第 36 卷 第 6 期 计 算 机 工 程 2010 年 3 月
Vol.36 No.6 Computer Engineering March 2010
·安全技术· 文章编号：1000—3428(2010)06—0149—03 文献标识码：A 中图分类号：TP391

# 基于加权信息增益的恶意代码检测方法

张小康，帅建梅，史 林
(中国科学技术大学自动化系，合肥 230027)

**摘 要：**采用数据挖掘技术检测恶意代码，提出一种基于加权信息增益的特征选择方法。该方法综合考虑特征频率和信息增益的作用，能够更加准确地选取有效特征，从而提高检测性能。实现一个恶意代码检测系统，采用二进制代码的 N-gram 和变长 N-gram 作为特征提取方法，加权信息增益作为特征选择方法，使用多种分类器进行恶意代码检测。实验结果证明，该方法能有效提高恶意代码的检测率和准确率。
**关键词：**数据挖掘；变长 N-gram；特征选择；信息增益

# Malicious Code Detection Method Based on Weighted Information Gain

ZHANG Xiao-kang, SHUAI Jian-mei, SHI Lin
(Department of Automation, University of Science & Technology of China, Hefei 230027)

**【Abstract】** Using data mining technology to detect malicious code, this paper proposes a feature selection method based on weighted information gain. This method can select effective features more correctly by combining the advantage of information gain with classwise frequency. A malicious code detection system is implemented which adopts binary N-gram and variable-length N-gram as the feature extraction method, weighted information gain as the feature selection method. Several classifiers are used to detect malicious code in the system. Experimental results prove that this method can effectively improve the detection and accuracy rate.
**【Key words】** data mining; variable-length N-gram; feature selection; information gain

## 1 概述

传统恶意代码检测技术主要基于签名和启发式方法，基于签名的检测方法是提取已知病毒样本的特征，通过搜索病毒库查找相匹配的恶意代码特征。这种方法有较高的检测率，但无法检测新出现的病毒。启发式算法利用专家定义一组行为来检测未知恶意代码，准确率高，但效率低。数据挖掘方法可以通过学习恶意代码与正常代码的区别，有效检测未知恶意代码。Kephart 最早提出了一种特征提取和选择方法，使用人工神经网络的方法来检测引导区病毒。Arnold 采用了同样的方法来检测 Win32 病毒。文献[1]提出数据挖掘技术检测未知恶意代码，分别提取 Win32 dll 文件调用、ASCⅡ字符串、字节序列等代码特征，使用多种分类算法，包括 RIPPER、朴素贝叶斯和多重朴素贝叶斯算法，其中检测未知恶意代码准确率最高的是用字节序列作为特征的多重朴素贝叶斯算法。文献[2]提出基于二进制代码的 N-gram 特征提取方法，使用 K-NN 算法来检测恶意代码。文献[3]用 N-gram 提取二进制代码的特征，利用信息增益进行特征选择，使用朴素贝叶斯、SVM、boosted J48 等算法实现分类，其中，boosted J48 的检测率最高。随着变长 N-gram 应用于入侵检测和文本分类，文献[4]提出用 N-gram 提取二进制代码的特征，利用类域频率进行特征选择，采用 J48 分类算法进行分类。

已有文献主要专注于特征提取的方法，然而按已有方法提取出的特征包含大量冗余特征。如何选取有效特征，对于提高检测性能是非常重要的，但这个方面却鲜有研究。本文提取二进制代码的 N-gram 和变长 N-gram 作为特征，提出一种基于加权信息增益的特征选择方法，选择有效特征供分类器学习。加权信息增益特征选择方法利用特征是否出现以及出现的频率这 2 个因素来综合评价一个特征所含有的信息量，弥补了信息增益只考虑特征出现与否的不足。结果表明，采用此方法能够更加准确地选取有效特征，检测性能优于基于信息增益特征选择方法的性能。

## 2 恶意代码检测模型结构

本文实现了一个基于数据挖掘方法的恶意代码检测系统，采用代码二进制序列的 N-gram 和变长 N-gram 作为特征，加权信息增益作为特征选择方法，采用多种分类算法实现恶意代码检测的系统。其模型如图 1 所示。

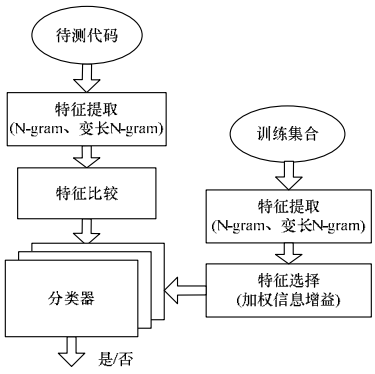


**图 1 恶意代码检测模型**

**基金项目：**国家“863”计划基金资助项目(2006AA01Z449)
**作者简介：**张小康(1983－)，女，硕士，主研方向：信息安全；帅建梅，副教授；史 林，硕士
**收稿日期：**2009-11-05 **E-mail：**zxkang@mail.ustc.edu.cn

训练部分首先选取一定数量的恶意代码和正常代码作为训练集合，提取代码二进制序列的 N-gram 和变长 N-gram 作为特征；然后进行特征选择，计算出每个特征对应的加权信息增益 $IW(j)$，并按照 $IW(j)$ 降序排序，选取前面若干个作为有效特征，根据每个训练样本是否包含这些特征，构成一个布尔向量空间，供分类器学习。检测部分提取待测代码二进制序列的 N-gram 和变长 N-gram 作为特征，根据每个待测代码是否包含训练部分选择的有效特征，构成一个布尔向量空间，分类器用几种分类算法对该向量空间进行分析，判断该待测代码是否为恶意代码。

假设待测代码样本数量为 $N$，则 $N = TP + TN + FP + FN$。其中，$TP$ 是恶意代码被正确分类的数目；$FP$ 是正常代码被错误标记为恶意代码的数量；$TN$ 是正常代码被正确分类的数目；$FN$ 是恶意代码被错误标记为正常代码的数量。一个恶意代码检测工具有如下 2 个评价指标：

(1)准确率 $\frac{TP+TN}{N}$，即所有被正确分类的代码数量在待测集合中所占的比例。

(2)检测率 $\frac{TP}{TP+FN}$，即被正确分类的恶意代码数目在待测集合的所有恶意代码中所占的比例。

## 3 特征提取方法

本文选择代码二进制序列作为表示形式，在此基础上使用 N-gram 和变长 N-gram 滑动窗口提取特征。下面分别介绍这 2 种算法。

### 3.1 N-gram 特征

N-gram 是由一个长度为 $N$ 的滑动窗口收集的一系列重叠的子字符串，这个窗口每次滑动一个单位长度。比如，08 00 74 ff 13 b2，其对应的 3-gram 为(08 00 74), (00 74 ff), (74 ff 13), (ff 13 b2)。N-gram 可以捕获到一些潜在的其他方法很难准确提取的特征，在恶意代码检测领域，N-gram 是广泛应用的特征提取方法。

N-gram 提取特征有 2 个缺点：

(1)N-gram 很难同时捕获不同长度的字节序列，当一个有意义的字节序列不是 $N$ 的倍数时，会产生边缘误匹配，从而无法提取这个特征。

(2)由 N-gram 产生的特征集合非常庞大，需要相当大的存储容量。本文在实现 N-gram 算法时采用 Trie 数据结构，节省存储空间，保证了特征生成和查找的快速准确性。

### 3.2 变长 N-gram 特征

变长 N-gram 也称段落，是一串有意义的连续字节序列，与 N-gram 不同，它的长度是不固定的，避免一个有意义的序列被拆开的可能性。提取有意义的段落，首先需要在一系列字节序列中寻找断点，相邻断点之间的连续序列就是一个段落。本文的段落分割算法采用了专家投票算法[5]。

专家投票算法假设有 2 个专家。一个是频率专家，负责测量每个子序列出现的频率，频率越高，那么这个子序列是一个段落的可能性越大，其中包含断点的可能性越小；另一个是熵专家，负责测量每个点的熵，如果一个元素后面连接的元素每次都各不相同，熵越大，那么与一个后面连接的元素总是固定的元素相比，是一个段落结尾的可能性更大。每个位置都有一个分数，在一个固定长度的滑动窗口中，频率最大以及熵最大的 2 个位置的分数加 1。最后，综合 2 个专家的结果，根据每个位置所得到的累加分数来判断可能的断点。将 2 个断点之间的连续序列作为一个特征提取出来。本文选择 $d$=4 的 Trie 数据结构实现专家投票算法。

假设字符串 string=(01 E8 B8 01 E8 B8 B8 01)，其 Trie 结构如图 2 所示。其中，左边叶节点 E8 在 Trie 中表示字节序列(01 E8 B8)，这个节点的数字 2 表示这个序列出现的次数。当长度为 3 的窗口从左至右划过 string 时，先遇到(01|E8|B8|)。第 1 个位置的频率是由此隔开的前后 2 个序列的频率之和，即 $f=f(01)+f(\text{E8 B8})$。$f(01)=f(01)+f(\text{E8 B8})$ 表示在 Trie 第 1 层的(01)的频率，$f(\text{E8 B8})$ 表示在 Trie 第 2 层并且父节点是(E8)的(B8)的频率。第 1 个位置的熵表示在 Tire 第 1 层的(01)的熵，第 2 个位置的熵表示在 Trie 第 2 层并且父节点是(01)的(E8)的熵，以此类推。这个窗口中频率和熵最大的位置分别被标记出来，对应的分数加 1。然后窗口向右滑动，直至结束。

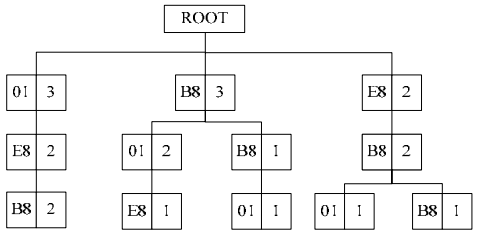


**图 2 string 的 Trie 结构($d$=4)**

Trie 中节点的频率为

$$P(x_0) = \frac{f(x_0)}{f(parent(x_0))} \tag{1}$$

节点的熵为

$$E(parent(x_0)) = -\sum_{i=0}^{m} P(x_i)\text{lb}P(x_i) \tag{2}$$

其中，$m$ 为 $x_0$ 的父节点拥有的子树个数。

当滑动窗口遍历整个 string 之后，可以根据分数找出一些局部极大值，对应的位置就为断点。string 的段落分割结果为((01 E8 B8) (01 E8 B8) (B8 01))。

## 4 加权信息增益

通过以上方法提取特征包含很多冗余特征，从中选取有利于区分代码类型的特征是必要的。本文提出加权信息增益的特征选择方法选择有效的一组特征。

信息增益也被称作平均互信息量。定义如下：

$$I(X;Y) = H(X) - H(X \mid Y) \tag{3}$$

其中，$H(X)$ 是 $X$ 的信息熵；$H(X/Y)$ 是已知 $Y$ 情况下 $X$ 的条件熵。上式表明，从 $Y$ 中获取关于 $X$ 的平均互信息量 $I(X;Y)$，等于获知 $Y$ 前后，关于 $X$ 的平均不确定性的消除。在恶意代码检测中，信息增益 $IG(j)$ 表示第 $j$ 个特征所传递的平均信息量，由式(3)可得：

$$IG(j) = \sum_{v_j \in \{0,1\}} \sum_{C_i} P(v_j, C_i)\text{lb}\frac{P(v_j, C_i)}{P(v_j)P(C_i)} \tag{4}$$

其中，$v_j$ 是第 $j$ 个特征属性的值，$v_j$=1 代表这个特征出现过，$v_j$=0 表示这个特征没有出现过；$C_i$ 表示第 $i$ 个类别，这里一共有 2 类：恶意代码和正常代码；$P(v_j, C_i)$ 表示在类 $C_i$ 中，第 $j$ 个特征值为 $v_j$ 的比例；$P(v_j)$ 表示在训练集中第 $j$ 个特征值为 $v_j$ 的比例；$P(C_i)$ 表示训练集中类 $C_i$ 所占的比例。信息增益越大，代表这个特征对分类越有用。

信息增益特征提取方法在计算信息增益时，把特征存在与否设置为布尔值，只考虑了其在代码中存在与否，忽略了特征出现的频率的作用。实际情况中，有些特征在 2 类代码

中出现的频率相差很多，这类特征对于正确检测恶意代码是非常有效的。本文综合考虑了特征频率和信息增益的作用，提出了一种基于加权信息增益的特征选择方法，能够更加准确地选出有效特征，提高恶意代码的准确率和检测率。

第 $j$ 个特征的加权信息增益定义如下：

$$IW(j)=\lambda_j\sum_{v_j\in\{0,1\}}\sum_{C_i}P(v_j,C_i)\mathrm{lb}\frac{P(v_j,C_i)}{P(v_j)P(C_i)} \tag{5}$$

其中，$\lambda_j$ 表示第 $j$ 个特征对应的权重。通过一个特征在 2 类代码中出现的平均次数比衡量此特征对于检测恶意代码的有效性，定义 $\lambda_j$ 如下：

$$\lambda_j=\begin{cases}\mathrm{lb}(1+g(\dfrac{f_{jM}N_B}{f_{jB}N_M})) & f_{jM}\neq 0, f_{jB}\neq 0\\ \mathrm{lb}(1+\dfrac{f_{jM}}{\alpha N_M}) & f_{jM}\neq 0, f_{jB}=0\\ \mathrm{lb}(1+\dfrac{f_{jB}}{\alpha N_B}) & f_{jB}\neq 0, f_{jM}=0\end{cases} \tag{6}$$

其中，$g(x)=\begin{cases}x & x\geq 1\\ \dfrac{1}{x} & 0<x<1\end{cases}$。

$f_{jM}$，$f_{jB}$ 分别表示第 $j$ 个特征在恶意代码与正常代码中出现的总次数；$N_M$，$N_B$ 分别表示训练集合中恶意代码和正常代码的个数；$\alpha$ 是调节参数，本文中 $\alpha$ 取 1。当 $f_{jM}$，$f_{jB}$ 都不为 0 时，$\lambda_j$ 与第 $j$ 个特征在 2 类代码中出现的平均次数比成正比；当 $f_{jM}$ 不为 0、$f_{jB}$ 为 0 时，$\lambda_j$ 与 $f_{jM}$ 成正比；当 $f_{jB}$ 不为 0、$f_{jM}$ 为 0 时，$\lambda_j$ 与 $f_{jM}$ 成正比。

加权信息增益 $IW(j)$ 越大，表示这个特征对正确分类恶意代码越有效。

## 5 实验仿真结果及分析

本文将 429 个正常代码与 408 个恶意代码作为训练样本集，所有恶意代码来源于网站 http: //vx.netlux.org，正常代码都是在一台新安装 Windows XP 的计算机上获得的。选择有效特征个数为 500[3]。分类算法由 WEKA 实现，实验采用十重交叉验证。

本文实现了采用 3-gram, 4-gram 以及变长 N-gram 算法时，分别使用信息增益和加权信息增益作为特征选择方法，选择朴素贝叶斯、SVM、J48 为分类器的恶意代码检测系统，对每种组合选择分类效果最好的检测结果，填入表 1。其中采用变长 N-gram 和加权信息增益的组合具有最高的检测率。由表 1 可以看出，采用加权信息增益的检测结果均优于信息增益的检测结果，4-gram 的性能优于 3-gram，变长 N-gram 的性能优于 N-gram。

**表 1　几种特征提取和选择方法组合的检测结果** (%)

| 算法 | 检测率 | 准确率 |
|---|---|---|
| 3-gram(IG) | 95.80 | 96.65 |
| 3-gram(加权 IG) | 97.90 | 98.21 |
| 4-gram(IG) | 96.74 | 97.37 |
| 4-gram(加权 IG) | 98.37 | 98.80 |
| 变长 N-gram(IG) | 97.90 | 98.68 |
| 变长 N-gram(加权 IG) | 98.83 | 99.16 |

当采用变长 N-gram 作为特征提取方法，加权信息增益作为特征选择方法时，各种分类算法的检测结果见表 2。其中 J48 算法的检测率最高。

**表 2　变长 N-gram 和加权 IG 组合使用各种分类器的检测结果** (%)

| 算法 | 检测率 | 准确率 |
|---|---|---|
| SVM | 97.90 | 94.74 |
| J48 | 98.83 | 99.16 |
| 朴素贝叶斯 | 95.79 | 96.42 |

## 6 结束语

本文基于数据挖掘技术，采用 N-gram 和变长 N-gram 作为特征提取方法，综合考虑特征频率和信息增益的作用，提出一种基于加权信息增益的特征选择方法，实现了一个恶意代码检测系统。实验结果表明，这种方法的检测性能优于基于信息增益的特征选择方法，能够更准确地选取有效特征。下一步将研究更为有效的特征选择方法，提高检测系统的性能；采用更大的训练和测试集合，进一步验证系统的性能。

## 参考文献


[1] Schultz M G, Eskin E, Zadok E, et al. Data Mining Methods for Detection of New Malicious Executables[C]//Proc. of the IEEE Symposium on Security and Privacy. Oakland, California, USA: IEEE Press, 2001: 38-49.

[2] Assaleh T A, Cercone N, Keselj V, et al. Detection of New Malicious Code Using N-grams Signatures[C]//Proc. of the 2nd Annual Conference on Privacy, Security and Trust. Ontario, Canada: [s. n.], 2004: 193-196.

[3] Kolter J Z, Maloof M A. Learning to Detect and Classify Malicious Executables in the Wild[J]. Journal of Machine Learning Research, 2006, 7: 2721-2744.

[4] Reddy D S, Dash S K, Pujari A K. New Malicious Code Detection Using Variable Length N-grams[C]//Proc. of the 2nd International Conference on Information Systems Security. Kolkata, India: [s. n.], 2006: 276-288.

[5] Cohen P, Heeringa B, Adams N M. An Unsupervised Algorithm for Segmenting Categorical Time Series into Episodes[C]//Proc. of the ESF Exploratory Workshop on Pattern Detection and Discovery. London, UK: [s. n.], 2002: 49-62.


编辑　顾逸斐

(上接第 148 页)

## 参考文献


[1] Garber L. Denial-of-Service Attacks Rip the Internet[J]. Computer, 2000, 33(4): 12-17.

[2] Paxson V. An Analysis of Using Reflectors for Distributed Denial-of-Service Attacks[J]. ACM SIGCOMM Computer Communication Review, 2001, 31(3): 38-47.

[3] Mirkovic J, Reiher P. A Taxonomy of DDoS Attack and DDoS Defense Mechanisms[J]. ACM SIGCOMM Computer Communication Review, 2004, 34(2): 39-53.

[4] Peng Tao, Leckie C, Ramamohanarao K. Survey of Network-based Defense Mechanisms Countering the DoS and DDoS Problems[J]. ACM Computing Surveys, 2007, 39(1): 1-42.

[5] Keromytis A D, Misra V, Rubenstein D. SOS: Secure Overlay Services[C]//Proc. of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. Pittsburgh, Pennsylvania, USA: [s. n.], 2002: 61-72.

[6] Tupakula U K, Varadharajan V. A Practical Method to Counteract Denial of Service Attacks[C]//Proc. of the 26th Australasian Computer Science Conference. Adelaide, Australia: [s. n.], 2003: 275-284.


编辑　顾逸斐