电子邮件分类中的特征选择*)

陈超兰　张自力

(西南大学智能软件与软件工程重点实验室　重庆 400715)

摘　要　电子邮件是互联网的最重要应用之一,尽管给人们日常工作和生活带来很大便利,但也带来了一种令人讨厌的副产品——垃圾邮件。对邮件进行分类已成为当前的一个研究热点,而如何进行邮件特征选择,是邮件分类中一个基本也是很重要的问题。本文在分析比较几种用于邮件分类的典型特征选择方法基础上,提出一种新的结合了 Mitra's 算法和顺序前进搜索法优点的邮件特征选择方法。实验结果表明该方法能够改进邮件分类的准确率,验证了本文方法的有效性和可行性。

关键词　垃圾邮件,邮件分类,特征选择

Feature Selection in E-mail Classification

CHEN Chao-Lan　ZHANG Zi-Li

(Key Lal. of Intelligent Software & Software Engineering, Sowth—West China Lniversity, Chongqing 400715)

Abstract　E-mail is one of the most popular services of the Internet. E-mail has brought us great convenience in our daily work and life. It has brought us an annoying byproduct——Spam. How to classify incoming E-mails and filter spam has attracted much attention. One fundamental yet important issue in E-mail classification is how to select the appropriate features. Based on the analysis and comparison of several typical feature selection methods for E-mail classification, a new method is proposed, which combines both Mitra's and Sequential Forward Selection. Experimental result shows that the proposed method can improve the precision of E-mail classification.

Keywords　Spam, E-mail classifying, Feature selection

1　引言

互联网 70%以上的应用是电子邮件,其中垃圾邮件(SPAM)是指未经收件人请求而发送的电子邮件。垃圾邮件破坏了电子邮件的正常流通秩序,对社会的危害十分严重。据统计,全世界每天的电子邮件其中 10%以上为垃圾邮件。根据调查,2004 年全球垃圾邮件达到 3.3 万亿封(2003 年为 1.6 万亿封),因此而造成的损失为 1190 亿美元(2003 年为 580 亿美元)(http://star-techcentral.com.tech/story.asp)。据 Ferris 公司的最新研究报告称,2005 年垃圾邮件给全球的生产力造成的损失和其他反垃圾邮件的投资将达到 500 亿美元。垃圾邮件占用邮件服务器大量网络资源、系统资源、存储资源;垃圾邮件攻击会导致系统瘫痪、服务中断;各种垃圾广告邮件,阻碍正常通讯,使人厌烦,导致用户投诉,并极其容易诱发经济犯罪等。如何有效地控制垃圾邮件的蔓延,成为亟待解决的一个问题。

目前一般的垃圾邮件解决方法是安装智能过滤器,主要是通过一定的过滤规则(如黑名单[1]、关键词等)来对垃圾邮件进行过滤,其过滤过程实质是一个分类过程,即将分类为垃圾邮件的一类邮件过滤掉,这在很大程度上抑制了垃圾邮件的泛滥。但它同时也带来了一系列的问题,过滤器有可能会把非垃圾邮件当作垃圾邮件过滤掉,给用户造成很大的损失。所以,如何提高邮件分类的准确率就成为邮件过滤研究中的一个重点也是一个难点问题。

要提高邮件分类的准确率,目前研究较多的是采用不同的分类算法[2~4]。然而,对于不同的分类算法而言,对邮件特征属性进行特征选择都是最基础也是很重要的一步。不同的特征选择算法选出的特征属性子集差异很大,对最后的邮件分类准确率产生很大的影响。

针对这个问题,本文在分析比较几种用于邮件分类的典型特征选择方法基础上,提出了一种新的邮件特征选择方法。该方法结合了 Mitra's 算法和 SFS 算法的优点,克服了 SFS 算法一旦选入某个特征就不能删除的缺点,实验结果表明该方法能够提高邮件分类的准确率。

2　邮件分类的总体流程

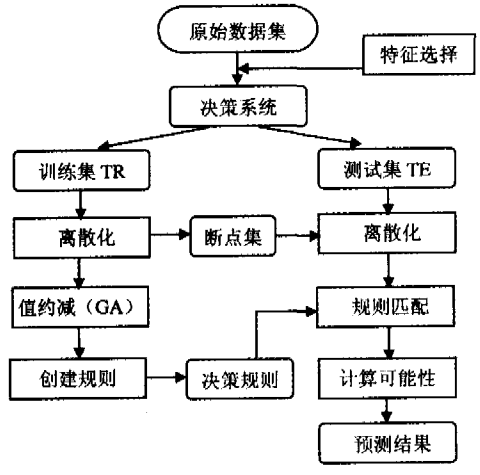

图 1　基于粗糙集邮件分类系统流程简图

*)受到重庆市自然科学基金资助。**陈超兰**　硕士研究生,主要研究方向:人工智能、邮件分类。**张自力**　博士,教授,主要研究方向:多代理系统,人工智能,混合智能系统等。

邮件的分类过程一般包括原始数据的输入、特征属性选择、构造决策系统。在决策系统构造过程中，主要包括训练过程和测试过程。训练过程通过一定的分类算法对数据集进行训练，获得决策规则，然后将决策规则应用到测试集，对测试集数据进行分类，完成测试过程，输出预测值。邮件分类的典型流程如图1所示[4]。

由图1可知，在构造决策系统之前就必须进行邮件特征选择，所以它是构建整个决策系统的基础。由于原始数据集中通常包含大量的属性，而将这些属性全部用来构建决策系统是不可能的。所以，特征属性子集选择的好坏将直接影响整个系统的分类准确率。

3 特征选择算法研究现状及在邮件分类中的相关应用

3.1 特征选择算法研究现状

特征选择是根据一定的评价准则从含有 N 个特征的集合中寻找出 $n(n<N)$ 个相关特征[7]。从原始特征集中选出最优特征子集是模式识别和机器学习等领域的一个关键问题，同时也是一个棘手问题。已证明，最优(最小)特征子集选择是NP难题[8]。

特征选择算法根据不同的评价准则，大体上可分为三类：过滤器模型，封装器模型以及混合模型[9]。当训练样本的类别已知，即监督的特征选择来说，一般是在错误率和特征子集的维数之间进行折中[10]。但该问题是属于NP难的问题，除了穷尽搜索之外，不能保证得到最优解，当原始特征数N较大时，如N>20，穷尽搜索实际上已经不可行。由于求最优解的计算量太大，人们一直在致力于寻找到较好次优解的算法。先后出现的顺序前进法(SFS)、顺序后退法(SBS)[7]等实际上都属于贪心一类的算法。Siedlecki和Sklansky[11]把遗传算法应用到特征选择中，获得了较好的效果，但遗传算法常出现过早收敛的问题。Pudil[12]等提出了顺序浮动前进法(SFFS)和顺序浮动后退法(SFBS)，其算法的解接近于最优解。

3.2 用于邮件分类的典型特征选择算法

特征选择算法虽然较多，但成功运用于邮件分类的，主要有如下几种：

(1)顺序前进法(SFS)。是最简单的自下而上搜索方法。设 U 为特征集，令 $X_0=\Phi$，设已选入 k 个特征构成一个大小为 k 的特征组 X_k。从未入选的D-k个特征中选择一个特征 x_1 加入 X_k，要求入选的特征组合在一起所得的评价准则 $J(X_{k+x_1})=\max_{x_i\in k-x_k} J(X_k+x_i)$ 令 $X_{k+1}+x_1$ 为新增特征组，直到特征数增加到指定的维数 n 为止。

上述作法的特点是每次选择一个特征加入特征组，所以算法的复杂度是多项式的，容易实现。其不足是一旦某个特征被选入，即使由于后面新加入的特征使它变为冗余也无法剔除。因此，算法不能保证得到的特征组是最优的，只能是次优的。

(2) 遗传算法(GA)。特征选择是一个组合优化问题，所以可以用遗传算法进行特征选择。基本遗传算法的步骤如下：

Step1 确定算法的运行参数：群体规模 M，终止迭代代数 T，交叉概率 P_c，变异概率 P_m。

Step2 给出初始化群体 $P(t)$，令进化代数 $t=0$，x_g 为任一个个体。

Step3 对 $P(t)$ 中每个个体计算适应度值 J，并将群体中最优解 x' 与 x_g 比较，如果 x' 的性能优于 x_g，则令 $x_g=x'$。

Step4 若终止条件被满足，则算法结束，x_g 是最后算法的结果。否则，转Step5。

Step5 从 $P(t)$ 中用比例算子选择个体构成配对库，对配对个体施行单点交叉操作，对产生的新个体施行变异操作，由此得到的新一代个体组成群体 $P(t+1)$。令 $t=t+1$，转Step3。

(3) 基于粗糙集理论的特征选择算法。粗糙集的属性约简，是在等价关系下，不改变系统信息的分类能力条件下，约去冗余属性，得到系统的满意约简乃至最小约简。这一理论的特点是：除了问题所需处理的数据之外，不需要额外提供任何外界信息或先验知识[13]，因此，在特征子集选择领域中的应用逐渐受到重视，已经成功地提出了一些基于粗糙集理论的相应算法[14]。

4 改进的Mitra's算法和SFS算法组合及实验结果

在各种电子邮件分类系统的研究中，特征属性的选择问题都是最基本也是非常重要的一环。在Zhao[4]等研究的基于粗糙集理论的电子邮件分类系统模型中，系统模型建立在决策系统的基础之上。在实验实现时，在该决策系统中，包含10个特征属性，1个决策属性。而从UCI(University of California, Irvine)机器学习数据库(www.ics.uci.edu/~mlearn/MLRepository.html)得来的实验数据中，共有4601个实例，每个实例分别由58个属性来描述，其特征数太大。在文[4]中，采用顺序前进法(SFS)来进行初始的特征选择，在分类系统中再用遗传算法进行属性约简。但是顺序前进法如前所述有它的缺点：一旦某个特征被选入，即使由于后面新加入的特征使它变为冗余也无法剔除。因此，算法不能保证得到的特征组是最优的，只能是次优的。

4.1 算法设计

为了得到更优的垃圾邮件特征子集，本文采用改进的Mitra's算法+SFS来进行邮件特征属性的选择。

1)第一步 用Mitra's算法来消除冗余特征。Mitra's算法是一种非搜索型，它是一种消除冗余特征算法，比其它那些通过搜索特征子集来完成特征选择的算法，具有相对低的时间复杂度，而且实现简单。在原算法的基础上，基于下面的分析做了一些修改：原算法在流程上存在一些缺陷，在每次删除冗余特征后，原算法没有重新对剩余的特征关联性进行排序，重新寻 r_i^k 找最小的特征 $F_{i'}$，从而使得流程不清晰，且存在误差。我们进行了一些修改，修改后的流程如下：

假设原始特征集为 $F=\{F_i, i=1,\cdots,N\}$，特征总数为 N，$\mathrm{dis}(F_i,F_j)$ 表示特征 F_i 与 F_j 间的距离(非关联性)，FR 为消除冗余的特征集，r_i^k 用来描述 F_j 与其在 FR 中的第 k 个最近邻特征之间的距离。

Step1：初始化，选取 $k\leqslant N-1$，$FR\leftarrow F$。

Step2：对所有特征，两两之间计算非关联性，构建一个上三角矩阵，其行列都代表特征，非零元素为特征之间的距离值，对每一个特征 $F_i\in FR$，计算 r_i^k。

Step3：保留 r_i^k 最小的特征 $F_{i'}$，并抛弃 F_i 的 k 个最近邻特征，LET$\varepsilon=r_i^k$。

Step4：IF $k>$cardinality(FR)-1：$k=$cardinality(FR)-1//cardinality(FR)表示FR的势，也就是维数。

Step5：IF $k=1$：GO TO Step9。

Step6：对每一个特征 $F_i \in FR$，计算 r_i^k，并寻找 r_i^k 最小的特征 $F_{i'}$。

Step7：WHILE $r_i^k > \varepsilon$ DO：
{$k=k-1$
$r_{i'}^k = \inf_{F_i \in R} r_i^k$
IF $k=1$；GO TO Step9}
END WHILE

Step8：GO TO Step3

Step9：输出 FR

2)第二步，用 SFS 算法选择满足合适维数的最优特征子集。将 Mitra's 算法的输出作为 SFS 算法的输入，采用 SFS 算法直到特征数增加到指定的维数 n 为止。这样可以在消除冗余特征的情况下，进一步降低特征维数，选出更优的特征子集。

4.2 实验结果

实验数据来自于 UCI 的垃圾邮件数据库，其中实例数为 4601(1813 个垃圾邮件=39.4%)。每个实例由 58 个属性来描述，其中条件属性 57 个，为连续值；决策属性一个，为名词性的类标签(1 表示垃圾邮件，0 表示非垃圾邮件)。在文[4]中，采用 SFS 算法来进行特征属性的选择，选出了 11 个特征属性。在我们的实验中，采用了 SFS、Mitra's、GeneticSearch、Mitra's+SFS 算法选择出不同的特征属性集，最后用 Weka 软件[15]中 NaiveBayes 分类算法来进行分类准确率等方面的比较。

根据 Mitra's 算法中 K 与所要达到的分类准确率(应该接近或好于原始特征集的分类准确率)，通过实验，取得较好的 K 值。各种算法通过实验选出特定维数的特征属性集，Mitra's+SFS 算法最终选出的特征属性为 10 个。主要包括 word_freq_remove('remove'在邮件中出现的频率)，word_freq_money，word_freq_free 等属性。实验最后对每一组特征属性集，分别用 Weka 软件中的 NaiveBayes 分类算法来对实例进行分类，采用交互验证法(10-fold-cross-validation)[16]作为评估方法，得到如下实验结果。

表 1 实验结果

算法	正确分类数	错误分类数	准确率	维数约减
SFS	4005	596	87.05%	81.03%
Mitra's(k=16)	3517	1084	76.44%	72.41%
GeneticSearch	3592	1009	78.07%	46.55%
Mita's+SFS(n=10)	4098	703	89.07%	82.76%

表 2 SFS 算法与 Mitra's+SFS 算法影响邮件分类查准率和查全率的比较

标准	非垃圾邮件查全率	非垃圾邮件查准率	垃圾邮件查全率	垃圾邮件准确率
SFS	95.6%	84.9%	91.6%	73.9%
Mitra's+SFS	92.2%	90%	84.3%	87.5%

从表 1 的实验结果可知：本文提出的特征属性选择方法，在邮件分类的特征属性选择方面，既有效地降低了特征维数又提高了分类的准确率。

从表 2 的比较结果可以看到，Mitra's+SFS 算法能够提高垃圾邮件和非垃圾邮件的的判定准确率。采用这种方法进行邮件特征属性的选择，与文[4]中直接采用 SFS 算法来进行垃圾邮件特征属性选择算法相比较，所增加的时间开销不大，却更能提高分类的准确率，能达到更好的分类效果。

结论 邮件分类是当前的一个研究热点问题，特征选择在邮件分类中起着基础和重要的作用。原始数据中的邮件特征属性集中存在大量的冗余和不相关特征，所以需要选择最优的特征子集来构建分类决策系统，然而最优特征子集选择算法是 NP 完全的。这些原因使得进一步研究邮件特征选择算法具有重要意义。本文在介绍几种常用特征选择算法的基础上，提出了改进的 Mitra's+SFS 的新方法。最后用实验结果证明通过该方法选出的邮件特征子集维数约减率高，应用在邮件分类中时，相比其它几种算法，使分类器具有更高的分类准确率，从而达到更好的分类效果。

参 考 文 献

1 陶卓彬，登元庆. 反垃圾邮件技术. 网络安全，2002

2 Drucker H, Member S, Wu Donghui, et al. Support Vector Machines for Spam Categorization. Transactions on Neural networks, 1999, 10(5)

3 丁文斌，李斌，罗浩. 基于改进贝叶斯的垃圾邮件过滤系统设计与实现. 计算机工程与应用，2005，18

4 Zhao Wenqing, Zhang Zili. An Email Classification Model Based on Rough Set Theory. The 2005 International Conference on Active Media Technology, 2005

5 王练，李云，汪血焰. 高维特征集选择模型研究. 重庆邮电学院学报，2005，17(1)

6 Parsons L, Haque E, Liu H. Subspace Clustering for High Dimensional Data: a Review. SIGKPO explorations, 2004, 6(1): 90～105

7 Molina L C, Belanceh L N. A Feature Selection Algorithm: a Survey and Experimental Evaluation. In: Proc. 2002 IEEE Intl. Conf. on Data mining, 2002. 306～313

8 陈彬，洪家荣，王亚东. 最优特征子集选择问题[J]. 计算机学报，1997，2(20)：133～138

9 Dash M, Liu H. Feature Selection for Classification. Intelligent Data Analysis, 1997. 131～152

10 张鸿宾，孙广煜. TABU 搜索在特征选择中的应用. 自动化学报，1999，7(4)，457～466

11 Liu H, Yu L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Trans. On Knowledge and Data Engineering, 2005, 17(3): 1～12

12 Pudil P, Novovicova, et al. Floating Search Methods for Feature Selection. Pattern Recognition Letters, 1994, 159110: 1119～1125

13 赵军，王国胤，吴中福，等. 基于粗集理论的特征子集选择算法. 计算机科学，2002，29(11)

14 王国胤. 粗糙集理论与知识获取. 西安：西安交通大学出版社，2001. 23～36

15 Witten I H, Frank E. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, 2000

16 Rivals I, Personnaz L. On cross-Validation for Model Selection. Neural computation, 1999, 11(4): 863～870