基于信息熵和决策分类技术的邮件识别研究

李 洋 赵 骅
(重庆大学经济与工商管理学院 重庆 400030)

摘 要 本文通过对电子邮件头信息和正文内容进行离散和特征化处理,将一封电子邮件用向量组的方式加以表达;进而使用基于信息熵的决策树分类技术构建一种垃圾邮件分类识别模型;最后通过实验对该模型做了相关的检验和测试。实验证明,该模型经过一定数量的垃圾邮件和正常邮件的对比学习后,能够进行垃圾邮件的识别,具有较好的效果。

关键词 决策树,信息增益,数据挖掘,垃圾邮件

Study on Email Classifying Based on Information Entropy and Determination Tree Technique

LI Yang ZHAO Hua
(College of Economics and Business Administration, Chongqing University, Chongqing 400030)

Abstract By disperseing and charactering an email, this paper uses a group of vectors to express an email. And bring forward a determination tree classifying model base on information entropy. And then followed with some experiments and tests. The results proved that the model can find out how to identify the new spams by learning and training from the spams and normals. So it shows that our model and method work well.

Keywords Data mining, Information entropy, Determination tree, Spam

1 引言

随着互联网应用的发展,电子邮件得到广泛的应用,已成为 Internet 网上最基本的服务之一,用户可以通过电子邮件与远程用户进行经济、方便和快捷的信息交流。然而,就在电于邮件逐渐成为一种不可缺少的重要信息交流工具的同时,也正在成为一种商业广告手段。用户在收到有用信息的同时,还必须花费大量时间和精力来处理从 Internet 上收到的各种各样的广告邮件,这些各种各样广告邮件就是人们所说的"垃圾"邮件,其主要来源为:商业广告,站点宣传,某些网络杂志,连环信的 E-mail,用垃圾邮件进行恶意攻击,以及一些政治宣传和淫秽内容的邮件等。目前国内外研究的垃圾邮件过滤技术有:基于规则的评定过滤技术、"白名单"验证过滤技术、分布式适应性黑名单过滤技术以及贝叶斯单词分布过滤技术等,但这些方法缺少自学习的能力。

2 基于信息熵和决策分类技术的邮件分类识别

2.1 模型的设计

本模型采用决策树学习算法,对邮件训练样本库中的邮件进行学习分类,提取出对垃圾邮件的识别规则。模型结构如图 1 所示。

邮件样本由垃圾邮件样本和正常邮件样本共同组成,数据挖掘学习方法就是通过统计分析垃圾邮件和正常邮件各自的结构特征、文本特征等信息来实现对正常邮件和垃圾邮件的分类识别。因此为了达到最好的挖掘效果,学习训练样本邮件的正常邮件与垃圾邮件的组成比例应该尽可能地贴近真实,反映真实的情况。

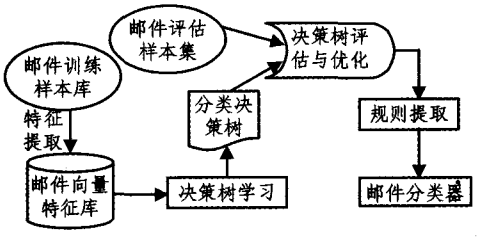

图 1 模型框架结构

2.2 样本邮件的特征化及向量表示

RFC822 规定了电子邮件在网络中传输的基本格式,由 20 多个常用字段、字段值和正文组成构成。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。电子邮件属于半结构化的文本信息,其中的字段标志及取值提供了邮件从发送、转发到最后投递过程中的许多信息,如:发送者地址、收信人地址、发信时间、发送程序、编码的格式等。这些信息能够用于帮助判断一封邮件是否为垃圾邮件。为方便处理这些信息,本文采用了向量空间模型(VSM) $(w_1, w_2, \cdots, w_n, C)$来表示一封样本邮件。向量表示中的属性 $w_1, w_2, \cdots, w_n$ 为有助于区别正常邮件与垃圾邮件的 n 个特征属性,属性 C 则为样本邮件的分类属性。分类属性 C 的取值定义为:正常邮件、垃圾邮件和无法确认。本文将一封电子邮件离散化后用 15 个特征属性来表示。特征属性取值及取值含义如表 1 所示。

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后(其中在提取正文特征时需要经过邮件解码),将这些信息以数据库字段的方式存入邮件向量数据库。

李 洋 硕士,主要从事计算机网络研究。

表 1 邮件特征属性

特征变量及说明		特征属性含义(取值)
W_1	收件人地址	本域(1)非本域(2)无域名(3)
W_2	HELO 域地址反向解析与发送 IP 比较	符合同属一 B 类地址(1)不符合(2)无法识别(3)
W_3	邮件正文格式(Content-Type 字段)	text/plain(1) text/html(2)
…	…	…
W_{14}	邮件发送方式	SMTP 方式(1)HTTP 方式(2)
W_{13}	正文 http 连接数	无(1)1～2 个(2)≥3(3)
C	邮件的分类属性	正常邮件(1) 垃圾邮件(2)无法确认(3)

2.4 基于信息熵的决策树邮件分类识别算法

基于信息熵的决策树学习算法采用贪心算法，自顶向下递归的构造决策树。树以代表训练样本的单个节点开始；如果样本都在同一个类，则该节点成为树叶，并用该类标记该节点；否则，算法使用称为信息增益的基于熵的度量作为启发信息，选择能够最好地将样本分类的属性；该属性成为该节点的测试属性。对测试属性的每个已知的值，创建一个分枝，并据此划分样本。算法使用同样的过程，递归地形成每个划分上的样本判定树。一旦一个属性出现在一个节点上，就不必考虑该节点的任何后代上。递归划分的步骤仅当下列条件之一成立时停止：① 给定节点的所有样本属于同一类。② 没有剩余属性可以用来进一步划分样本，在此情况下，使用多数表决。并将给定的节点转化成树叶，并用样本集 samples 中的多数所在的类标记。③ 分枝 test_attribute=a_i 没有样本，在这种情况下，以 samples 中的多数所在的类创建一个树叶。

算法开始的时候只有一颗空的决策树，并不知道如何根据属性将实例进行分类，所要做的就是根据训练实例集构造决策树来预测如何根据属性对整个实例空间进行划分。设此时训练实例集为 X，目的是将训练实例分为 n 类。设属于第 I 类的训练实例个数是 C_i，X 中总的训练实例个数为 $|X|$，若记一个实例属于第 i 类的概率为 $P(C_i)$，则：

$$P(C_i)=\frac{C_i}{|X|}$$

此时决策树对划分 C 的不确定程度为：$H(X,C)=-\sum P(C_i)\log P(C_i)$，简记为 $H(X)$。

$$H(X/a)=-\sum_i\sum_j P(C_i;\ a=a_j)\log P(C_i/a=a_j)$$
$$=-\sum_j P(a=a_j)\sum_i P(C_i/a=a_j)\log P(C_i/a=a_j)$$

决策树学习的过程就是使得决策树对划分的不确定程度逐渐减小的过程。若选择测试属性 a 进行测试，在得知 $a=a_j$ 的情况下属于第 i 类的实例个数为 C_{ij} 个。记 $P(C_i/a=a_j)=\frac{C_{ij}}{|X|}$，即 $P(C_i/a=a_j)$ 为在测试属性 a 的取值为 a_j 时它属于第 i 类的概率。此时决策树对分类的不确定程度就是训练实例集对属性 X 的条件熵：

$$H(X_j)=-\sum_i P(C_i/a=a_j)\log P(C_i/a=a_j)$$

而在选择测试属性 a 后伸出的每个 $a=a_j$ 叶节点 X_j 对于分类信息的信息熵为：

$$H(X/a)=\sum_j P(a=a_j)H(x_j)$$

这样就确定了属性 a 对于分类提供的信息量为 $I(X;a)$：

$$I(X;a)=H(X)-H(X/a)$$

如果 $H(X/a)$ 的值越小则 $I(X;a)$ 的值越大，说明选择测试属性 a 对于分类提供的信息越大，选择 a 之后对于分类的不确定程度越小，所以在该节点上应该选者测试属性 a 进行划分。邮件分类决策树学习算法如下所示，其中：*samples* 为训练样本，由邮件特征向量属性表示；*attribute_list* 为候选属性的集合为。*Generate_decision_tree* 为决策树分类学习函数。

算法

输入：*samples*；*attribute_list*；
输出：邮件分类决策树。
① 创建结点 N；
② if *samples* 都在同一个类 C then 返回 N 作为叶结点，以类 C 标记；
③ if *attribute_list* 为空 then 返回 N 作为叶结点，标记为 *samples* 中的多数类；
④ 选择 *attribute_list* 中具有最高信息增益的属性 *test_attribute*；
⑤ 标记结点 N 为 *test_attribute*；
⑥ for each *test_attribute* 中的已知属性值 a_i，由结点 N 长出一个条件为 *test_attribute*=a_i 的分枝；
⑦ 设 S_i 是 *samples* 中 *test_attribute*=a_i 的样本的集合；
⑧ if S_i 为空 then 加上一个树叶，标记为“无法确认”类；
⑨ else 加上一个由 Generate_decision_tree(S_i，attribute_lis - test_attribute)返回的节点。

2.5 邮件决策分类树的评估与优化

上述算法通过对邮件训练样本库学习得到一颗邮件分类决策树，此外还需对该分类树进行测试评估，并通过邮件测试样本库的测试与评估对决策分类树进行修改和优化。测试评估时，对学习生成的决策分类树的每个分类结果不是“无法确认”的叶结点都设置一个二维变量(K_1，K_2)，变量 K_1 用于记录该点分类正确的测试邮件数目，变量 K_2 则记录在该点分类错误的测试邮件数目。并计算该点的分类错误率 error=$K_2/(K_1+K_2)$，对于那些分类结果为“垃圾邮件”，并且错误率大于可接受阀值的修改该点分类结果为“正常邮件”。

2.6 邮件分类规则的提取

邮件分类器在对邮件的分类识别和过滤时，将学习得到的分类决策树转化为分类规则，使用该规则对垃圾邮件进行过滤识别。由分类决策树提取的规则是以 IF-THEN 的形式所表示。

提取规则时，从树的树根到树叶的每条路径创建一个规则，该规则是一个合取项集，路径上的每个属性一值对应于规则的一个合取项。这些合取项构成规则的前件(“IF”部分)。沿每条路径最终所到达的叶节点，便是该规则对邮件的分类识别的预测，叶节点的分类属性预测形成规则的后件(“THEN”部分)。

2.7 实验测试

本文做了多次实验来测试模型的效果，训练样本通过学校网络中心的邮件服务器收集，其中训练样本 2000 封(包括正常邮件和垃圾邮件)、测试样本 1000 封，并利用重庆大学工商管理学院局域网为测试环境，在网关上对进出学院的邮件进行监听收集，再用于识别测试。共做了 20 次测试，其中最好情况率为 72%，最坏情况为 37%，平均识别率为 57%。测试结果如图 2 所示。

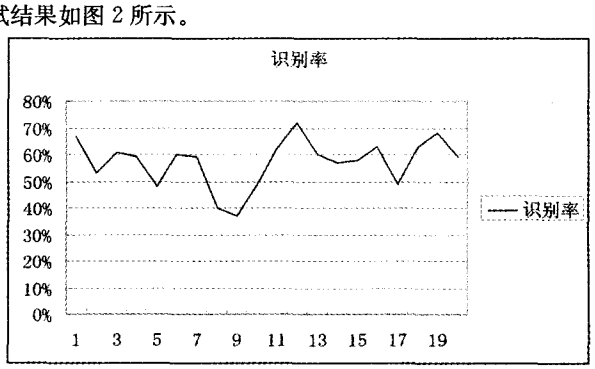

图 2

针对指令乱序变形技术的归一化研究

金 然 魏 强 王清贤
(信息工程大学信息工程学院 郑州 450002)

摘 要 新出现的恶意代码大部分是在原有恶意代码基础上修改转换而来。许多变形恶意代码更能自动完成该过程,由于其特征码不固定,给传统的基于特征码检测手段带来了极大挑战。采用归一化方法,并结合使用传统检测技术是一种应对思路。本文针对指令乱序这种常用变形技术提出了相应的归一化方案。该方案先通过控制依赖分析将待测代码划分为若干基本控制块,然后依据数据依赖图调整各基本控制块中的指令顺序,使得不同变种经处理后趋向于一致的规范形式。该方案对指令乱序的两种实现手段,即跳转法和非跳转法,同时有效。最后通过模拟测试对该方案的有效性进行了验证。

关键词 变形恶意代码,归一化,恶意代码检测

Research on Normalization towards Instructions Reordering Metamorphism Technique

JIN Ran WEI Qiang WAN Qing-Xian
(Information Engineering Institute, Information Engineering University, Zhengzhou 450002)

Abstract Much of apparently new malware comes from transformed known malware. Metamorphic malware could even complete this process automatically. The mutable signature makes the traditional detection method based on it difficult to detect metamorphic malware. Combining normalization idea with the traditional detection technology is a promising approach to resolve the problem. This paper proposes a normalization scheme towards instructions reordering metamorphism technique. In the scheme, the inspected code is firstly partitioned into some basic control blocks based on control-dependency analysis, then the instructions order in each block is adjusted according to the data-dependency graph. After the variants of malware are normalized according to the scheme, they tend to have the same form. The scheme is applicable to both jump method and non-jump method which are two implementations of instructions reordering. Testing has been conducted to validate the feasibility of the scheme.

Keywords Metamorphic malware, Normalization, Malware detection

1 引言

据统计,新出现的恶意代码中大部分是在原有恶意代码基础上修改转换而来[1]。变形[2](Metamorphism)技术是其中一种高级转换技术,它采用程序变换方法,在保持语义等价的前提下,改变代码形态。近年来出现的多种变形病毒(如Win32. Evol, Win32. Zmist, Win32. Metaphor 等)[2]更能自动完成这一过程。由于变换前后的代码不论是在静态文件中,还是动态执行在内存里,都不呈现相同的特征码,因此给基于特征码的检测方法带来了极大挑战[3]。

采用归一化方法,结合传统的基于特征码检测技术,是一种应对变形恶意代码的思路[4]。其主要思想是先将待测代码进行归一化处理,将其变换为规范形式,然后再使用匹配特征码的方法对规范代码进行检测。图1显示了该方法的整个过程:已知M为恶意代码,将其归一化后提取代码特征;对于待检测代码C1,…,Cn,同样先进行归一化,然后提取输出结果特征码与M的进行匹配,以此来判断待检测代码是否为M的变种。

恶意代码常用变形技术有:等价指令替换,插入垃圾代码和指令乱序,其中指令乱序有两种不同实现方式:跳转法和非跳转法。许多变形恶意代码都采用了其中一种或多种技术[2],因此研究针对这些变形技术的归一化策略有着重要意义。Christodorescu 和 Bruschi 分别在文[5,6]中针对这些变形技术讨论了相应的归一化方案,不过对于指令乱序,他们的方案仅对跳转法有效。本文专门针对指令乱序进行探讨,提出了一种适用于跳转法和非跳转法的归一化方案。通过模拟

金 然 博士研究生,主研方向:网络安全;魏 强 博士研究生,主研方向:网络安全;王清贤 教授,博导,主研方向:算法分析,网络安全。

结束语 测试表明,利用信息熵和决策分类技术对邮件进行分类识别,通过优选构建训练样本和测试样本能够通过学习得到能识别垃圾邮件的过滤规则,具有一定的智能性。该模型的识别准确率与训练样本的选择和构成、邮件特征选择以及评估阈值相关。阈值越小,准确率越高,但是将正常邮件误判为垃圾邮件的情况也较多;反之则相反。

参 考 文 献

1 赵晓明,郑少仁. 电子邮件过滤器的分析与设计. 东南大学学报,2001(9)
2 史忠植. 知识发现. 清华大学出版社,2002(1)
3 丁岳伟. 基于 SMTP 协议电子邮件的还原. 小型微型计算机系统,2002(3)
4 曹麒麟,张千里. 垃圾邮件与反垃圾邮件技术. 人民邮电出版社,2003
5 陈文伟、邓苏、张维明. 数据挖掘与知识发现综述. 计算机世界报,1997, 24期专题版
6 Postel J B. RFC821 simple mail transfer protocol. USA: IETF, 1982. 4~18
7 Freed N, Borenstein N. RFC2045 Multipurpose Internet Mail Extensions(MIME) part one: format of Internet message bodies. USA: IETF, 1996. 6~27