基于多 Bayes 网的垃圾邮件智能过滤研究*)

欧 阳 韩逢庆
(重庆工学院网络信息中心 重庆400050)

摘 要 在分析朴素 Bayes 方法用于垃圾邮件自动过滤中存在的一些问题基础上,提出了一种新的基于多 Bayes 网的垃圾邮件自动过滤方法。该方法利用多个 Bayes 网构成的多个分类器同时对邮件进行分类,当前邮件被认定是垃圾邮件当且仅当全部分类器都判断它为垃圾邮件。这种多个分类器同时工作及分类临界值的使用在一定程度上减少了将有用邮件误判为垃圾邮件的可能性。该方法还引入动态学习机制,在邮件分类过程中能够补充训练样本,满足不同用户的邮件分类标准。

关键词 电子邮件,垃圾邮件,多 Bayes 网

A Way to Filter Junk Mail Intelligently Based on Multi-Bayesian Networks

OU Yang HAN Feng-Qing
(Chongqing Institute of Technology, Chongqing 400050)

Abstract Discussing some problems in filtering junk mail with Bayesian networks, a new way for filtering junk mail intelligently based on multi-Bayesian networks is proposed. Multi-classifiers are come from multi-Bayesian networks and mail is processed by Multi-classifiers. A mail is judged Spam if and only if every judging result is Spam. This idea and critical value can reduce the error probability of classifying. The adopted dynamic learning is introduced in the new method that is the multi-classifiers can be supplemented training examples so that satisfy different demand.

Keywords E-mail, Spam, Multi-Bayesian networks

0 引言

现在电子邮件(E-mail)已成为信息交互的重要工具,人们用它交流思想、传输文件、发表意见等。据 IDC 调查,2000年全球日平均发送邮件超过100亿封,到2005年将达350亿封以上。随着 E-mail 的日益普及,我们注意到网络管理面临着新问题——垃圾邮件的泛滥。

所谓垃圾邮件主要有两类,一类是名目繁多的商业广告,另一类是非法团体为其政治、经济等目的而进行的"网络宣传"。后者的危害性显然远远大于前者。垃圾邮件耗费了有限的网络资源,反动邮件严重破坏了社会稳定。另外,它还侵犯了个人隐私,浪费了用户大量时间。所以垃圾邮件的智能分析、自动过滤是目前研究的一个热点。

目前邮件过滤的主要方法有如下三种:

(1)安全认证方法,也就是用户 A 向用户 B 发送邮件时,必须到用户 B 的邮件服务器上先进行登记,得到授权,否则邮件服务器拒绝接收。这样虽然有效地防止未经认证的用户发来邮件,具有很高的安全性,但影响了邮件的易用性;

(2)基于规则的方法,很多时候是基于关键词匹配的邮件过滤,虽然能够处理邮件头和正文,但是实质还是生硬的二值判断,局限在二维空间上进行处理,缺少可信度的知识,同时要求用户自己定义规则,对用户的素质要求高,用户需要花费很多时间定义自己的规则,如果用户的兴趣发生变化,规则也要进行很大的改变,另外规则的纯粹人工定制,可能考虑并不周全;

(3)基于文本分类的方法,可以采用 Bayes、LLSF、SVM、KNN 等等方法,实际上邮件过滤也是一类文本分类问题,即将邮件分类成"垃圾"(无用)和"有用"两类。按照这一思路,可以区分合法邮件和垃圾邮件。

目前邮件自动过滤中以 Bayes 方法[1,2]用得较多,这是因为它以概率论中的 Bayes 定理作为理论基础,计算简便,并且是有监督分类。

为了简化计算,朴素 Bayes 方法中假设各个分类特征是相对独立的,计算出邮件分别属于"垃圾"类和"有用"类的条件概率,判定邮件属于条件概率大的那一类。

本文针对邮件自动过滤中朴素 Bayes 方法[3~5]存在的一些问题,提出了一种基于多贝叶斯网的邮件自动过滤方法,并且进行了试验比较。本文第1节介绍朴素 Bayes 方法用于邮件自动过滤的基本思想;第2节指出其中存在的问题;第3节提出一种基于多贝叶斯网的邮件自动过滤方法和算法分析;第4节给出了邮件分类器的实现;最后是结束语。

1 邮件自动过滤的朴素 Bayes 方法

1.1 贝叶斯网

贝叶斯网是一种表达概率分布的有向无环图,在该图中的每一节点表示一随机变量,图中两节点若存在着一条弧,则表示这两节点对应的随机变量不独立,两节点间若没有弧,则说明这两个随机变量是相对独立的。网中任一节点 X_i 均有一相应的条件概率表 CPT(Conditional Probability Table),用以表示节点 X_i 在其父节点取各可能值时的条件概率。若节点 X_i 无父节点,则 X_i 的 CPT 为其先验概率分布。贝叶斯网的结构及各节点的 CPT 定义了网中各变量的概率分布。

*)基金项目:重庆市教委科技基金资助(030601)。欧 阳 硕士,副教授,主要从事网络计算研究。

1.2 贝叶斯分类器

贝叶斯分类器即是用于分类工作的贝叶斯网。该网中应包含一表示分类的节点 C，变量 C 的取值来自于类别集合 $\{C_1,C_2,\cdots,C_m\}$。节点 C 还有一组子节点 $X=(X_1,X_2,\cdots,X_n)$ 用于反映分类的特征。

对于这样的一贝叶斯分类器，若某一待分类的样本 D，其分类特征值为 $x=(x_1,x_2,\cdots,x_n)$，则样本 D 属于类别 C_k 的概率为 $P(C=C_k|X=x)$，因而样本 D 属于类别 C_k 的条件满足：

$$P(C=C_k|X=x)=\max\{P(C=C_i|X=x)|1\leqslant i\leqslant m\} \quad (1)$$

由贝叶斯公式

$$P(C=C_k|X=x)=\frac{P(X=x|C=C_k)P(C=C_k)}{P(X=x)} \quad (2)$$

式中的分母 $P(X=x)$ 和类别 C_k 无关，所以贝叶斯分类仅计算概率 $P(X=x|C=C_k)$ 和 $P(C=C_k)$。其中 $P(C=C_k)$ 一般可由领域专家的经验得到，而 $P(X=x|C=C_k)$ 的计算则要困难得多。特别是对特征数 n 较大，而特征变量之间相关程度较高时，其计算将极其费时。为简化计算，可假定各特征变量 X_i 是相对独立的，这就是朴素的贝叶斯分类器。因而有

$$P(X=x|C=C_k)=\prod_{i=1}^{n}P(X_i=x_i|C=C_k) \quad (3)$$

则

$$P(X=x|C=C_k)=\frac{P(C=C_k)\prod_{i=1}^{n}P(X_i=x_i|C=C_k)}{P(X=x)} \quad (4)$$

1.3 贝叶斯邮件过滤器

贝叶斯邮件过滤器是一种贝叶斯分类器，将邮件分成有用的和无用的两类（C_1：有用邮件类，C_2："垃圾"邮件类）。一个类的定义是通过设置它的样本集合来实现的，首先构造关键字集合 $A=\{x_1,x_2,\cdots,x_n\}$，该关键字集合包含了最能体现"垃圾"邮件特征的字、词（如"免费"、"Free"、"欢迎"等）或特殊的字符串（如"!!!!"、"$ $ $ $"），并且排除"的"、"了"、"the"等信息含量极低的常用字。

设某个邮件 E 对应的特征向量 $\boldsymbol{T}=\{t_1,t_2,\cdots,t_n\}$，其中

$$t_i=\begin{cases}1, & x_i \text{ 在邮件中出现}\\ 0, & \text{否则}\end{cases},i=1,2,\cdots,n$$

因而对于具有特征向量 T 的邮件 E，其为垃圾邮件的概率为：

$$P(C=C_2|X=T)=\frac{P(C=C_2)\prod_{i=1}^{n}P(X_i=t_i|C=C_2)}{P(X=T)} \quad (5)$$

有用邮件的概率为：

$$P(C=C_1|X=T)=\frac{P(C=C_1)\prod_{i=1}^{n}P(X_i=t_i|C=C_1)}{P(X=T)} \quad (6)$$

其中 $P(C=C_1)$，$P(C=C_2)$，$P(X_i=t_i|C=C_1)$，$P(X_i=t_i|C=C_2)$ 即贝叶斯网的条件概率表 CPT，均可通过从预分类的训练集中学习得到。则

$$\text{邮件 E}=\begin{cases}\text{"垃圾"邮件}, & \text{当 } P(C=C_1|X=T)<1/2\\ \text{有用邮件}, & \text{否则}\end{cases} \quad (7)$$

整个算法由分类器训练和分类工作两部分组成，分类器训练就是用给定类别（C_1类或者 C_2类）的训练集样本来训练分类器，得到贝叶斯网的条件概率表 CPT。分类工作是使用训练完成的分类器对未知类别邮件 E 进行分类，分类标准见(7)式。

2 朴素 Bayes 方法存在的问题

上述贝叶斯邮件过滤器采用监督学习方式，并且假定各特征变量 X_i 是相对独立的。其中存在以下几个问题：

(1)如果预分类的训练集样本不够充分，将会导致分类偏差较大。例如：关键字集合 A 包含某个关键字 k，即 $x_i=k$，并且包含 k 的预分类训练样本全部是 C_2("垃圾"邮件)，则一个包含 M 的新邮件 E 将被判为"垃圾"邮件。这是因为：$P(C=C_1|X=T)=0$，$P(C=C_2|X=T)=1$。

(2)由于体现"垃圾"邮件特征的字、词或特殊的字符串数量巨大，因此，关键字集合 A 非常庞大，每个邮件的特征向量 x 都是稀疏向量，使得预分类的训练集样本不充分，降低条件概率表 CPT 的可信度。

(3)在设计邮件过滤模块的时候，面对的一个十分棘手的问题是误承认率(定义：被误作为垃圾邮件的正常邮件在所有的垃圾邮件之中所占的比例大小)的大小，理想情况下误承认率要尽量地小甚至为0，但这是十分困难的。对于垃圾邮件的误承认就是指在对垃圾邮件过滤的时候，同时也将正常的合法的邮件过滤了，发生这种情况有时候用户是不允许的。对于用户来说有时宁愿多收到一些垃圾邮件也不希望被邮件过滤器过滤掉一封正常邮件。

(4)随着时间推移，关键字集合 A 中的字、词或特殊的字符串需要不断更新，即预分类的训练集也需要经常扩充，这种情况在贝叶斯邮件过滤器未加考虑。

(5)如果用户认定邮件 E 是"垃圾"邮件，且对应 A 的特征向量全为零，则分类系统将其判为有用邮件。但如果这个邮件多次重复出现，则每次都会被分类系统误判，会导致用户无法容忍。出现此类情况时，分类器的分类规则应该调整。

(6)除了反动邮件外，不同的用户区分"垃圾"邮件的标准有所区别，对同一个邮件，有用户认为是正常邮件，可能另外的用户认为是"垃圾"邮件，上述贝叶斯邮件过滤器无法满足这种差异。

这几个问题将会在接下来的第3节一一解决。

3 基于多贝叶斯网的邮件过滤器

3.1 多贝叶斯网邮件过滤器

本节提出一种新的基于多 Bayes 网的邮件过滤器，其基本思想如下：

为了克服问题(1)，在选择预分类训练样本时应该使包含 k 的预分类训练样本同时有 C_1类和 C_2类邮件。

对问题(2)，根据关键字的相关程度将可能的"垃圾"邮件预分成若干类，即关键字集合 A 分成不相交的子集 $A_1,A_2,\cdots,A_k$，每个 A_i 对应一个贝叶斯网 BN_i、预分类训练样本集合 B_i 及其条件概率表 CPT_i。

对问题(3)，考虑到将一有用邮件分类成"垃圾"邮件的危害要比将"垃圾"邮件分类成有用邮件的危害大得多，所以可设定分类临界值 $\lambda<0.5$作为"垃圾"邮件和有用邮件的边界，此时邮件按如下方式分类：

$$\text{邮件 } E=\begin{cases}\text{"垃圾"邮件}, & \text{当 } P(C=C_1|X=T)<\lambda\\ \text{有用邮件}, & \text{否则}\end{cases} \quad (8)$$

改进后的分类算法仍然包括分类器训练、分类工作和分类器更新学习三部分，流程如下：

1)分类器训练　根据"垃圾"邮件的性质以及关键字的相关程度将关键字集合 A 分成不相交的若干类，即子集 $A_1,A_2,\cdots,A_k$，每个 A_i 对应预分类训练样本集合 B_i 和一个贝叶斯网 BN_i，用训练集 B_i 中的样本来训练分类器，得到贝叶斯网 BN_i 的条件概率表 CPT_i，分类器训练完成后和每个用户的邮箱绑

定在一起。

2)分类工作　使用训练完成的多个贝叶斯网 BN, 分类器同时对未知类别邮件 E 进行分类，过程如下：

对于每个 A_i，$i=1,2,\cdots,k$，计算对应的特征向量 T_i，令 $M=\{i|T_i\neq 0\}$，则有2种情形：M 非空、M 空。当 M 非空时，则对 M 中的任意 I，按照(8)式用贝叶斯网 BN, 对邮件 E 进行分类，分类标准如下：如果对 M 中的任意 I，BN, 对邮件 E 都判定为"垃圾"邮件，则判定邮件 E 为"垃圾"邮件，否则为有用邮件。如果 M 是空集，即邮件 E 对应每个 A_i 的特征向量全为零，则分类系统应该将其判为有用邮件。

3)分类器更新学习　如果分类系统判定某邮件 E 为有用邮件，但经用户 A 认定是"垃圾"邮件，则需要对绑定在用户 A 上的分类器进行更新学习。

如果 M 是空集，此时先由用户指定该邮件的体现"垃圾"邮件特征的字、词或特殊的字符串，系统管理员对其校正。然后分类系统中加入一个新的关键字集合 A_{k+1}，邮件 E 就作为 A_{k+1} 的预分类训练集样本。

如果 M 非空，对 M 中的任意 I，只要 BN, 对邮件 E 都判定为有用邮件，将邮件 E 加入到 BN, 的训练样本集合 B_i，调整 CPT$_i$。

3.2　算法分析

3.2.1　*训练样本的饱和度*　设 A 包含有 n 个关键字特征，对应的预分类训练集样本数量 N，定义训练样本的饱和度 $d(A,n)=N/2^n$，一般地，训练样本的饱和度 $d(A,n)\leqslant 1$，饱和度越大，则训练样本越充分。

新算法将一个分类器分成多个分类器后，每个分类器的饱和度比改进前都有提高，这是因为：

定理1　每个 A_i 包含有 n_i 个关键字特征，对应的预分类训练集样本数量 N_i，在 N_i 相差不多的情况下，有 $d(A_i,n_i)>d(A,n)$。

证明略。

3.2.2　*分类器更新学习*　如果分类系统判定某邮件 E 为有用邮件，但经用户认定是"垃圾"邮件，则需要对分类器更新学习。

若 BN, 对邮件 E 判定为有用邮件，但经用户认定是"垃圾"邮件，则将邮件 E 加入到 BN, 的训练样本集合 B_i，调整 CPT$_i$。

设邮件 E 加入 B_i 之前有：

$$P(X_i=t_i|C=C_2)=p_1/q_1<1,\ P(C=C_2)=p_2/q_2<1$$

将邮件 E 加入 B_i 之后，$P(X_i=t_i|C=C_2)'=\dfrac{1+p_1}{1+q_1}$，$P(C=C_2)'=\dfrac{1+p_2}{1+q_2}$

则

$$P(X_i=t_i|C=C_2)<P(X_i=t_i|C=C_2)',$$

$$P(C=C_2)<P(C=C_2)'$$

所以当邮件 E 加入 B_i 之后，由 BN, 计算的 $P(C=C_2|X=T)$ 增加，重复加入若干次后，有

$$P(C=C_2|X=T)\geqslant 1-\lambda$$

4　邮件分类器的实现

按上述思想，我们设计了基于多贝叶斯网的邮件过滤器系统，该系统总体结构如图1所示，其中包括：

训练样本集：其中的电子邮件已经人工分类，用作训练的样本。分类器训练：根据样本集中的邮件计算的条件概率表 CPT$_i$。邮件过滤器：若干个训练完成的贝叶斯分类器。目前，该系统已经实现，并且取得了较好的分类效果。

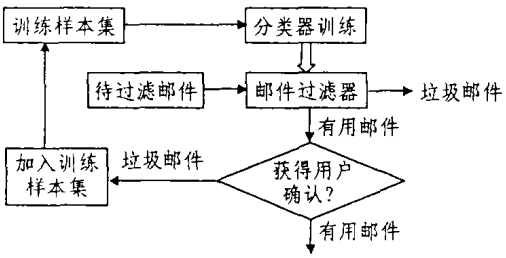

图1

结束语　本文提出了一种新的基于多 Bayes 网的垃圾邮件智能过滤算法，新的过滤算法采用多过滤器方式，提高了训练样本的饱和度。在过滤器的工作过程中能够根据用户需要动态补充训练样本，实现分类器的再训练，能够满足不同用户的邮件分类标准。用于"垃圾"邮件的过滤效果较好。

参 考 文 献

1　Pearl J. Fusion, propagation, and structuring in belief networks [J]. Artificial Intelligence, 1986,29:241～288

2　Friedman N. Bayesian network classifier[J]. Machine Learning, 1997,29:131～163

3　陈华辉，等. 一种基于贝叶斯网的"垃圾"邮件过滤器. 微机发展，2000,13(4)

4　刘洋，等. 垃圾邮件的智能过滤系统设计探讨. 微机发展，2003,13(4)

5　周颜军，等. 基于贝叶斯网络的分类器研究. 东北师大学报自然科学版，2003,35(2)