贝叶斯网络模型在反垃圾邮件中的应用①

Bayesian Network in Anti-Spam System

张兵利　裴亚辉　(河南科技大学 电子信息工程学院 河南 洛阳 471003)

摘　要： 近些年，网络上的垃圾邮件肆意横行，令人深恶痛绝，因此反垃圾邮件就成了亟待解决的问题。而贝叶斯网络理论的研究为反垃圾邮件指出了一个明确方向。贝叶斯推断理论提供一种概率手段，为数据建模提供了个统一的框架，而且它为算法的分析提供了理论基础。本文在对贝叶斯网络分类模型形式化描述的基础上，设计了一个基于贝叶斯分类器的反垃圾邮件模型。实验证明，利用基于贝叶斯分类器的反垃圾邮件模型对邮件进行分类时可以获得较高的准确率和不错的查全率。

关键词： 数据挖掘 信度网 贝叶斯网络 贝叶斯分类器 反垃圾邮件

J.Pearl 在 1986 年提出的一种基于概率的不确定推理网络，1988 年正式提出了贝叶斯网络[1]。贝叶斯分类模型建立在经典的贝叶斯概率理论与贝叶斯网络技术基础上。它是从传统的统计学中分离出来的，对不确定性问题进行处理的一个有力工具。

1　贝叶斯网络模型的描述

贝叶斯网络(BN)，又称为信度网，由一个有向无环图(Directed Acylic Graph, DAG)和条件概率表(Conditional Probability Table, CPT)组成[2]。

贝叶斯网络分类模型(BNC)的形式化的描述如下：

n 元随机变量 $X=\{X_1,X_2,...,X_n\}$ 的贝叶斯网络模型是一个二元组 $B=(B_s,B_p)$。$Bs=(X,E)$ 是一有向无环图 (directed acrylic graph, DAG)，其中 $X=\{X_1,X_2,...,X_n\}$ 为结点集，每个结点可看成取离散或连续值的变量(本文限定其只取离散值)；E 是有向边的集合，每条边表示两结点间依赖关系，依赖程度由条件概率参数决定。称 Bs 为 BN 模型网络结构。$Bp=\{P(X_i/\prod X_i),X_i\in X\}$ 是贝叶斯网络模型的一组条件概率分布的集合。在各结点取离散值的情况下，B_P 为一组条件概率表(conditional probability tables, CPTs)的集合。∏Xi 是在 Bs 中 Xi 所有父结点的集合，表示结点 Xi，在其父结点某一取值组合状态下的条件概率分布。这说明，在贝叶斯网络模型中，结点的取值依赖于其父结点的取值状态。

这里，学习贝叶斯网络的问题描述为：给定 Xi 中的一组实例构成的训练集合 $D=X=\{X_1,X_2,...,X_n\}$，找到一个与 D 匹配最好的网络 B。这样，学习贝叶斯网络的问题转化为优化问题。这时类变量和属性变量不加区别。

实际处理这个问题的方法是在可能的网络构成的空间中进行启发式搜索。搜索成功的关键是确定一个合理的评分函数，评价网络对训练数据的匹配程度，以指导搜索。

有两种主要的评分函数[3]：贝叶斯评分函数和最小描述长度原理(MDL：minimal description length)评分函数。它们是渐进正确的，即随着样本数目的增加，得分最高的网络将任意逼近样本的概率分布。

2　构造贝叶斯网络的方式

一般情况下，构造贝叶斯网有三种不同的方式：

①由领域专家确定贝叶斯网的变量(有时也称为影响因子)节点，然后通过专家的知识来确定贝叶斯网络的结构，并指定它的分布参数。这种方式构造的贝叶斯网完全在专家的指导下进行，由于人类获得知识的有限性，导致构建的网络与实践中积累下的数据具有很大的偏差[4]。

②由领域专家确定贝叶斯网络的节点，通过大量

① 基金项目:国家自然科学基金项目(60743008);河南科技大学教改项目(G2003-21);河南科技大学实验技术开发基金项目(SY0304016)
收稿时间:2008-08-12

的训练数据，来学习贝叶斯网的结构和参数。这种方式完全是一种数据驱动的方法，具有很强的适应性，而且随着人工智能、数据挖掘和机器学习的不断发展，使得这种方法成为可能。如何从数据中学习贝叶斯网的结构和参数，已经成为贝叶斯网络研究的热点。

③由领域专家确定贝叶斯网络的节点，通过专家的知识来指定网络的结构，而通过机器学习的方法从数据中学习网络的参数。这种方式实际上是前两种方式的折衷，当领域中变量之间的关系较明显的情况下，这种方法能大大提高学习的效率。

但不管是那种模型，建立贝叶斯网络，大多都需要经过下面几步：

第一步：建立模型有关的变量及其解释，需要做下面的工作：

①确定模型的目标，即确定相关问题的解释；

②确定与问题有关的观测值，并确定其中值得建立模型的子集；

③将这些观测值组织成互不相容的变量。

第二步：建立一个条件独立的有向无环图。

第三步：局部概率分布。在离散的情形，需要为每一个变量的各个父节点的状态指派一个分布。

以上各步可能交叉进行，不是简单的顺序进行可以完成的。

3　贝叶斯网络模型在反垃圾邮件中的应用

当我们在贝叶斯网络中把其中代表类别变量的节点作为根节点，其余所有变量都作为它的子节点时，贝叶斯网络就变成了分类器[5]，这恰恰可以用来解决目前令人头疼的垃圾邮件问题。

根据以上对贝叶斯分类器的工作原理的讨论，我们设计了一个基于贝叶斯分类器的反垃圾邮件模型。如图 1。

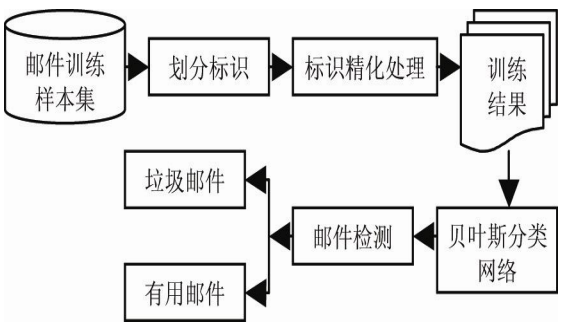

图 1　基于贝叶斯分类器反垃圾邮件模型

这个基于贝叶斯网络的反垃圾邮件模型是以电子邮件为研究对象，通过对电子邮件样本集进行训练，通过标识的划分、精化处理获得能区分是否是垃圾邮件的特征模式(训练结果)，再以此模式为基础得到贝叶斯分类器从而对邮件进行检测，找到有用邮件，过滤掉垃圾邮件。

3.1 对邮件进行训练

对邮件进行训练之前，先把邮件人工分成有用邮件和垃圾邮件两类，这样可以避免把垃圾邮件作为有用邮件学习或把有用邮件作为垃圾邮件学习。对邮件的训练，我们是对发件人地址、主题和正文三个部分分别学习，统计标识在这三个部分中出现的次数。

主要步骤包括三部分：首先对标识(token)进行划分。我们采用的方法是将收件人 To，发件人 From，主题 Subject 分别进行统计，而不是使用“Subject*free”的方式。其次是对标识数量进行精化处理，找出非用词并放入 stop list(非用词表)中，保留其他的标识以备后续工作使用。最后是读入待训练的垃圾邮件和有用邮件文件，分别对垃圾邮件和有用邮件进行处理并将统计的结果分别存入三个文本文件中，以便比较时使用。

3.2 对邮件进行检测

第一步： 标识出现的概率的计算

读入待检测邮件，按照标识划分的方法，分别取出 Body,From,Subject 部分中的除非用词表中的标识外的所有标识，并从训练结果中取出该标识在有用邮件和垃圾邮件中出现的次数，并计算：

$b(w)$=包含标识 w 的垃圾邮件数/总的垃圾邮件数　(1)

$g(w)$=包含标识 w 的有用邮件数/总的有用邮件数　(2)

$$p(w)=b(w)/(b(w)+B(w)) \quad (3)$$

这样计算出来的 $p(w)$可以粗略地表示随机选择的包含标识 w 的邮件是垃圾邮件的概率。

第二步：将单个概率组合成一个整体

一封邮件是由多个标识构成，对于每个标识，我们用概率来表示它在垃圾邮件中出现的概率，由此可知，一封邮件可由一组概率来表示，作为一个整体，我们将这些单个概率通过不同的算法组合起来，使之形成一个整体，当其值大于阈值时说明邮件为垃圾邮件，当其值小于阈值时说明邮件为有用邮件。

3.3 实验结果及评估

在本实验中，我们收集了一些公共邮件样本集，

包括 PUI 语料、SpamAssassin 语料和 CDSCE 中文语料，共计 5857 封邮件，其中垃圾邮件 4773 封，合法邮件 1084 封。我们将整个语料随机分成了均等的 4 组，并对各个组中的邮件进行分类并对其结果进行评估，评估的标志主要是分类的准确程度，分类准确程度的参照物是通过算法判断后对邮件的分类结果与人工分类结果的接近程度，算法的分类结果与人工分类结果越相近，分类的准确程度就越高，这里使用两个指标对结果进行评估:

①查全率(recall)：是人工分类结果应有的文本中分类系统吻合的文本所占的比率，其数学公式可以表示如下:

$$查全率=\frac{正确过滤掉的邮件数}{应过滤掉的垃圾邮件数} \quad (4)$$

其数值越高，表示漏网的垃圾邮件就越少。

②准确率(precision)：是所有判断的邮件中与人工分类结果吻合的邮件所占的比率。其数学公式可以表示如下：

$$准确率=\frac{正确过滤掉的邮件数}{实际过滤掉的邮件数} \quad (5)$$

其数值越高，表示将合法邮件误判为垃圾邮件的可能性越小。

我们的实验结果如表 1：

表 1　实验结果统计

组别	应该过滤掉的垃圾邮件数	实际过滤的邮件数	正确过滤掉的邮件数	查全率	准确率
1	1078	1085	1047	0.9712	0.9650
2	1245	1259	1192	0.9574	0.9468
3	1156	1146	1093	0.9455	0.9536
4	1294	1206	1101	0.8509	0.9129

从实验结果看出，利用基于贝叶斯分类器的反垃圾邮件模型对邮件进行分类时可以获得较高准确率，但查全率有的较高，有的却比较低。查全率低意味着有用邮件被误判为垃圾邮件的可能性较大，这将给用户带来更大的损失，因为相对于用户来说，有用邮件比垃圾邮件要重要得多。

查全率有时高有时低的主要原因是：待检测邮件的标识特征如果太少，就不能全面表现出邮件的内容，造成区分度不够；如果标识特征太多的话，又会有一些无关的特征，也就是引入了分类噪声。并且实验用的贝叶斯方法的前提是假设特征之间相互独立，特征数量增加后，特征之间相互依赖的机会增大，独立性变小，因为邮件文本内容中各个词汇之间本来就不是完全独立的。

4　贝叶斯网络模型在反垃圾邮件中的应用

目前，常用的反垃圾邮件分类方法有：关键字检测、基于规则评分的过滤技术、贝叶斯过滤、决策树、粗糙集等，而其中贝叶斯过滤与关键字检测的应用最为广泛和普及。

与关键字检测相比较贝叶斯过滤有以下优势：

①贝叶斯方法是把邮件信息作为一个整体来考虑的。换句话说，贝叶斯过滤是一个很智能的方法，它检测邮件信息的各个方面。与之相对的是，关键字检测只是通过一个简单的词判断一个邮件是否是垃圾邮件。

②贝叶斯过滤器是自适应的。通过对新的垃圾邮件和新的外发有效邮件进行检测，贝叶斯过滤器可以不断地调整和适应新的垃圾邮件技术[6]。

③贝叶斯方法可以满足国际化的需要，它可以适用于任何语言。而大多数关键字列表只使用于英语，而对于非英语的地方是毫无用处的。贝叶斯过滤中的智能化技术考虑到某些语言表达方式差异性的现象。从而使得它可以捕捉到更多的垃圾邮件。

④与关键字检测相比，贝叶斯过滤是难以被欺骗的[7]。

贝叶斯算法之所以有以上优势，是因为贝叶斯过滤器不必预先设定规则，不需要分析邮件句法或内容含义，而是用户根据自己所接收的垃圾邮件和非垃圾邮件的统计数据来创建的，这意味着垃圾邮件发送者无法猜测出过滤器是如何配置的，从而可以有效阻止垃圾邮件。

5　小结

近些年，网络上的垃圾邮件肆意横行，令人深恶痛绝，因此反垃圾邮件就成了亟待解决的问题。而贝叶斯网络理论的研究为反垃圾邮件指出了一个明确方向。贝叶斯推断理论提供一种概率手段，为数据建模提供了个统一的框架，而且它为算法的分析提供了理

论基础。尽管关于贝叶斯网的理论研究还很不完善，应用研究还处于起步阶段，但在许多领域中已显现出令人瞩目的效果，可以预见随着技术的进步，贝叶斯网模型将越来越发挥极其重要的作用。

参考文献

1 Graham P. Better Bayesian Filtering. Proceedings of SpamConference.Cambridge, 2003:112－118.

2 刘伟娜,霍利民,张立国.贝叶斯网络精确推理算法的研究.微计算机信息, 2006,22(9):92－94.

3 林士敏,田凤占,陆玉昌.贝叶斯网络的建造及其在数据采掘中的应用.清华大学学报(自然科学版), 2001,41(1):49－52.

4 周颜军,王双成,王辉.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):21－27.

5 Fu ZW, Sarac I. A Computational Study of Naive Bayesian Learning in Anti-spam Management. Structural, Syntactic, and statistical Pattern Recognition, 2004:824－830.

6 Oezguer L, Guengoer T, Guergen F. Spam Mail Detection Using Artificial Neural Network and Bayesian Filter. International Conference on Intelligent Data Engineering and Automated Learning(IDEAL 2004). 20040825-20040827. Exeter. GB. 2004: 505－510.

7 何绍华.运用贝叶斯方法过滤垃圾邮件.现代计算机, 2004,(5):30－33.