率。给定阈值 α 和 $\beta(0<\beta<\alpha<1)$，进行三支决策分类，对于能够判定类别的邮件进行即时处理，分别划分到相应的正域 POS 和负域 NEG 中；而对于不能确定类别的邮件并将其判定为可疑邮件并划分到边界域 BND 中。

(2)如果在属性子集 $C_j(1\leqslant j<m)$ 下的三支决策划分边界域不为空，即当前的信息不足以对所有的邮件做出决策，则按属性重要度大小顺序在属性子集 C_j 中增加一个属性，即属性子集变为 C_{j+1}，然后继续(1)中的操作以进行进一步的决策，直到边界域为空或者所有属性信息都已使用完全。

综上所述，邮件过滤算法的算法流程如下。

输入：一个电子邮件的邮件头信息表 S=(U,Atr=C∪D,V,f)，阈值 α 和 β

输出：邮件分类情况

步骤 1 数据预处理：根据表 2 中的邮件头信息特征提取相应特征并取值。

步骤 2 属性重要性计算：根据式(12)计算每个条件属性 $c_i\in C$ 相对决策属性 D 的重要性并按大小排序为 $C=\{c_1,c_2,\cdots,c_m\}$。将 C 划分为 m 个属性子集，即 $C_1,C_2,\cdots,C_m$。$C_1=\{c_1\}$，$C_2=C_1\cup\{c_2\}$，$C_m=C_{m-1}\cup\{c_m\}$。

步骤 3 邮件分类：计算贝叶斯概率 $P(Y|x_i(C_j))$ 并进行序贯三支决策。

```
for (j=0;j<m;j++)
{ for (i=0;i<n;i++)
  {
    if (P(Y|xi(Cj))≥α) xi∈POS(Y);
    else if(P(Y|xi(Cj))≤β) xi∈NEG(Y);
    else xi∈BND(Y);
  }
  if(BND(Y)=null)break;}
```

3　仿真实验

3.1　算法评价标准

通常借用文本分类的相关指标对垃圾邮件过滤的性能进行评价。具体地，假设待测试的邮件集合中共有 N 封邮件，令 N_L 和 N_S 分别表示分类好的正常邮件和垃圾邮件的个数，$n_{l\to s}$ 表示将正常邮件分到垃圾邮件的个数，$n_{s\to l}$ 表示将垃圾邮件分到正常邮件的个数，$n_{l\to l}$ 表示正确查出的正常邮件个数，$n_{s\to s}$ 表示正确查出的垃圾邮件个数，令 $N_{L+S}=N_L+N_S$。常用的衡量垃圾邮件过滤系统性能的指标如下。

1）召回率（Recall）：$Rec=n_{s\to s}/(n_{s\to s}+n_{s\to l})$，即垃圾邮件检出率。该指标反映了过滤系统发现垃圾邮件的能力，召回率越高，“漏网”的垃圾邮件就越少。

2）精确率（Precision）：$Pre=n_{s\to s}/(n_{s\to s}+n_{l\to s})$，即垃圾邮件检对率。精确率反映了过滤系统“找对”垃圾邮件的能力，精确率越大，将合法邮件误判为垃圾邮件的可能性越小。

3）准确率（Accuracy）：$Acc=(n_{s\to s}+n_{l\to l})/N$，即对所有邮件（包括垃圾邮件和合法邮件）的判对率。如果不考虑三支决策中边界域 BND 的判别情况，则准确率表示为 $Acc2=(n_{s\to s}+n_{l\to l})/N_{L+S}$。

4）错误率（Error rate）：$Err=(n_{s\to l}+n_{l\to s})/N$，即对所有邮件（包括垃圾邮件和合法邮件）的判错率。如果不考虑三支决策中边界域 BND 的判别情况，则错误率表示为 $Err2=(n_{s\to l}+n_{l\to s})/N_{L+S}$。

5）F 值：$F=2Rec\times Pre/(Rec+Pre)$，F 值实际上是召回率和精确率的调和平均，它将召回率和精确率综合成一个指标。

3.2　实验过程及结果

为了验证本文提出的邮件过滤方法的有效性，从 trec06p 邮件库[18]中选取语料，trec06p 邮件库一共有 37822 封邮件，包括正常邮件 12910 封（占比 0.34）和垃圾邮件 24912 封。实验选取其中的 30000 封邮件作为训练集（占比约 4/5），剩下的 7782 封邮件作为测试集。具体实验过程如下：

（1）根据邮件头的特征提取信息，即根据表 2 所描述的属性进行特征提取，形成邮件决策表。其中决策属性 $D=\{0,1\}$，0 代表垃圾邮件，1 代表正常邮件。

（2）在训练集中（30000 封邮件），根据邮件头中的 From 信息确定发件人信息（每个发件人有多于 1 封的邮件），然后根据发件人信息，每次随机选取发件人信息表中 1/10 的发件人数据作为训练集，如此训练 10 次得出先验概率 $P(Y)$、后验概率 $P(x|Y)$ 以及后验概率 $P(x|\neg Y)$ 的平均值，随机选取并取平均值是为了避免随机性。比如在表 1 中，对于属性 c_1，取值为 0 时的后验概率 $P(0|Y)=1/3$，$P(0|\neg Y)=1/2$，取值为 1 时的后验概率 $P(1|Y)=2/3$，$P(1|\neg Y)=1/2$，而先验概率 $P(Y)=3/5$。从概率统计的角度来看，在实验中对大量的邮件随机选取数据来进行训练并取平均值可以很好地表现数据的真实特征。

（3）在测试集中（7782 封邮件），根据邮件头中的 From 信息确定发件人信息，然后根据发件人信息，每次随机选取发件人信息表中 1/10 的发件人数据，经多次实验人工设定阈值 $\alpha=0.8$，$\beta=0.2$。分别采用文献[6]中的方法、文献[7]中的方法和本文方法进行测试，每种方法测试 10 次（本文方法会记录实际用到的属性个数并取平均值）并取得平均值，得出的实验结果如表 3 所列。

表 3　邮件集 Trec06p 的测试结果/%

评价指标	文献[6]中的方法	文献[7]中的方法	本文方法
Rec	92.56	94.67	**95.64**
Pre	77.36	89.45	88.84
Acc	76.42	88.71	**88.76**
Err	23.58	11.29	**11.24**
Acc2	52.85	70.33	**88.76**
Err2	16.31	8.95	11.24
F 值	84.09	91.97	**92.09**
BND	30.84	21.72	0.00

通过对测试结果的分析，可以得出以下结论：

（1）文献[6]中的方法是增加属性信息后进行序贯三支决策来分类邮件，但是并没有考虑增加属性的顺序，即没有考虑属性的重要度问题。从与本文方法的对比中可以看出，按照属性重要度大小顺序进行序贯三支决策的邮件判断精确率、准确率和错误率都要优于属性无序的序贯三支决策，即在序贯决策过程中增加属性的顺序对实验的效果会有很大的影响。

（2）文献[7]中的方法是取得数据的全部属性信息后，不采用序贯决策的手段，而是对全部属性信息计算贝叶斯概率，然后进行三支决策分类邮件。从与本文方法的对比中可以看出，本文方法的召回率、准确率和 F 值等指标优于文献[7]中的方法，其他指标略差于文献[7]中的方法，同时文献[7]中的方法对于其中约 20%的邮件不能判断类别。即采用序贯决策和不采用序贯决策的实验结果的大部分指标相差不大，但不采用序贯决策时由于只对邮件判断一次，因此对部分邮件不能判断类别。

（3）采用文献[6]中的方法和文献[7]中的方法会有约 20%～30%的邮件不能判断类别，即实验中邮件头信息的 12 个属性即使全部都用上也不能将所有邮件判断完全，而本文方法平均只用了 3 个属性的信息（实验中得出的实际用到的属性个数）就可以将所有邮件判断完毕，表明本文所定义的属性重要度是有意义且有效的。

（4）从本文方法的测试结果中可以看出，几个指标均有较优的表现，尤其是在召回率指标上体现得更为明显，说明本文方法发现垃圾邮件的能力较强。

结束语　本文提出了一种基于邮件头信息的三支决策垃圾邮件过滤方法，定义了新的属性重要性，并用该度量方法对邮件头信息属性依据重要性大小进行排序，然后按属性重要性的大小顺序对邮件计算贝叶斯概率并进行三支决策，当信息较少以致不足以决策时，增加属性信息以帮助进一步的决策，直到得到最后的邮件分类。通过在标准邮件集上与其他邮件过滤算法进行测试对比，证明该方法能有效地根据邮件头信息对垃圾邮件进行过滤，从而提高垃圾邮件过滤的准确性。下一步的工作将分析不同的属性重要性度量方法对过滤结果的影响、阈值 α 和 β 的取值优化问题，以及结合邮件头和邮件内容信息来判断邮件的类别。

（下转第 114 页）

[3] LIAN Z X, YU J, XU L M. Improved LMMSE Channel Estimation Algorithm[J]. Computer Science, 2014, 41(4): 53-56. (in Chinese)
练柱先,余江,徐丽敏. 一种改进的 LMMSE 信道估计算法[J]. 计算机科学,2014,41(4):53-56.

[4] COVER T M, GAMAL A E. Capacity theorems for the relay channel [J]. IEEE Trans. Inf. Theory, 1979, 25(5): 572-584.

[5] LANEMAN J N, TSE D N C, WORNELL G W. Cooperative diversity in wireless networks: Efficient protocols and outage behavior [J]. IEEE Trans. Inf. Theory, 2004, 50(12): 3062-3080.

[6] ADINOY A, FAN Y, YANIKOMEROGLU H, et al. Performance of selection relaying and cooperative diversity[J]. IEEE Trans. on Wireless Commun., 2009, 8(12): 5790-5795.

[7] ZOU Y L, ZHENG B Y. Adaptive Cooperative Transmission Scheme Based on Distributed Relay Selection [J]. ACTA Electronica Sinica, 2009, 37(1): 13-20. (in Chinese)
邹玉龙,郑宝玉. 基于分布式中继选择的自适应协作传输方案[J]. 电子学报,2009,37(1):13-20.

[8] RIIHONEN T, WERNER S, WICHMAN R. On the feasibility of full-duplex relaying in the presence of loop interference[C]// 10th IEEE Workshop on Signal Processing Advances in Wireless Communications. 2009: 275-278.

[9] LUO X Y. Study on key technology for front-end of receiver and transmitter in full duplex communication system [D]. Chengdu: University of Electronic Science and Technology of China, 2013. (in Chinese)
罗馨逸. 全双工通信系统收发前端关键技术研究[D]. 成都:电子科技大学,2013.

[10] KRIKIDIS I, SURAWEERA H A, SMITH P J, et al. Full-duplex relay selection for amplify-and-forward cooperative networks[J]. IEEE Trans. Wireless Commun., 2012, 11(12): 4381-4393.

[11] YANG K. Efficient full-duplex relaying with joint antenna relay selection and self-interference suppression [J]. IEEE Trans. on Wireless Commun., 2015, 14(7): 3991-4005.

[12] CHEN L. Optimal power allocation for dual-hop full-duplex decode-and-forward relay [J]. IEEE Commun. Lett., 2015, 19(3): 471-474.

[13] KWON T, LIM S, CHOI S, et al. Optimal duplex mode for DF relay in terms of the outage probability[J]. IEEE Trans. Veh. Technol., 2010, 59(7): 3628-3634.

[14] ZHONG F J, ZHAO Y C. Study on performance of full-duplex relay selection strategy [J]. Journal of Southwest Jiaotong University, 2015, 50(5): 912-917. (in Chinese)
仲福建,赵永驰. 全双工中继选择策略的性能研究[J]. 西南交通大学学报,2015,50(5):912-917.

(上接第 77 页)

参 考 文 献

[1] Internet Society of China. China anti-spam survey report in the first quarter of 2014 [J]. China Internet, 2014(7): 59-67. (in Chinese)
中国互联网协会. 2014 年第一季度反垃圾邮件调查报告[J]. 互联网天地,2014(7):59-67.

[2] CHEN Z X. Review of spam filtering technology[J]. Application Research of Computers, 2009, 26(5): 1612-1615(in Chinese)
陈志贤. 垃圾邮件过滤技术研究综述[J]. 计算机应用研究,2009,26(5):1612-1615.

[3] YAO Y Y. Three-Way Decision: An Interpretation of Rules in Rough Set Theory[M]// Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2009: 642-649.

[4] YAO Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180(3): 341-353.

[5] YAO Y Y. An Outline of a Theory of Three-Way Decisions[M]// Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012: 1-17.

[6] LI J L, DENG X F, YAO Y Y. Multistage Email Spam Filtering Based on Three-Way Decisions[M] // Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2013: 313-324.

[7] ZHOU B, YAO Y Y, LUO J G. Cost-sensitive three-way email spam filtering [J]. Journal of Intelligent Information Systems, 2014, 42(1): 19-45.

[8] JIA Y X, SHANG L. Three-Way Decision Versus Two-Way Decisions on Filtering Spam Email[M] // Transactions on Rough Sets XVIII. Springer Berlin Heidelberg, 2014: 69-91.

[9] DENG W B, HONG Z Y. Two stage email filtering method based on rough set[J]. Journal of Computer Applications, 2010, 30(8): 2006-2009, 2048.

[10] PAWLAK Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[11] YAO Y Y. Decision-Theoretic Rough Set Models[C]// International Conference on Rough Sets and Knowledge Technology. New York: Springer-Verlag, 2007: 1-12.

[12] MARINOFF L. The Middle Way: Finding Happiness in a World of Extremes[M]. New York: Sterling, 2007.

[13] 贾修一,等. 三支决策理论与应用[M]. 南京:南京大学出版社,2012.

[14] 于洪等. 三支决策:复杂问题求解方法与实践[M]. 北京:科学出版社,2015.

[15] PAWLAK Z. Rough sets: Theoretical aspects of reasoning about data[M]. London: Kluwer Academic Publishers, 1991.

[16] LIANG J Y, QIAN Y H. Information granules and entropy theory in information systems [J]. Science in China-Series, 2008, 51(3): 1427-1444.

[17] YE J, ZHU H S, LI M. Kind of attribute importance defined method and its application in attribute reduction [J]. Application Research of Computers, 2016, 33(7): 2075-2078. (in Chinese)
叶军,朱华生,黎敏. 一种属性重要性定义方法及其在约简中的应用[J]. 计算机应用研究,2016,33(7):2075-2078.

[18] Text Retrieval Conference. 2006 TREC Public Spam Corpora [EB/OL]. [2016-04-21]. http://plg.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/foo06.