

基于模糊识别恶意代码检测技术的研究

王小英, 陈丹琪, 刘庆杰, 潘志安

(防灾科技学院, 河北 三河 065201)

摘 要: 在恶意代码检测的过程中, 假设恶意代码隐藏的较深, 很难对恶意代码特征进行完整、准确的提取. 利用传统算法进行恶意代码检测, 恶意代码的分布情况都是未知的, 没有充分考虑到不同类别代码特征之间的差异性, 降低了恶意代码检测的准确性. 为此, 提出基于模糊识别的恶意代码检测方法. 根据支持向量机相关理论, 提取恶意代码特征, 并将上述特征作为恶意代码识别的依据. 建立模糊识别辨别树, 计算识别对象属于恶意代码的概率, 实现恶意代码的检测. 实验结果表明, 利用改进算法进行恶意代码检测, 能够极大提高检测的准确性.

关键词: 模糊识别; 恶意代码; 支持向量机

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2014)06-0189-04

Based on the Fuzzy Recognition of Malicious Code Detection Technology Research

WANG Xiao-ying, CHEN Dan-qi, LIU Qing-jie, PAN Zhi-an

(Institute of Disaster Prevention, Sanhe 065201, China)

Abstract: In the process of malicious code detection, it is assumed that the malicious code hidden deep, it is difficult to complete and accurate of malicious code features are extracted. Malicious code detection using the traditional algorithm, the distribution of malicious code is unknown, without fully considering the characteristic differences between different categories code, reduces the accuracy of the malicious code detection. To this end, the malicious code detection based on fuzzy recognition method is proposed. According to the theory of support vector machine (SVM) related to extract characteristics of malicious code, and will be the basis of the above characteristics as malicious code identification. Establish a fuzzy recognition to identify tree, calculated to identify objects belong to the probability of malicious code, for detecting malicious code. The experimental results show that the improved algorithm of malicious code detection, can greatly improve the accuracy of detection.

Key words: fuzzy recognition; malicious code; support vector machine (SVM)

1 引言

随着网络技术的快速发展, 给人们的生活带来了很大的方便, 但随之也带来负面的影响, 比如: 网络中恶意欺诈行为等, 这主要是网络恶意代码连续不断的增加造成的^[1]. 若网络被恶意代码攻击, 用户的电脑可能成为肉机, 用户的资料信息等可能泄密, 导致意想不到的后果, 针对公司而言, 这种攻击带来的危害, 显得更加严重, 比如: 可能造成公司所存客

户银行卡等信息泄露, 给公司带来致命打击^[2]. 因此, 对这种危害的预防显得十分重要, 预防的本质是对网络中恶意代码的检测, 近来, 各国该领域专家都对此进行了深入的研究, 取得了一定的成效^[3]. 近年来, 通过研究分析发现, 各国专家提出的恶意代码检测方法有典型的有: 基于 K 均值聚类算法的检测方法, 基于最小二乘法的恶意代码检测方法以及基于人工智能私密敏感挖掘的检测方法^[4-6]. 这三类方法中, 以最后一种最为常用^[7]. 因为恶意代码的检测还

收稿日期: 2013-09-12; 修回日期: 2013-11-01

基金项目: 中央高校基本科研业务费专项资金(ZY20140214); 中国地震局教师科研基金项目(20120108)

有很多不完善的地方,因此,该领域诸多专家都在进行完善^[8].

基于人工智能私密敏感挖掘的检测方法^[9]是最常用到的恶意代码检测方法,但随着恶意代码制造者不断的迭代更新代码库,以基于人工智能私密敏感挖掘的检测方法检测的准确性越来越低,尤其对伪装的恶意代码检测准确率 20% 以下,因此,需要对传统方法进行改进^[10].

在传统算法的基础上,本文提出了基于模糊识别的恶意代码检测方法.根据支持向量机相关理论,提取恶意代码特征,并将上述特征作为恶意代码识别的依据.建立模糊识别辨别树,计算识别对象属于恶意代码的概率,实现恶意代码的检测.通过仿真实验,结果表明,由于模糊识别的引入,本文算法的恶意代码准确检测率大大提高,应用范围很大.

2 恶意代码检测基本原理

在恶意代码检测过程中,包含大量的网络操作行为数据,这些数据中包括 n 个操作变量,可以用 $\{y_1, y_2, \dots, y_n\}$ 表示,这些变量对应的变量阈值能够用 a_j 表示,则网络操作行为数据简化数学模型可以用下述公式进行描述:

$$\begin{cases} G_1 = b_{11} a_1 + b_{12} a_2 + \dots + b_{1n} a_n \\ G_2 = b_{21} a_1 + b_{22} a_2 + \dots + b_{2n} a_n \\ \dots \quad \dots \quad \dots \\ G_n = b_{n1} a_1 + b_{n2} a_2 + \dots + b_{nm} a_n \end{cases} \quad (1)$$

式中, G_j 与 G_k 之间,相关性很弱,网络操作行为数据中的主成分用 G_j 表示,通过 G_j ,网络操作特征权值系数矩阵表示如下:

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix} \quad (2)$$

敏感性数据挖掘方法的引入,可以有效地检测网络操作行为中的恶意代码,对检测到的恶意代码进行处理,可以保障网络用户的数据不受到侵犯.

本文引入网络操作特征相关数据的敏感性概念,它是恶意代码中比较有代表性的特征,是对恶意代码特征的约束.对此概念的描述用敏感性变化率,表示如下:

$$X_{C(y_j)} = \lim_{\Delta C(y_j) \rightarrow 0} \frac{\Delta F_{\text{avg}}}{\Delta C(y_j)} = \frac{\partial F_{\text{avg}}}{\partial C(y_j)} \quad (3)$$

敏感性变化率求出后,恶意代码特征值的计算如下式表示:

$$X_{C(y_j)} = \lim_{\Delta C(y_j) \rightarrow 0} \frac{\Delta F_{\text{avg}}}{\Delta C(y_j)} = \frac{\partial F_{\text{avg}}}{\partial C(y_j)} \quad (4)$$

要获取恶意代码的敏感性,最重要的是获取恶意代码特征的归属概率,其公式如下所述:

$$X_{C(y_j)} = \frac{\sum_j [\Delta C_{10}(l) + \Delta C_{01}(l)]}{\Delta c(y_j)} \quad (5)$$

恶意代码之间的特征具有弱相关性,这些特征之间可以近似构成线性组合,公式描述如下:

$$F = F_{\text{avg}} + \sum_{j=1}^{J_{\text{num}}} \Delta \chi(y_{y_j}) X_{C(y_j)} \quad (6)$$

根据上面阐述的方法,能够对恶意代码进行检测,从而高效率的识别出恶意代码,从而增加网络的安全性,对用户数据提供安全保障.

3 恶意代码检测优化方法的提出

利用传统算法进行恶意代码检测,无法避免由于恶意代码隐藏较好造成的恶意代码特征之间存在较大差异性的缺陷.为此,提出基于模糊识别的恶意代码检测方法.

3.1 提取恶意代码特征

模糊支持向量机方法是一种以统计学为基础的数据分类方法,利用这种方法进行恶意代码检测,需要用到超平面数据分类的相关知识.

利用下述公式,能够计算恶意代码二次规划的对偶规划:

$$\begin{cases} \max \sum_{k=1}^m \beta_k - \frac{1}{2} \sum_{j=1}^m \sum_{k=1}^m z_j z_k \beta_j \beta_k (y_j \cdot y_k) \\ \text{s. t. } \sum_{k=1}^m z_k \beta_k = 0 \\ 0 \leq \beta_k \leq v(y_k) D k = 1, 2, \dots, m \end{cases} \quad (7)$$

根据上面阐述的方法,能进行一次转换,即把最优超平面问题转换为二次规划中,对偶规划的求解问题.上述对偶规划的最优解表示如下:

$$\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_m^*)^T \quad (8)$$

对偶规则的最优解求出之后要,模糊最优分类函数可以利用下述公式描述:

$$g(y) = \text{sgn}\{(x^* \cdot y) + c^*\} \quad (9)$$

通过分析可知,模糊最优分类函数要满足如下约束:

$$\begin{aligned} x^* &= \sum_{k=1}^m \beta_k^* z_k y_k \\ b^* &= y_i - \sum_{j=1}^l y_j \alpha_j (x_j \cdot x_i) \\ j &\in \{j | 0 < \beta_j^* < v(y_j) D\} \end{aligned} \quad (10)$$

利用下述公式,能够获取模糊支持向量机最优分类函数:

$$g(y) = \operatorname{sgn}\left\{\sum_{k=1}^m \beta_k^* z_k L(y, y_k) + c^*\right\} \quad (11)$$

式中:

$$c^* = z_j - \sum_{k=1}^m z_k \beta_k L(y_k, y_j) \\ j \in \{j | 0 < \beta_j^* < v(y_j) D\} \quad (12)$$

根据上面阐述的方法,能够对恶意代码特征进行分类处理,从而检测出恶意代码特征。

3.2 实现恶意代码检测

设置 p 用于描述全部恶意代码的数量, $\{t_1, t_2, \dots, t_p\}$ 用于描述恶意代码构成的数据集。这些恶意代码特征的关联性可以用 $P(t_1), P(t_2), \dots, P(t_p)$ 表示。利用下述公式能够计算恶意代码特征的信息熵:

$$I(T) = q_1 Y(t_1) + q_2 Y(t_2) + \dots + q_p Y(t_p) \\ = - \sum_{k=1}^p q(t_k) \log_2 q(t_k) \quad (13)$$

设置 ω 表示粒子群辨别树恶意代码的属性数量,对上述子属性分割处理,得到大量的子区域。相应子区域的信息熵计算公式如下:

$$L(b_i) = - \sum_{k=1}^m q_k \log_2(q_k) \quad (14)$$

根据下述公式可以计算恶意代码在模糊识别时的信息量:

$$L(r, n) = - \frac{r}{r+n} \log_2 \frac{r}{r+n} - \frac{n}{r+n} \log_2 \frac{n}{r+n} \\ B(D) = \sum_{k=1}^e \frac{r_k + n_k}{r+n} L(r_k, n_k) \quad (15)$$

利用下述公式进行恶意代码的模糊识别:

$$\operatorname{win}(D) = L(r, n) - B(D) \quad (16)$$

根据上面阐述的方法,能够建立粒子群辨别树,通过运算获取恶意代码出现的概率,从而实现恶意代码的模糊识别。

4 实验结果及分析

利用 MATLAB 进行恶意代码检测仿真实验。实验环境能够用表 1 进行描述。

表 1 实验环境数据表

名称	型号
CPU	P4 3.0
内存	4 GB
操作系统	Windows XP
编程语言	JAVA

分别利用传统算法和改进算法进行恶意代码检测,检测的效率能够用图 1 进行描述。

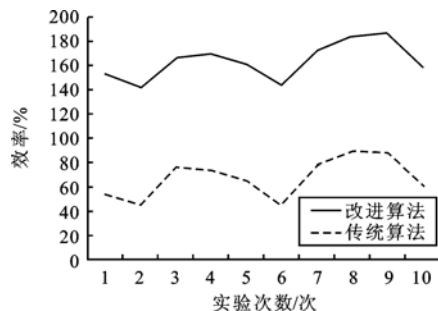


图 1 不同算法的检测效率对比

将上述实验数据进行整理分析,能够得到表 2。

表 2 不同方法的检测效率对比

实验次数	传统算法	改进算法
1	54%	98%
2	46%	96%
3	77%	89%
4	74%	95%
5	65%	95%
6	45%	99%
7	78%	94%
8	89%	94%
9	88%	98%
10	61%	97%

根据上述实验能够得知,随着恶意代码检测时间的不断延长,利用不同算法进行恶意代码检测的效率都在降低。

为了进一步验证改进算法的有效性,针对不同类型的恶意代码进行实验。恶意代码类型能够用表 3 进行描述。

表 3 恶意代码类型数据表

恶意代码类型	数量/个
И	45
Й	44
III	37
II	66

针对上述恶意代码类型,分别利用传统算法和改进算法进行恶意代码检测,检测的结果能够用图 2 进行描述,根据图中结果可知:本文算法恶意代码检测准确率比传统算法平均高出 20%,并且随着恶意代码类型的不断增加,改进算法的检测准确率更

加优越于传统算法。

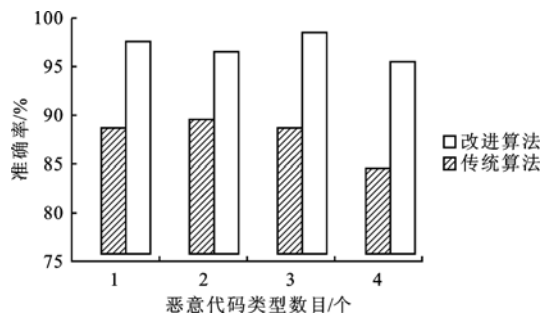


图 2 不同算法检测准确率对比

在上述实验过程中,利用不同算法进行恶意代码检测的实验结果能够用表 4 进行描述。

表 4 不同方法的检测结果对比 %

衡量标准	传统算法	改进算法
平均误报率	46	9
平均检测率	37	7
平均漏检率	29	4

根据上述实验能够得知,利用改进算法进行恶意代码检测,检测的效率和准确率都非常高,充分展示了改进算法的优越性,说明改进算法能够为用户提供优质的网络安全服务。

5 结束语

针对传统算法无法避免由于恶意代码隐藏较好造成的恶意代码特征之间存在较大差异性的缺陷,提出基于模糊识别的恶意代码检测方法。根据支持向量机相关理论,提取恶意代码特征,并将上述特征作为恶意代码识别的依据。建立模糊识别辨别树,计算识别对象属于恶意代码的概率,实现恶意代码的检测。实验结果表明,利用改进算法进行恶意代码检测,提高检测效率,降低检测误差,满足网络运行的

安全需求。

参考文献:

- [1] Perrig A, Stankovic J, Wagner D. Security in wireless sensor networks[J]. CACM, 2004(47):53-57.
- [2] Onat I, Miri A. An intrusion detection system for wireless sensor networks[J]. Wireless And Mobile Computing Networking And Communications, 2005(3):253-259.
- [3] 史晓红, 张艳宜. 机器学习应用于恶意代码检测的研究[J]. 科技通报, 2013, 10(29): 21-23.
- [4] B Waters. Efficient identity-based encryption without random oracles[C]// Advances in Cryptology-EUROCRYPT 2005. Berlin: Springer-Verlag, 2005:114-127.
- [5] 王丽苹, 安娜, 吴晓南, 等. 入侵检测系统中的行为模式挖掘[J]. 通信学报, 2004, 25(7):168-172.
- [6] 罗敏, 王丽娜, 张焕国. 基于无监督聚类的入侵检测方法[J]. 电子学报, 2003, 31(11):1713-1716.
- [7] 刘韬, 皮国强. 人工免疫算法在网络入侵检测中的应用[J]. 计算机仿真, 2011(11):91-94.
- [8] 肖海军, 洪帆, 张昭理, 等. 基于融合分类和支持向量机的入侵检测研究[J]. 计算机仿真, 2008, 25(4):130-132.
- [9] 卿斯汉, 蒋建春, 马恒太, 等. 入侵检测技术研究综述[J]. 通信学报, 2004, 25(7):19-21.
- [10] 汪兴东, 余堃, 周明天, 等. 基于 BP 神经网络的智能入侵检测系统[J]. 成都信息工程学院学报, 2005, 20(1):1-4.

作者简介:

王小英 女, (1979-), 硕士研究生, 讲师. 研究方向为网络与信息安全、密码学。

陈丹琪 女, (1972-), 硕士研究生, 副教授. 研究方向为软件工程、软件测试。

刘庆杰 男, (1978-), 讲师. 研究方向为网络工程。

潘志安 男, (1983-), 硕士研究生, 讲师. 研究方向为网络安全。