DOI: 10.3969/j.issn.1673-4785.201108007
网络出版地址:http://www.cnki.net/kcms/doi/10.3969/j.issn.1673-4785.201108007.html

基于文本聚类的网络攻击检测方法

杨晓峰[1],李伟[1,2],孙明明[1],胡雪蕾[1]
(1.南京理工大学 计算机科学与技术学院,江苏 南京 210094; 2.哈佛医学院 Dana-Farber 癌症研究所,波士顿 马萨诸塞州 02115,美国)

摘 要:针对 Web 服务应用的攻击是近年来网络上广泛传播的攻击方式,现有的攻击检测算法多采用监督学习的方法确定正常行为和攻击行为的分类边界;但由于监督检测模型在检测之前需要复杂的学习过程,往往会降低系统的实用效果。因此,根据现实中正常访问样本和攻击样本在数量和分布上的差异,提出了一种基于文本聚类的非监督检测算法。算法首先采用迭代聚类过程聚类样本,直至聚为一类;同时根据异常与正常样本的分布规律,在聚类过程中选择最优的最大类别作为正常样本类,将其余的作为异常样本类。最优方案的选择采用了使得分类误差最小的原则确定。实验表明,与多种经典检测方法相比,该方法省去了复杂的学习过程,增强了方法的适应性,具有较好的检测率和误报率。

关键词:网络攻击;网络攻击检测;文本聚类;非监督检测算法
中图分类号:TP393 **文献标志码**:A **文章编号**:1673-4785(2014)01-0040-07

中文引用格式:杨晓峰,李伟,孙明明,等.基于文本聚类的网络攻击检测方法[J]. 智能系统学报, 2014, 9(1): 40-46.
英文引用格式:YANG Xiaofeng, LI Wei, SUN Mingming, et al. Web attack detection method on the basis of text clustering [J]. CAAI Transactions on Intelligent Systems, 2014, 9(1): 40-46.

Web attack detection method on the basis of text clustering

YANG Xiaofeng[1], LI Wei[1,2], SUN Mingming[1], HU Xuelei[1]
(1. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China; 2. Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA)

Abstract: The attacks aiming at Web service applications within the past several years have become more widely-propagated, and the present attack detection algorithms mostly use the supervision study to determine the border between normal the behavior and attack behavior; however, for the supervision and detection model, before the detection, a complex studying process is necessary, this will lower the practical effects of the system. Therefore, on the basis of the realistic difference between the normal visit specimen and the attack specimen on the aspects of quantity and distribution, an unsupervised detection algorithm based on text clustering is proposed. In the algorithm, firstly, the iteratively clustered process is applied to cluster specimens, until reaching a category; in addition, according to the distribution law of the abnormal and normal specimens, in the clustering process, the optimal maximum category is considered as the normal specimen category and the others are considered as an abnormal specimen category. The optimal scheme is determined on the basis of the principle of the minimum classification error. The experiment shows that, in comparison with many traditional detection methods, the method used in this paper omits complex study processes and improves adaptability; the detection rate and the false positive rate are excellent.

Keywords: Web attack; Web attack detection; text clustering; unsupervised detection algorithm

随着 Web 应用的不断普及,网络服务为越来越多的用户使用。由于许多网络应用服务开发者安全意识的缺失,致使网络服务程序中存在大量的安全漏洞,这使得 Web 服务器成为黑客攻击的主要目标之一。最新的 CVE 漏洞趋势报告[1]显示,跨站脚本攻击(XSS)、SQL 注入(SQL-inject)和远程文件包含(RFI)等基于 HTTP 协议[2]的网络攻击已经成为近年来互联网上最主要的攻击方式。

为了防御 Web 攻击,网络安全工作者研究和开

收稿日期:2011-08-29. 网络出版日期:2014-02-24.
基金项目:国家自然科学基金资助项目(60705020);江苏省自然科学基金资助项目(BK207594).
通信作者:李伟. E-mail: liweinust@hotmail.com.

发了多种方法[3-5]。入侵检测系统(intrusion detection system, IDS)是防御网络攻击的重要手段[6]。入侵检测方法主要分为两大类:误用检测(misuse detection)和异常检测(anomaly detection)。误用检测通过规则的方式来定义攻击行为的特征,例如 Snort 检测系统[7]。随着攻击方式的日益增加,误用检测方法的弊端越来越明显:攻击的增加导致检测规则的增加,这使得及时准确的更新、维护规则库越来越困难;而且,误用检测只能检测已知的攻击方式,对新的未知攻击方式则无能为力[8]。异常检测可以有效地克服误用检测的上述缺点。异常检测定义正常行为的特征,通过统计分析、数据挖掘的方法,学习得到正常行为的特征模型,当网络行为显著偏离正常的行为模式时则识别为异常行为。近年来提出了很多针对 Web 攻击的异常检测模型[9-10],这些模型多采用监督学习来学习正常访问行为的特征模式,利用学习样本的分布来确定正常和攻击行为的边界。然而,许多异常检测方法问题在于:1)模型在开始检测之前需要多次学习,通常需要大量的计算资源;2)由学习得到的正常行为模型需要进一步做精细的泛化处理,使得模型能够尽可能代表未学习过的正常样本,而泛化的困难性很多时候会大大地限制该模型的应用效果。

基于 Web 访问的统计研究表明,正常样本占总体样本的大多数,且行为模式相似;而攻击样本只占一小部分,且行为模式各异[11]。本文依据正常样本与攻击样本的统计差异,提出了一种基于文本聚类的非监督检测算法:采用层次聚类的过程逐步聚合样本,并用贝叶斯原则指导聚类过程,最终将正常和攻击样本聚合到不同的类别之中。本文方法在系统设计方面具有非监督学习方法的特点,省去了复杂的模型学习过程,简化了检测流程,以提高算法的适应性。在与多种经典检测方法的比较实验中,本文方法取得了较高的检测率,可以较好地抑制误报率。

1　相关研究工作

异常检测方法最早被应用于传统入侵检测系统的设计。Mahoney 和 Chan[12]从正常的网络数据包序列建立马尔可夫模型来检测新的未知攻击方式。Portnoy 等[11]提出了基于 TCP 协议的聚类方法。Warrender 等[13]分析了特殊应用程序的正常系统调用序列以识别恶意程序。Sengar 等[14]专门为 VoIP(voice over Internet protocol)协议的通信设计了 IDS 系统,系统为协议生成相应的状态机以检测通信行为是否为攻击。周东清等[15]提出基于隐马尔可夫模型的 DDoS(分布式拒绝服务)检测攻击方法。

针对 Web 攻击,Kruegel 和 Vigna[9-10,16]提出了多种异常检测模型,它们是长度模型、字符分布、NFA 和枚举类型。这些模型分别解析了 HTTP 协议请求,针对 HTTP 请求或请求中的属性进行检测,给出异常程度评价。它们都使用了有监督的方式来训练正常请求的模型参数和最佳的分类阈值。与此类似,本文研究的是基于 HTTP 协议的 Web 攻击检测。

本文利用现实情况中正常样本占总访问量的绝大多数、行为特征类似,而攻击样本数量少、个体行为差别大的规律,提出了基于聚类的非监督检测方法,省去了训练样本的标记和训练过程,提高了系统的适应能力。需要特别指出的是,类似的分布规律也同样出现在 IDS 告警处理的研究中[17],这也是方法有效性成立的基础和方法设计的准则。

2　基于文本聚类的检测方法

2.1　算法假设

HTTP 请求一般由多个参数组成,参数中间用字符"&"隔开,每个参数以"属性名=属性值"的形式组织。参数值中放入恶意的代码是常见的针对特定程序和特定属性的攻击方式,未经充分检查的参数值可以引起 Web 服务端的信息泄露、服务崩溃,甚至导致 Web 服务器劫持。在包含多个参数的请求中,任意一个参数的属性值被检测到含有攻击代码则判定请求为攻击。虽然属性值可以取包括数字、字符和特殊符号等多种形式,但本文都看作是一个文本字符串。

在对现实 Web 访问数据的分析中,发现 90%以上的访问请求都是正常的,恶意的攻击行为占总请求量的很小一部分[11]。正常访问和恶意攻击的概率密度函数符合如下的特点(如图 1 所示):1)正常访问占绝大多数,攻击占很小一部分;2)正常访问的参数形式之间变化很小,具有很好的聚类特性;3)恶意的攻击与正常的样本模式之间有较大的差异,聚类特性较差。

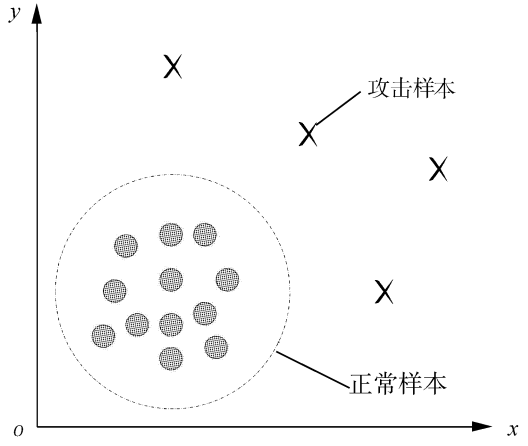


图 1　样本分布示例

Fig.1　Illustration of sample distribution

Web 服务的同一个页面或程序通常会被许多不同的用户以同样的形式访问;而且一个参数可以接受的数据类型一般都比较规范,例如属性名为“生日”的参数形式会是一个日期型字符串,不同的合法输入值之间比较相似。这就使得正常样本具备了上述特点 1)和 2)的分布特性。恶意攻击则来自于少数的恶意客户端,同样的攻击方式没有必要多次重复,同时攻击的形式也没有规律性,决定了上述攻击样本的分布特性。

2.2 距离定义

本文处理的数据是字符串样本,而传统的欧式距离只适用于向量型的数据类型,本节将定义字符串数据间的距离度量。

2.2.1 样本间距离

样本间距离是定义所有距离度量的基础。给定字符串样本 $x_1 = k_1k_2\cdots k_m$, $x_2 = l_1l_2\cdots l_n$,其中 $k_i = x_1(i)$ 是 x_1 中第 i 个字符,$l_j = x_2(j)$ 是 x_2 中第 j 个字符,$m = L(x_1)$、$n = L(x_2)$ 分别代表 x_1 和 x_2 的字符串长度。样本 x_1 和 x_2 的距离 $d(x_1,x_2)$ 由下面 3 个部分组成。

1)字符串长度差异。

字符串长度差异是字符串间最基本的差异度量,定义为

$$d_1(x_1,x_2) = |L(x_1) - L(x_2)|$$

2)字符集差异。

样本 x 的字符集是组成样本字符串的所有字符的集合,用 $A(x)$ 表示样本 x 的字符集。例如,$x =$ “12332”时,$A(x) = \{1,2,3\}$。字符集差异用于描述字符串在字符选择上的差异,当 2 个字符串的字符集在数量和类型方面存在较大不同,特别是出现不同类型的字符时,需给出惩罚。

定义字符集距离之前需要定义字符间的距离,字符 k、l 的距离 $\hat{d}(k,l)$ 定义如下:

$$\hat{d}(k,l) = \begin{cases} 0, & k = l \\ 1, & k \neq l, k \text{ 和 } l \text{ 属于同一字符类型} \\ 10, & k \neq l, k \text{ 和 } l \text{ 属于不同字符类型} \end{cases}$$

表 1 为算法中使用的 3 种字符类型,分别为数字类型、小写字符和大写字符、其他的字符。例如字符“/”和“-”不属于同类字符,字符间距离为 10。

表 1 字符类型定义

Table 1 Definition of character types

字符类型	字符范围
数字	0~9
小写字符和大写字符	a~z、A~Z
其他	/.-=’……

字符集 A 和字符 b 间的距离用 $\bar{d}(A,b)$ 表示,其中 $\bar{d}(A,b) = \bar{d}(b,A)$,定义为

$$\bar{d}(A,b) = \min_{i=1}^{n} \hat{d}(A(i),b)$$

式中:n 为字符集 A 中字符的数量,$A(i)$ 表示 A 中第 i 个字符。

字符集距离定义为字符和字符集间的距离之和,即字符串样本 x_1 和 x_2 间的字符集距离为

$$d_2(x_1,x_2) = \sum_{i=1}^{n} \bar{d}(A(x_1),x_2(i)) + \sum_{i=1}^{m} \bar{d}(A(x_2),x_1(i)) = \sum_{i=1}^{n} \min_{j}^{m} \hat{d}(A(x_1)(j),x_2(i)) + \sum_{i=1}^{m} \min_{j}^{n} \hat{d}(A(x_2)(j),x_1(i))$$

3)字符组合顺序差异。

字符组合顺序是用 2 个或多个连续字符为单位,描述字符串间相似度的度量,例如字符串“abcd-abcd”和“dcbadcba”在长度和字符集方面都相同,但字符的组合顺序大不相同。$G(x)$ 表示字符串样本 x 中连续出现 2 个字符的集合,$\#(G(x))$ 表示集合 $G(x)$ 中元素的数目,则 x_1 和 x_2 的字符组合顺序差异定义为

$$d_3(x_1,x_2) = \#(G(x_1) \cup G(x_2)) - \#(G(x_1) \cap G(x_2))$$

$d_3(x_1,x_2)$ 是 x_1 和 x_2 中不同的 2-grams 的数量。

因此,x_1 和 x_2 的距离为

$$d(x_1,x_2) = d_1(x_1,x_2) + d_2(x_1,x_2) + d_3(x_1,x_2) = |L(x_1) - L(x_2)| + \sum_{i=1}^{n} \min_{j}^{m} \hat{d}(A(x_1)(j),x_2(i)) + \sum_{i=1}^{m} \min_{j}^{n} \hat{d}(A(x_2)(j),x_1(i)) + \#(G(x_1) \cup G(x_2)) - \#(G(x_1) \cap G(x_2))$$

2.2.2 样本与类间距离

样本 x 与聚类 $C = \{x_1,x_2,\cdots,x_n\}$ 间的距离定义为

$$d(x,C) = \frac{1}{n}\sum_{i=1}^{n} d(x,x_i) \tag{1}$$

2.2.3 类间距离

类 $C_1 = \{x_{11},x_{12},\cdots,x_{1m}\}$ 与类 $C_2 = \{x_{21},x_{22},\cdots,$

$x_{2n}\}$间的距离为

$$d(C_1,C_2)=\frac{1}{n}\sum_{i=1}^{n}d(x_{2i},C_1)=$$

$$\frac{1}{m\times n}\sum_{i=1}^{n}\sum_{j=1}^{m}d(x_{2i},x_{1j})$$

2.2.4　类内距离

聚类 $C=\{x_1,x_2,\cdots,x_n\}$ 的类内距离为

$$D(C)=d(C,C)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d(x_i,x_j)\qquad(2)$$

2.3　基于聚类的检测算法

如图 2 所示,本文基于层次聚类的 Web 攻击检测算法包括 4 个部分组成:层次聚类算法、迭代聚类算法、最优方案选择算法和检测算法。

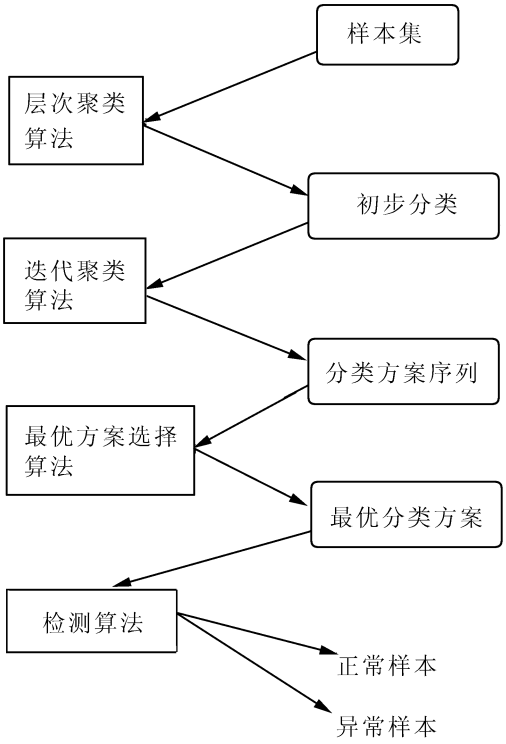


图 2　基于层次聚类的检测算法

Fig.2　Agglomerative clustering based detection method

2.3.1　层次聚类算法

层次聚类算法旨在形成初步的 2 类聚类,分别代表正常的聚类和异常的聚类,并且使得聚类符合 2.1 节中特点 1)的条件。

样本集 $X=\{x_1,x_2,\cdots,x_n\}$ 的层次聚类的算法如下:

1)初始化,将每一个样本作为一个聚类,得到聚类集合 $C=\{C_1,C_2,\cdots,C_n\}$,其中 $C_i=\{x_i\}$;

2)选择 C_i、C_j,满足 $d(C_i,C_j)=\min_{k\neq l}d(C_k,C_l)$;

3)合并 C_i、C_j,更新聚类集合 $C=C-C_i-C_j+C_i\cup C_j$;

4)选择 C_M,满足 $\#(C_M)=\max_i\#(C_i)$;

5)如果 $\#(C_M)\leqslant 50\%\times\sum_i\#(C_i)$,则转到步骤 2);

6)将除 C_M 以外的所有聚类合成一类 $C_N=\cup_i C_i, C_i\in C-C_M$;

7)记聚类 C_M 和 C_N 为 C_1、C_2;

8)结束。

层次聚类算法得到了聚类 C_1 和 C_2,其中 C_1 包含整个样本集 50%以上的样本,根据对正常样本和异常样本分布的假设,将 C_1 和 C_2 作为初步的正常样本和异常样本聚类,继续迭代聚类。

2.3.2　迭代聚类算法

迭代聚类是以层次聚类中的初步分类方案为基础继续聚类过程,每一个聚类步骤将产生一个分类方案,这些分类方案序列是 2.3.3 节中算法的基础。

C_1 和 C_2 是迭代聚类的数据源。根据样本数据的分布特性,C_1 仅仅包含了正常样本中聚类特性最好的一部分样本,C_2 还含有大量的正常样本。通过迭代聚类,将 C_2 中的正常样本逐步聚合到 C_1 中来。

C_1 和 C_2 是对样本集的一种分类方案,将集合 $S=\{C_1,C_2\}$ 称为一种分类方案,迭代聚类的结果是一个分类方案的序列。

迭代聚类算法如下:

1)初始化,$N=0$;

2)记录分类方案 $S_N=\{C_1,C_2\}$,$N=N+1$;

3)从 C_2 中选择样本 x,满足 $d(C_1,x)=\min_i d(C_1,x_i), x_i\in C_2$;

4)将 x 从 C_2 移到 C_1,更新 C_1、C_2,$C_1=C_1\cup\{x\}$,$C_2=C_2-\{x\}$;

5)如果 $C_2\neq\varnothing$,则转到步骤 2);

6)结束。

2.3.3　选择最优方案算法

本小节算法旨在选择满足 2.1 节中特点 2)和 3)的分类方案。选择最优聚类方案时,考虑使得分类误差最小,最小分类误差计算如下:

$$e=P(C_1\mid C_2)+P(C_2\mid C_1)$$

根据切比雪夫不等式,得到样本 x 分布在均值 μ 距离 t 以外的概率上界:

$$P(|x-\mu|>t) < \frac{\sigma^2}{t^2}$$

所以图 3 中,

$$P(C_1 \mid C_2) = \frac{1}{2}P(|x'-\mu_2|>|x-\mu_2|) < \frac{\sigma_2^2}{2\times|x-\mu_2|^2}$$

$$P(C_2 \mid C_1) = \frac{1}{2}P(|x'-\mu_1|>|x-\mu_1|) < \frac{\sigma_1^2}{2\times|x-\mu_1^2|}$$

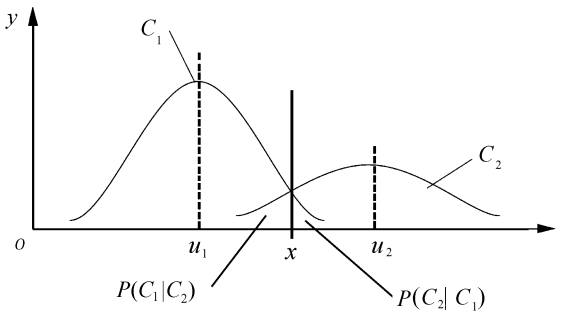


图 3 两类问题中的最小分类误差原则

Fig.3 Minimum error principle in 2-class situation

最小误差的上界 $\bar{e}$ 为

$$e < \bar{e} = \frac{1}{2}\left(\frac{\sigma_1^2}{|x-\mu_1|^2}+\frac{\sigma_2^2}{|x-\mu_2|^2}\right) \quad (3)$$

用 $\bar{e}$ 近似最小分类误差 e 来衡量分类方案的优劣。

给定 $S_i=\{C_1,C_2\}$,x 为 S_{i-1}到 S_i迭代聚类时从 C_2移到 C_1中的样本,式(3)中 C_1和 C_2的方差 σ_1^2 和 σ_2^2 分别用式(2)的类内距离$D(C_1)$和 $D(C_2)$来计算,而 x 到类均值的距离 $|x-\mu_1|$ 和 $|x-\mu_2|$ 分别用式(1)中的样本与类间距离$d(x, C_1)$和 $d(x,C_2)$计算,$\bar{e}$ 的计算公式为

$$\bar{e} = \frac{1}{2}\left(\frac{D(C_1)}{d(x,C_1)^2}+\frac{D(C_2)}{d(x,C_2)^2}\right)$$

在分类方案序列 $S_1,S_2,\cdots,S_N$中,选择使得 $\bar{e}$ 最小的分类方案 S^*,作为最优的分类方案。

2.3.4 检测算法

最优的分类方案 $S^*=\{C_1,C_2\}$,则检测算法直接根据样本所属聚类的类别进行分类:

$$f(x)=\begin{cases}0, & x\in C_1\\ 1, & x\in C_2\end{cases}$$

式中:0 表示正常样本,1 表示异常样本。

3 实验和讨论

标准的网络攻击样本库数量较少,其中比较著名的包括 MIT 林肯实验室的 DARPA 数据集[18-19]和 KDDCup'99 数据集[20]。因为这些数据集只含有少量的 Web 攻击,同时数据收集时间较早,无法有效地验证本文方法,所以采用笔者收集的域名为"njust.edu.cn"2010 年上半年的网站 log 数据。经过人工和程序辅助的检查,用于本文实验的数据集含有大量的恶意攻击访问,其中包括 SQL 注入、跨站脚本攻击(XSS)、文件包含、漏洞溢出等。

表 2 列出了 NJUST 数据集中 3 个子数据集的大致情况,数据集 *A* 样本数最小,正常的样本格式规整、形式单一;数据集 *B* 中含有大量的攻击样本,数量几乎与正常的样本数量相当;数据集 *C* 的样本总量最大,但攻击样本的比率最低。

表 2 实验数据集信息

Table 2 Information of NJUST datasets

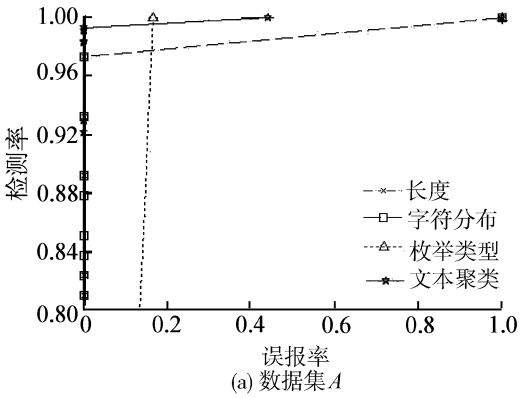
数据集	样本数	正常样本	攻击样本
A	6 119	4 512	1 607
B	11 148	6 050	5 098
C	85 700	83 391	2 309

用文献[10-12]中的长度模型、字符分布模型和枚举类型模型与本文的文本聚类方法作比较。长度、字符分布和枚举类型的分类原则均采用原文中阈值高于最高正常样本异常度 10%的经验策略。

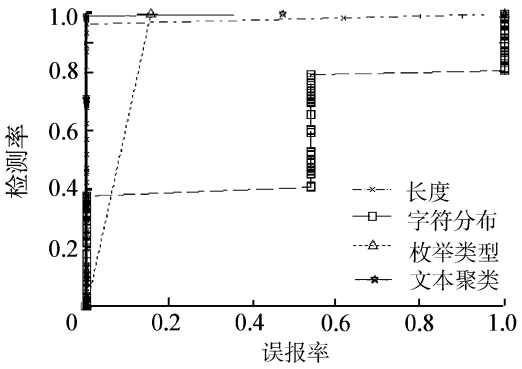
实验目的是比较本文方法和经典方法的 ROC 曲线,即模型的分类能力。与其他模型直接输出测试样本的异常测度不同,文本聚类 Web 攻击检测算法只能给出样本的分类结果。在本文方法的层次聚类和迭代聚类过程中,取每一个最大聚类中的样本为正常样本,其他样本为异常样本。这样每一个聚类的中间步骤都作为一次分类,把这些分类结果的检测率和误报率绘制成文本聚类方法的 ROC 曲线。聚类方法的 ROC 曲线也反应了聚类的分类性能,可以用来验证 2.3.3 节中最优分类方案选择算法是否可以找到聚类方法的最优分类点。

NJUST 数据集的 3 个子数据集分别被随机均分成训练集和测试集,用训练集训练模型,测试集测试模型的检测性能。因为文本聚类方法不需要事先训练,因此直接在测试集上测试聚类方法。

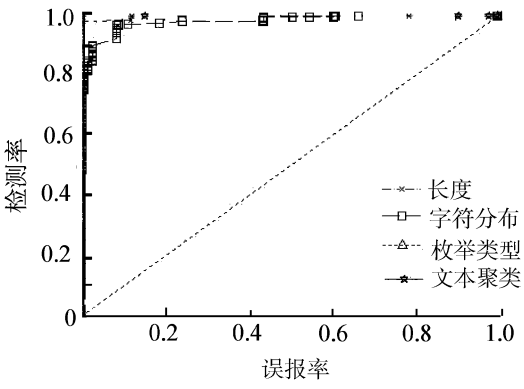
4 种检测方法在 *A*、*B*、*C* 3 个数据集上的 ROC 曲线对比结果依次见图 4,本文方法在图中称为"文本聚类"。



(a) 数据集 *A*



(b) 数据集 *B*



(c) 数据集 *C*

图 4　数据集 *A*、*B*、*C* 上的 ROC 曲线对比

Fig.4　ROC comparisons on dataset *A*, *B* and *C*

从结果中可以看出:

文本聚类方法和长度模型都具有很高的检测率和较低的误报率,同时这 2 种方法在不同数据集上的性能表现都很稳定,最好的检测率和误报率几乎没有变化。

字符分布模型的性能根据数据集的不同起伏很大,特别是在具有大量攻击样本的数据集 *B* 上。说明字符分布模型对正常样本的形式要求很高,在正常样本形式简单和规则的条件下性能尚可,否则检测性能将大幅下降。

枚举类型在数据集 *A* 和 *B* 上性能差不多,这个模型容易产生很高的误报率。这是因为枚举类型模型不具有样本的泛化能力,只是机械地记录学习过的样本,遇到新的样本一律判定为异常行为。枚举类型模型在数据集 *C* 上的 ROC 曲线是连接点(0,0)和点(1,1)点的直线,表明模型对数据集 *C* 的数据没有任何分类能力。这是因为模型在检验数据类型是否为枚举类型时假设检验失败,从而判定模型无法检测数据集 *C* 的数据。

文本聚类方法在 3 个数据集上均表现出了最高的检测率和几乎为 0 的误报率,证实了聚类方法的有效性。在数据集 *B* 上有大量攻击样本的存在,攻击样本和正常样本的比率很高,这不满足 2.1 小节中二者比率很小的算法成立假立,说明虽然算法成立的假设是很严格的条件,但在适当放宽其中的一部分条件时,聚类算法仍然能够有效地检测攻击样本。

本文方法在实验中比经典模型显示出检测率高、误报率低和稳定性高的特点,并且无需事先训练。这是因为算法设计运用了现实数据统计规律,相当于使用了样本分布的先验知识,从而使得本文方法在 3 个实验中可以进行有效的检测。

4　结束语

针对 Web 攻击检测研究表明,正常的 Web 访问行为占访问样本的大多数且行为模式类似、具有很好的聚类性;而攻击行为占访问样本少数且行为模式各异、聚类性差。基于这一规律,提出了一种基于文本聚类的非监督检测算法。另外,本文算法在 NJUST 现实数据集上取得了较高的检测率,从而证实了该方法的有效性。

本文算法基于非监督学习思想,省去了多数异常检测方法中复杂的学习训练过程,增强了方法的应用性。该方法尚存的缺点是聚类速度较慢,对于边界样本难于判断其类别,这将在未来的工作中研究改进。

参考文献:

[1] CHRISTEY S, MARTIN R A. Vulnerability type distributions in CVE [EB/OL]. [2011-08-20]. http://cwe.mitre.org/documents/vuln-trends.html.

[2] FIELDING R, GETTYS J, MOGUL J, et al. RFC-2616: hypertext transfer protocol-HTTP/1.1[S]. Montreal: Internet Engineering Task Force (IETF), 1999.

[3] INGHAM K L, SOMAYAJIB A, BURGEA J, et al. Learning DFA representations of HTTP for protecting web applica-

tions[J]. Computer Networks, 2007, 51(5): 1239-1255.

[4]CORONA I, ARIU D, GIACINTO G. HMM-Web: a framework for the detection of attacks against web applications [C]//IEEE International Conference on Communications. Dresden, Germany, 2009: 1-6.

[5]DURY A, HALLAL H H, PETRENKO A. Inferring behavioural models from traces of business applications [C]// IEEE International Conference on Web Services. Los Angeles, USA, 2009: 791-798.

[6]BACE R. Intrusion detection[M]. [S.l.]: Macmillan Publishing Co. Inc., 2000: 1-4.

[7]ROESCH M. Snort-lightweight intrusion detection for networks[C]//Proceedings of the 13th USENIX Conference on System Administration. Seattle, USA, 1999: 229-238.

[8]CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): artical no. 15.

[9]KRUEGEL C, VIGNA G. Anomaly detection of web-based attacks[C]//Proceedings of the 10th ACM Conference on Computer and Communications Security. Washington, DC, USA: ACM, 2003: 251-261.

[10]KRUEGEL C, VIGNA G, ROBERTSON W. A multi-model approach to the detection of web-based attacks[J]. Computer Networks, 2005, 48(5): 717-738.

[11]PORTNOY L, ESKIN E, STOLFO S. Intrusion detection with unlabeled data using clustering[C]//Proceedings of ACM CSS Workshop on Data Mining Applied to Security. Philadelphia, USA, 2001: 5-8.

[12]MAHONEY M V, CHAN P K. Learning nonstationary models of normal network traffic for detecting novel attacks [C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2002: 376-385.

[13]WARRENDER C, FORREST S, PEARLMUTTER B. Detecting intrusions using system calls: alternative data models[C]//Proceedings of IEEE Symposium on Security and Privacy. Oakland, USA, 1999: 133-145.

[14]SENGAR H, WIJESEKERA D, WANG H, et al. VoIP intrusion detection through interacting protocol state machines[C]//Proceedings of International Conference on Dependable Systems and Networks. Philadelphia, USA: IEEE/IFIP, 2006: 393-402.

[15]周东清,张海锋,张绍武,等.基于HMM的分布式拒绝服务攻击检测方法[J].计算机研究与发展, 2005, 42(9): 1594-1599.
ZHOU Qingdong, ZHANG Haifeng, ZHANG Shaowu, et al. A DDos attack detection method based on hidden Markov model[J]. Journal of Computer Research and Development, 2005, 42(9): 1594-1599.

[16]INGHAM K L, INOUE H. Comparing anomaly detection techniques for HTTP[C]//Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection. Gold Goast, Australia, 2007: 42-62.

[17]JULISCH K. Clustering intrusion detection alarms to support root cause analysis[J]. ACM Transactions on Information and System Security, 2003, 6(4): 443-471.

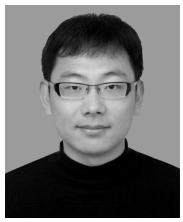
[18]HAINES J W, LIPPMANN R P, FRIED D J, et al. 1999 DARPA intrusion detection system evaluation: design and procedures, TR-1062[R]. Lexington, USA: Lincoln Laboratory, Massachusetts Institute of Technology, 2001.

[19]LIPPMANN R P, HAINES J W, FRIED D J, et al. The 1999 DARPA off-line intrusion detection evaluation[J]. Computer Networks, 2000, 34(4): 579-595.

[20]The UCI KDD Archive. KDD Cup 1999 data[EB/OL]. (1999-10-28)[2011-08-20]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

作者简介:

杨晓峰,男,1982年生,博士研究生,主要研究方向为网络安全、机器学习。



李伟,男,1978年生,博士,主要研究方向为复杂网络、模式识别、机器学习。

孙明明,男,1981年生,讲师,主要研究方向为模式识别、机器学习。