# 粗糙集与决策树在电子邮件分类与过滤中的应用

邓春燕[1,3],陶多秀[2],吕跃进[3]
DENG Chun-yan[1,3],TAO Duo-xiu[2],LV Yue-jin[3]

1.广西河池学院 计算机与信息科学系,广西 宜州 546300
2.广西大学 电气工程学院,南宁 530004
3.广西大学 数学与信息科学学院,南宁 530004
1.Department of Computer and Information Science,Hechi University,Yizhou,Guangxi 546300,China
2.College of Electrical Engineering,Guangxi University,Nanning 530004,China
3.College of Mathematics and Information Science,Guangxi University,Nanning 530004,China
E-mail:dchy7072@126.com

**DENG Chun-yan,TAO Duo-xiu,LV Yue-jin.Application of rough set and decision tree in e-mail classification and filtering.Computer Engineering and Applications,2009,45(16):138-140.**

**Abstract**: Spam identification and filtering is one of the hot issues.And the rough set is a new data analysis tool to deal with ambiguity and uncertainty of knowledge;it has been successfully applied to many areas of classification.Combining rough sets with decision tree,a spam filtering solution based on rough sets and decision tree (RS-DT) was proposed.The feasibility of the solution was indicated by the experiments on the public email corpus.Comparison experiments were also made between SVM classifier, Bayes classifier and RS-DT model.The results show that the RS-DT model can not only reduce the error rate of judging the normal email as spam,but also improve adaptive learning of the filtration system.
**Key words**: spam;rough set;data mining;decision tree

**摘 要** 垃圾邮件的识别与过滤是目前研究的热点问题之一。而粗糙集是一种新的处理模糊和不确定性知识的数据分析工具,已被成功地应用到许多有关分类的领域。将粗糙集与决策树结合,提出一个基于 RS-DT 的邮件分类方案与模型,并进行了实验及结果分析。通过与朴素贝叶斯模型及 SVM 的比较,表明提出的基于 RS-DT 的模型可以降低把正常邮件错分为垃圾邮件的比率,提高过滤系统的自学习能力。
**关键词** 垃圾邮件 粗糙集 数据挖掘 决策树
**DOI**:10.3778/j.issn.1002-8331.2009.16.040 **文章编号**:1002-8331(2009)16-0138-03 **文献标识码**:A **中图分类号**:TP182

## 1 引言

随着计算机技术的迅猛发展,网络覆盖与应用的范围日益广泛,电子邮件正被越来越多的人所使用。然而,电子邮件在给人们带来极大便利的同时,也带来了不少干扰,比如:垃圾邮件的不时侵袭。垃圾邮件(Spam 或 Junk Mail),也称为 UCE(Unsolicited Commercial E-mail,不请自来的商业邮件)或 UBE(Unsolicited Bulk E-mail,不请自来的大量电子邮件),一般被概括为:向新闻组或他人电子信箱发送的未经用户准许、不受用户欢迎的、难以退掉的电子邮件或电子邮件列表。从事此类活动的人员叫作垃圾邮件制造者(Spammer),其主要来源为:商业广告,站点宣传、恶意攻击的病毒携带邮件等。垃圾邮件一方面影响用户的工作和生活,因为垃圾邮件会大量吞食用户的邮箱空间,占用用户的时间,而且会带来负面信息、一些病毒程序或带有病毒的网络连接,甚至可获得用户的私人信息或者导致用户机器中毒;另一方面还影响了网络的正常运行,垃圾邮件大量占用带宽,严重影响邮件服务器工作效率,甚至造成网络阻塞。国内外研究的垃圾邮件过滤的方法主要有:查表法(黑白名单,RBI)、溯源法、统计法(如贝叶斯分类器[1])、质询-回应技术,但这些方法都缺少自学习能力,而文献[2]则只考虑邮件头,忽略了邮件内容的一些重要特征。

粗糙集理论是 20 世纪 80 年代初 Pawlak 首次提出[3]的一种新的处理模糊和不确定性知识的数据分析工具。由于在应用中不需要先验知识等特点,目前它已成为国内外学术研究的热点,在理论研究、图像处理、模式识别、机器学习、知识获取、数据挖掘和决策分析等许多领域得到了广泛应用[4-5]。

本文提出一种基于粗糙集-决策树(RS-DT)的邮件分类方

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70861001);广西研究生科研创新项目(the Innovative Scientific Research Project of Guangxi Graduate No.2008105930701)。
作者简介:邓春燕(1971-),女,讲师,主要研究领域为数据挖掘、粗糙集理论与方法;陶多秀(1985-),女,在读硕士研究生,主要研究领域为数据挖掘、商业智能、管理决策;吕跃进(1958-),男,教授,主要研究领域为数据挖掘、预测与决策。
收稿日期:2008-12-15 修回日期:2009-02-25

案,将邮件头的特征和邮件文本的特征结合进行分析,扩展了现有的条件属性,随后进行粗糙集属性约简,降低向量空间维数,减少了特征数,降低了待分类邮件数据集测试数据集的向量空间的生成时间,从而提高分类速度,然后采用决策树算法建树,提取规则,对邮件进行分类,识别出垃圾邮件并进行过滤,以提高分类正确率,降低把正常邮件错划为垃圾邮件的比率,提高过滤系统的自学习能力。

## 2 电子邮件机理

电子邮件是半结构化的文件。RFC2822 规定了电子邮件在网络传输中的基本格式,RFC1341 在 RFC822 的基础上又扩充了多用途因特网邮件扩展 MIME (Multipurpose Internet Mail Extensions) 协议,这二者定义了目前正广泛使用的邮件格式。电子邮件通常包括邮件头和正文。RFC2822 为邮件头定义了 20 多个标准字段,主要有:Message-ID 邮件的唯一标识符,From 表示邮件的发送方地址,To 表示邮件收件人,Subject 表示邮件的主题,Date 表示邮件的创建时间等。这些字段表征了一封邮件的属性,可用于识别和分类邮件。

## 3 粗糙集基本理论简介

### 3.1 信息系统、决策表

信息系统是粗糙集理论及其应用紧密相关的一个概念。

一个信息系统[4]定义为一个四元组 $S=(U,A,V,f)$,其中 $U$ 是对象的非空有限集合,称为论域;$A$ 为属性的非空有限集;$V=\bigcup_{a\in A}V_a$,$V_a$ 是属性 $a$ 的值域;$f:U\times A\to V$ 是一个信息函数,$\forall a\in A, x\in U, f(x,a)\in V_a$。这种定义方式使对象的知识可以方便地以数据表格的形式描述。

若在四元组 $S=(U,A,V,f)$ 中,$A=C\cup D$,且 $C\cap D=\varnothing$,$C$ 为条件属性集,$D$ 为决策属性集,则将这样的信息系统称为决策表或决策信息系统。

### 3.2 不可区分关系

不可区分关系(或称不分明关系[5])在 $S=(U,A,V,f)$ 中,设 $P\subseteq A$,且 $P\neq\varnothing$,$P$ 中所有等价关系的交集 $\cap P$ 称为 $P$ 上的一种不可区分关系(或称不分明关系),记作 $IND(P)$,即

$$IND(P)=\{(x,y)\in U\times U: a(x)=a(y), \forall a\in P\}$$

$$[x]_{IND(P)}=\bigcap_{a\in P}[x]_a$$

其中 $[x]_a$ 表示属性(关系) $a$ 中包含元素 $x\in U$ 的概念或等价类。

### 3.3 属性约简

粗糙集属性约简是将决策表中对决策分类不必要的属性进行约简,在一定程度上去掉决策表中的冗余信息,对于每一条实例来说可能仍然有不必要的属性存在,因此在不引起冲突的条件下,可将每一条实例中的该属性删除。

把邮件向量空间模型看成是决策表,作为粗糙集处理模块的输入,通过离散算法对连续性数据进行离散化,再通过约简算法对新生成的决策表进行约简,降低向量空间维数,减少了特征数,降低了待分类邮件数据集测试数据集的向量空间的生成时间,从而提高分类速度。

## 4 决策树简介

决策树[6]是一个可以自动对数据进行分类的树型结构,是树型结构的知识表示,便于直接转换为决策规则。决策树构造的输入是一组带有类别标记的例子,构造的结果是一棵二叉树或多叉树。内部节点是属性,边是该属性的所有取值,有几个属性值,就有几条边,树的叶子节点是类别标记。决策树的生成是一个从根节点开始、从上到下的递归过程,一般根据分而治之的思想,通过不断地将训练样本分割成子集来构造决策树。

## 5 RS-DT 邮件过滤解决方案

对邮件样本集进行数据预处理,包括:分词处理、特征选取、生成向量空间;然后用粗糙集属性约简,进行数据降维,形成新的向量空间,把这新空间作为决策表,当作输入,用决策树进行分类,提取规则,达到分类邮件,识别和过滤垃圾邮件的目的。方案如图 1 所示。

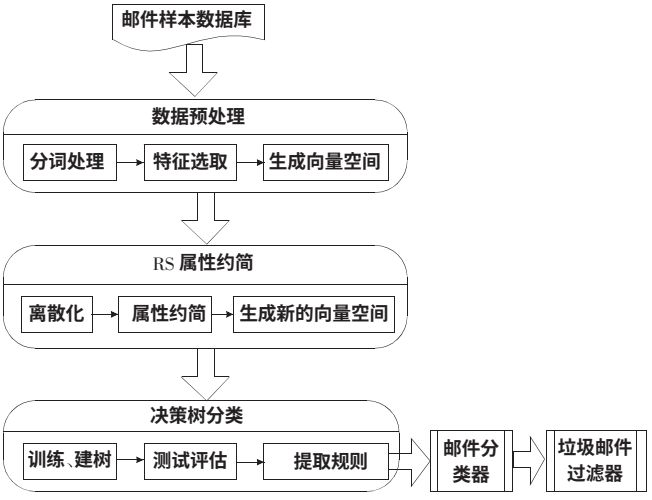



图 1 模型结构示意图

### 5.1 邮件决策表

样本邮件经过邮件头特征提取、附件特征提取以及正文特征提取(其中在提取正文特征时先经过邮件解码)后,选取了 25 个特征属性。

根据粗糙集对信息系统的定义,将邮件的集合作为论域,每篇邮件作为论域中的对象,特征项集作为条件属性集,即各特征项为条件属性,特征项在邮件文本中的权重作为该属性在对象上的属性值,邮件所属类别作为决策属性,则这样的决策表就是一个待分类的电子邮件决策系统。其中每行代表一封邮件,每列代表一个属性,$C_1$ 到 $C_n$ 的值代表邮件的条件属性,$D$ 代表决策属性,取值 $V_D=\{0,1,2\}$,分别表示正常邮件、可疑邮件、垃圾邮件。示例如表 1 所示。

表 1 邮件分类决策表示例

| $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | … | $C_{25}$ | $D$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 0 | 0.12 | … | 0 | 0 |
| 0 | 2 | 5 | 1 | 1 | 2 | 0.80 | … | 5 | 1 |
| 0 | 3 | 4 | 2 | 0 | 1 | 0.20 | … | 5 | 1 |
| … | … | … | … | … | … | … | … | … | … |
| 1 | 2 | 2 | 1 | 1 | 1 | 0.76 | … | 6 | 2 |

$C_1$:发送者域名与 IP 是否一致,取值 $V_{C_1}=\{1,0\}$,分别表示一致、不一致。

$C_2$:发送者地址,按邮箱所在服务器进行编码。163、新浪的可分别编为 1、2 等。

$C_3$:同时收到该封邮件的收件人个数,取值为整数。

$C_4$:邮件类型,取值 $V_{C_4}$={0,1,2}表示直接发送、回复、转发。

$C_5$:标题是否包含有意义的词,$V_{C_5}$={0,1}表示无、有。

$C_6$:附件数目,取值为整数。

$C_7$:字符! 出现的频度。

$C_8$:字符$出现的频度。

…

$C_{25}$:正文 http 连接数。

## 5.2 数据预处理

决策表离散化:由于粗糙集较适合于处理离散化的属性值,故对上述的一些取值较多的属性需要进行离散化处理。为了与分类的决策树结合较好,采用信息熵的方法[7]来离散化。

条件属性约简:为使空间降维,采用基于可辨识矩阵的 Johnson 算法[8]进行属性约简。

## 5.3 决策树建树算法

采用自顶向下方法递归地构造决策树。按照经典的 ID3 建树算法思想[6],根据信息熵的原理,选取使信息增益最大者作为分裂属性进行建树。目的是使得决策树对划分的不确定程度减小。

Generate_decision_tree (*Si*,attributelist,test_attribute) //由给定的训练数据产生一棵邮件分类决策树

输入:邮件训练样本集 samples,候选条件属性的集合 attributes_list,即 $C_1$ 到 $C_n$

输出:一棵决策树

基本步骤如下:

(1) 创建节点 *N*;

(2) if 邮件训练样本集 samples 都在同一个类 C then;

(3) 返回 *N* 作为叶节点,以类 C 标记;

(4) if attributes_list 为空 then;//算法终止的一个条件

(5) 返回 *N* 作为叶节点,标记为 samples 中最普通的类,即出现最多的类;

(6) 选择 attributes_list 中具有最高信息增益的属性作为 test_attribute;

(7) 标记节点 *N* 为 test_attribute;

(8) for each test_attribute 中的已知值 *ai*;

(9) 由节点 *N* 长出一个条件为 test_attribute=*ai* 的分枝;

(10) 设 *Si* 是 samples 中 test_attribute=*ai* 的样本的集合;//一个划分

(11) if *Si* 为空 then //算法终止的另一个条件

(12) 加上一个叶节点,标记为 samples 中最普通的类;

(13) else 加上一个由Generate_decision_tree (*Si*,attributelist,test_attribute) 返回的节点,继续分裂节点。 //递归划分

## 5.4 规则提取

决策树建立后,可以从中导出分类规则,称为规则抽取。分类规则以"IF-TNEN"形式表示。从根到叶节点的每条路径创建一个规则,规则的前件("IF"部分)是从根节点出发到达该叶节点路径上所有的中间节点构成的一个"与"判断(沿着给定路径上的每个属性-值对的一个合取项),而规则的后件("THEN"部分,结论)就是叶节点的类别。规则举例说明,如:

rule1:IF $C_1$=0 and $C_8$≥0.8 then $D$=2

其实际意义为:如果发送者域名和 IP 地址不一致且字符$出现的频度≥0.8,则这封邮件为垃圾邮件。

## 5.5 垃圾邮件过滤评价指标说明

召回率 Recall:$Re=\frac{D}{B+D}$,系统发现垃圾邮件的能力,即垃圾邮件检出率。这个指标反映了发现垃圾邮件的能力,召回率越高,漏网的垃圾邮件就越少;

垃圾邮件的检对率 Precision:$Pre=\frac{D}{C+D}$;

精确率 Accuracy:所有判别正确的邮件数与所有进行分类的邮件数的比率 $Accur=\frac{A+D}{N}$,*N* 为邮件总数;

错误率 Error rate:$Err=1-Accur$ 等。

为了衡量将正常邮件划分为垃圾邮件的比率,还采取了一种新的评估指标 *F*1:即 $F1=\frac{C}{A+C}$。

表 2 垃圾邮件过滤评价指标说明

| 垃圾邮件过滤评价指标说明 | 实际为正常邮件 | 实际为垃圾邮件 |
|---|---|---|
| 分类判定为正常邮件 | A | B |
| 分类判定为垃圾邮件 | C | D |

# 6 实验及结果分析

实验数据集采用 UCI 机器学习数据库[9]中的垃圾邮件数据库 Spam E-mail Database。数据集包含 4 601 个实例,每个实例分别用 58 个特征属性来描述,其中 57 个为邮件特征属性,1 个为类别标识,即最后一列,其中垃圾邮件 1 813 封。决策表离散化、属性约简等均按文中所述的对应方法。实验中将样本划分为训练集和测试集,采用 10 折-交叉验证,进行多次实验,取平均结果。

此外,作者还收集了 3 个常用个人邮箱的 100 封邮件。其中正常邮件 70 封,垃圾邮件 20 封,可疑邮件 10 封。按文中提取邮件的特征属性来构造决策表,记为 DataSet(100)。

由实验结果(如表 3 所示)可知,将粗糙集与决策树结合,能较大限度地降低属性空间维数,而且正确率、召回率、精确率变化不大,而将正常邮件划分为垃圾邮件的比率 *F*1 有较大程度的降低。

表 3 实验结果

| 语料 | 评价指标 | SVM/(%) | RS-DT/(%) | 贝叶斯/(%) |
|---|---|---|---|---|
| Spam E-mail Database | Recall | 86.53 | 85.87 | 86.21 |
| | Precision | 79.67 | 84.10 | 85.22 |
| | Accuracy | 86.00 | 85.33 | 84.97 |
| | F1 | 8.22 | 5.23 | 6.58 |
| DataSet (100) | Recall | 94.80 | 94.24 | 92.79 |
| | Precision | 92.58 | 94.71 | 94.20 |
| | Accuracy | 94.60 | 94.20 | 92.70 |
| | F1 | 4.62 | 2.67 | 4.85 |

# 7 结束语

垃圾邮件过滤是目前的一个热点问题,本文研究了基于粗糙集和决策树结合的垃圾邮件过滤技术,将邮件头的特征和邮件文本的特征结合进行分析,扩展了现有的条件属性,提出了一种 RS-DT 邮件分类方案,并进行了实验,给出结果及其与 SVM、朴素贝叶斯的比较,实验表明基于 RS-DT 的模型能有效降低错误识别率,尤其是将正常邮件判为垃圾邮件的比率。

(下转 184 页)

这里 $Z_{x_1}(\cdot)$ 和 $Z_{x_2}(\cdot)$ 是沿 $x_1$、$x_2$ 方向的一阶偏导数，$w_i$ 是待估计点周围的局部分析窗。梯度的局部主方向与这个矩阵的特征向量有关。

比较式(6)、(7)和(8)，似乎可以由局部协方差矩阵来估计矩阵 $\sum$ 及参数 $\rho$、$\sigma_1$、$\sigma_2$，以确定如图 1 所示的平均区域及各像素权值。然而协方差矩阵的估计结果可能是欠秩或不稳定的，在这种情况下难以直接求矩阵的逆。对于这个问题，有两种方法解决：(1) 使用矩阵算法中的迭代分解技术求解；(2) 使用局部多尺度技术来估计局部方向[5]。

该滤波器很好地满足了像素及权值选择的 3 个准则，兼顾了图像平滑去噪和边缘保护的问题。当然，其代价是计算开销的增加。

## 5 实验

为验证结合空间和像素距离加权的自适应高斯平滑滤波器的效果因素，以下分别进行了几个实验。

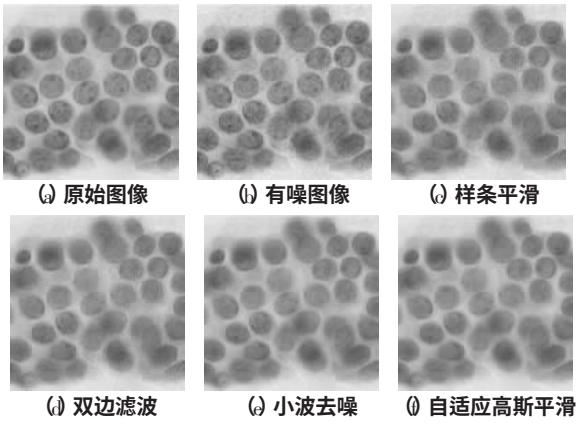
实验 1 前述三种算法的比较。图 2(a)是 256×256 的原始 Lena 图像，在其上加入均值为 0，方差 25 的高斯白噪声，得到图 2(b)的有噪图像。图 2(c)、(d)、(e)分别是经高斯平滑滤波、梯度倒数加权滤波、自适应高斯平滑滤波等三种平滑去噪方法后得到的结果。可以看出，图 2(e)中图像的恢复比较好。

(a)原始图像 (b)有噪图像 (c)高斯平滑滤波

(d)梯度倒数加权滤波 (e)自适应高斯平滑滤波

图 2 三种滤波去噪效果

实验 2 与其他方法的比较。这里对医学图像应用自适应高斯平滑滤波与几种常见去噪方法进行比较，实验图片大小为 256×256。实验中对图像加入的是 $SNR$=6 的高斯白噪声。自适应高斯平滑滤波与几种常见去噪方法结果见图 3。通过实验结果可以看到自适应高斯平滑滤波能够较好地对医学图像去噪。



(a)原始图像 (b)有噪图像 (c)样条平滑

(d)双边滤波 (e)小波去噪 (f)自适应高斯平滑

图 3 不同去噪方法的实验比较

## 6 结语

本文研究了结合空间和像素距离加权的自适应高斯平滑滤波器，其结合了高斯平滑滤波器和梯度倒数加权滤波器的特点，充分考虑了图像的局部空间距离和像素距离。因而，在降噪的同时，自适应地保留了图像的局部边缘特性。分析和实验显示该方法是有效的。

对于如何确定二维高斯函数的参数，以及如何简化确定局部边缘的方向的计算量，都是需要进一步研究和探讨的问题。

## 参考文献：


[1] 李弼程，彭天强，彭波.智能图像处理技术[M].北京：电子工业出版社，2004.

[2] 阮秋琦.数字图象处理[M].2 版.北京：电子工业出版社，2007.

[3] 余庆军，谢胜利.基于人类视觉系统的各向异性扩散图像平滑方法[J].电子学报，2004，32(1)：17-20.

[4] 魏丹，陈淑珍，陈彬，等.梯度倒数加权平滑算法的改进与实现[J].计算机应用研究，2005(3)：153-154.

[5] Feng X，Milanfar P.Multi-scale principal components analysis for image local orientation estimation[C]//Proceedings of the 36th Asilomar Conference on Signals，Systems and Computers，Pacific Grove，CA，November 2002.


(上接 140 页)

在今后的工作，将对以下方面进行进一步研究：

(1) 文本特征的选取，结合文本数据挖掘，选取更能反映邮件特征、有助于分类正常邮件和垃圾邮件的条件属性。

(2) 实验数据的规模问题。收集更具规模的数据集，尝试将此方法应用到更多的数据集中来进行实验和改进。

(3) 考虑将粗糙集与其他方法的结合以提高性能。

## 参考文献：


[1] Saharni M，Dumais S，Heckerman D，et al.A Bayesian approach to filtering junk e-mail[C]//Proceeding of AAAI Workshop on Ixaming for Text Categorization，1998：55-62.

[2] 李志君，王国胤，吴渝.基于 Rough Set 的电子邮件分类系统[J].计算机科学，2004(3)：58-60，66.

[3] Pawlak Z.Rough set theory and its applications to data analysis[J].Cybernetics and Systems，1998，29：661-688.

[4] 张文修.粗糙集理论与方法[M].北京：科学出版社，2002.

[5] 王国胤.Rough 集理论与知识获取[M].西安：西安交通大学出版社，2002.

[6] 李雄飞，李军.数据挖掘与知识发现[M].北京：高等教育出版社，2004.

[7] Fayyad U M，Irani K B.Multi-interval discrerization of continuous-valued attributes for classification learning[C]//Proceedings of the 13th International Joint Conference on Artificial Intelligence，Morgan Kaufmann，1994：1022-1027.

[8] Rosetta.http://www.idi.ntnu.no/~aleks/thesis/.

[9] Spam E-mail database.http://www.ics.uci.edu/~mlearn/MLRepository.html.

[10] 廖明涛，张德运，李金库.基于朴素贝叶斯和层次聚类的两阶段垃圾邮件过滤方法[J].微电子学与计算机，2007，24(8)：1-3，7.