基于区间值2型模糊集的伪装入侵检测算法

曾剑平[1], 郭东辉[1,2]
(1. 厦门大学物理系 EDA 实验室, 福建厦门 361005; 2. 厦门大学电子工程系, 福建厦门 361005)

摘 要: 由于正常用户的行为本身是变化的, 且伪装用户的行为可能看起来是正常的, 这种不确定性使得现有的伪装检测算法很难正确判断用户身份的真实性, 从而限制了现有算法的实际应用推广. 本文合理地选择用户的行为特征, 并建立相应的用户可信度计算方法, 采用区间值模糊集对多个特征进行可信度综合计算得到用户的最终可信度, 将该值与设定的阈值比较从而判断用户是否属于伪装. 理论分析及实验结果显示, 与普通模糊计算相比, 区间值模糊计算能有效表示及处理伪装检测中的不确定性, 因而能得到比较理想的检测效果.

关键词: 伪装检测; 区间值模糊集; 不确定性; 可信度
中图分类号: TP309 文献标识码: A 文章编号: 0372-2112 (2008) 04-0777-04

Masquerade Intrusion Detection Algorithm Based on Interval Type-2 Fuzzy Set

ZENG Jiarr-ping[1], GUO Dong-hui[1,2]
(1. *EDA Laboratory, Physics Department, Xiamen University, Xiamen, Fujian 361005, China;*
2. *Department of Electronic Engineering, Xiamen University, Xiamen, Fujian 361005, China*)

Abstract: User activities are normally in variant forms or may be aberrant in some cases. This kind of uncertainty leads to the difficulty for current intrusion detection algorithm in deciding whether the user is masquerading or not. In the proposed algorithm, the user features are properly selected and the corresponding user trustworthiness computation methods are introduced. Different types of trustworthiness are integrated with interval type-2 fuzzy set, thus user trustworthiness is got and applied to a threshold-based decision. Theory analysis and experiments show that the proposed algorithm can handle the uncertainties that exist in user actirity or user model, so better detection performance can be achieved, compared with the detection algorithm based on ordinal fuzzy set.

Key words: masquerade detection; interval type-2 fuzzy set; uncertainty; trustworthiness

1 引言

伪装入侵所造成的潜在危害比针对网络或主机所产生的破坏要严重得多. 目前被人们用以建立用户行为模型有: HMM[2,7], SVM[4], ECM[5], 序列数据库模型[6,7], Markov[8]等. 这些模型大都是以正常用户访问应用服务系统时应用程序的系统调用和操作的行为序列数据为基础的. 2001 年 Schonlau 等人[8]为伪装入侵检测的研究提供了一个真实环境下的系列调用数据集, 并在这个数据集上进行了许多测试分析工作. 他们采用了以 Markov 模型为主的几种不同建模方法来分析用户行为的某个特征, 所获得的检测效果均不理想. 从检测率看, 最好的结果是 69.6%, 但是相应的误报率也达到了 6.7%; 从误报率看, 最小的是 1.4%, 但相应的检测率只有 39.4%. 随后, 卡内基梅隆大学的 R. Maxion 教授等人也提出了一些改进模型算法[9], 但其对用户行为单一特征检测的效果改进还是很有限的. 这些工作说明, 采用单一用户特征的检测算法很难在实际应用中取得较好的检测效果.

对用户的多个行为特征进行检测分析通常可以比较好地解决这个问题, 如文献[10, 11]针对网络的入侵检测, 从 TCP 数据流中提取两种不同特征, 并用普通模糊逻辑对结果进行综合; 文献[12]从安全审计记录中提取了系统调用、文件操作以及用户进程信息三种特征, 并分别建立了三个不同的用户模型, 最后采用普通模糊逻辑对三个输出进行融合; 文献[13, 14]针对网络入侵中协议数据流, 采用模糊神经网络对多个判断结果进行综合. 在这些检测算法中, 规则中模糊变量的隶属函数往往不容易确定, 另一方面, 由于伪装行为在某些时刻与正常用户的行为非常相似, 这种不确定性也直接限制了检测效果的提高.

为了有效解决单特征检测算法的缺点以及处理上述不确定性, 本文采用多特征, 建立相应的可信度计算方法, 并提出了采用基于区间值 2 型模糊集对这些可信度进行融合, 得到最终的用户可信度. 区间值隶属度函数在表达上更符合实际需要, 判断、推理所产生的结果用区间值模糊集表示更能反映人类推理的模糊性和不确定性[1,3], 更能适应于伪装检测中的不确定性处理.

2 区间值2型模糊集的表示及计算

2 型模糊集是 Zadeh 于 1975 年提出的[15], 是普通模糊集(又称 1 型模糊集)的扩展. 基于 2 型模糊集的计算

收稿日期: 2006-06-19; 修回日期: 2007-12-02
基金项目: 国家人事部留学人员创业基金; 福建省自然科学基金(No. A0410007)

通常能得到比 1 型模糊集计算更合理的结果. 而区间值 2 型模糊逻辑[1]是一种特殊的 2 型模糊逻辑, 是一种比较实用的 2 型模糊逻辑.

2.1 基于区间值的模糊集表示

闭区间[0, 1]中的某个闭区间 $a=[a^-, a^+]$ $(0\leqslant a^-\leqslant a^+\leqslant 1)$ 称为[0, 1]上的区间值. 可以用集合[I]表示[0, 1]上区间值的全体, 则区间值 2 型模糊集的定义[1]如下表述, 即:

定义 **1** 设 X 是一非空普通集合, 称映射 $A: X\rightarrow [I], x\rightarrow [A^-(x), A^+(x)]$ 为 X 上的区间值 2 型模糊集, 其中 $x\in X$, $A^-(x)\leqslant A^+(x)$, $A(x)$ 是区间值 2 型模糊集的隶属函数, $A^-(x)$ 与 $A^+(x)$ 分别表示模糊集的区间值$[A^-(x), A^+(x)]$的下界与上界.

如果采用积分符号表示区间值模糊集, 则 A 可以写成:

$$A=\int_{x\in X}[A^-(x), A^+(x)]/x$$

其中, 积分号表示论域 X 中任意值 x 对应的模糊集.

2.2 基于区间值 **2** 型模糊集的计算

交、并运算是两种主要的模糊集合运算, 其中对于两个定义在论域 X 上的区间值 2 型模糊集 A_1 和 A_2 的交运算可表示为[16]:

$$A_1(x_1)\cap A_2(x_2)=[A_1^-(x_1)\cap A_2^-(x_2), A_1^+(x_1)\cap A_2^+(x_2)] \tag{1}$$

其中: x_1, x_2 分别是 X 中的值.

如果 T 是[0, 1]上的 t-模算子, 且 $a=[a^-, a^+]$, $b=[b^-, b^+]$ 分别为[0, 1]上的两个区间值, 则[0, 1]上的区间值 t-模算子的数学表达式可以如下表示[3]:

$$T^{(i)}(a, b)=[T(a^-, b^-), T(a^+, b^+)] \tag{2}$$

这样, 类似于式(1)的区间值 2 型模糊集复合运算可以表示为:

$$C(A_1, A_2)=\int_{x\in X}T^{(i)}(A_1(x), A_2(x))/x \tag{3}$$

假设 A_1、A_2、A_3 是三个区间值 2-型模糊集, 则它们的复合运算结果与顺序无关, 即有:

$$C(C(A_1, A_2), A_3)=C(C(A_1, A_3), A_2)=C(C(A_2, A_3), A_1) \tag{4}$$

3 算法设计与分析

为了提高伪装检测算法的检测性能, 我们选择命令特征(SEQ)、用户登录的计算机标识(CID)、用户登录的时间点(LTIME). 这三个特征之间的相关性小, 有利于提高检测算法的性能.

3.1 算法描述

在确定好三个用户特征之后, 伪装检测算法首先根据这三个特征计算得到三个不同的用户可信度, 即可以将特征值映射到一个相同的域上, 即: 可信度. 接着, 在该域上对不同的可信度进行模糊融合, 最后对融合结果进行反模糊化处理并与设定的阈值进行比较判断, 得到用户是否是伪装的结论. 算法描述如下:

算法 **1** 基于区间值模糊计算的伪装检测算法

输入: CID, LTIME, SEQ 的特征值、用户模型、阈值

处理:

(1) 根据用户模型及三个特征值分别计算相应的用户可信度

(2) 将可信度转换成三个区间值模糊集

(3) 运用复合运算定理进行区间值模糊集的融合运算, 得到一个可信度区间值模糊集

(4) 采用重心法对模糊集进行反模糊化处理得到用户可信度

(5) 将用户可信度与阈值进行比较, 得到检测结果

输出:

当前用户是否是伪装.

下面, 将介绍这个算法中的几个关键问题, 包括: 用户特征及其可信度模糊计算、可信度的模糊融合等以及参数取值分析.

3.2 用户特征及其可信度的模糊计算

假设 CID 的概率分布为 $P_1(X)$, 则某次登录点 x 对应的可信度定义为:

$$Tr_1(x)=\frac{P_1(X=x)}{\max\limits_{\text{所有}x}(P_1(X=x'))} \tag{5}$$

假设已知某个用户在各个登录时间段$[ti_1, ti_2]$的概率分布, 记为 $P_2(X)$. 设某用户的登录时间点为 x, 则相应的可信度定义为:

$$Tr_2(x)=\frac{P_2(ti_1<x\leqslant ti_2)}{\max\limits_{\text{所有}ti_1, ti_2}(P_2(ti_1<x\leqslant ti_2))} \tag{6}$$

由序列数据建立用户的 HMM(Hidden Markov Model)模型, 记为 λ 则对于某次观察到的序列 x, 它对应的可信度 $Tr_3(x)$ 按照下式计算:

$$Tr_3(x)=\frac{P(x|\lambda)-\min\limits_{\text{所有}x}(P(x'|\lambda))}{\max\limits_{\text{所有}x'}(P(x'|\lambda))-\min\limits_{\text{所有}x'}(P(x'|\lambda))} \tag{7}$$

其中, x' 是一个长度与 x 相同的序列, $P(x|\lambda)$ 是序列 x 相对于模型的概率.

当正常用户的数据中包含伪装记录、或者不完整时, 由此而建立的用户模型通常是不准确的, 因此, 特征值对应的可信度就存在不确定性, 为了尽量减小这种不确定性对检测算法的影响, 将上述可信度转换成区间值模糊集的形式, 以便对这种不确定性进行处理. 定义可信度区间值模糊集:

定义 **2** 可信度区间值模糊集是具有如下形式隶属度函数的区间值 2 型模糊集,

$$UC(x, \xi\ \sigma, k)=[UC^-(x, \xi\ \sigma-k), UC^+(x, \xi\ \sigma+k)] \tag{8}$$

其中　$UC^{-}(x, \xi, \sigma, k) = e^{-\frac{(x-\xi)^2}{2\times(\sigma-k)^2}}$

$UC^{+}(x, \xi, \sigma, k) = e^{-\frac{(x-\xi)^2}{2\times(\sigma+k)^2}}$

ξ反映了不同的可信程度, $0 \leqslant \sigma, k \leqslant 1$, σ和k 决定了隶属函数的区间下界与上界值.

根据定义 2, 对于三个观测到的特征值 x_1, x_2, x_3, 可以将它们转换成用户可信度论域上的模糊集, 即:

$$UC_1(x) = UC(x, Tr_1(x_1), \sigma_1, k_1) \quad (9)$$

$$UC_2(x) = UC(x, Tr_2(x_2), \sigma_2, k_2) \quad (10)$$

$$UC_3(x) = UC(x, Tr_3(x_3), \sigma_3, k_3) \quad (11)$$

其中, $\sigma_1, k_1, \sigma_2, k_2, \sigma_3, k_3$ 与相应的特征值及用户模型所表现出的不确定性是相对应的.

3.3　用户可信度的模糊融合

用户可信度模糊融合对可信度论域 x 上的三个区间值模糊集 $UC_1(x)$、$UC_2(x)$、$UC_3(x)$综合运算, 得到一个可信度模糊集. 这个计算过程可以看作是三个模糊集的复合运算.

根据复合运算, 可以写出融合的计算公式,

$$UC' = C(C(UC_1, UC_2), UC_3)$$

3.4　不确定性参数取值分析

可信度区间值模糊集中的 σ_1, σ_2, σ_3 所确定的模糊集与普通模糊计算中的模糊集是一致的, 反映了特征值的模糊性. 而 k_1, k_2, k_3 与相应的用户模型中所存在的不确定性是对应的. 因此, 算法 1 不但可以处理特征值本身的不确定, 也可以处理用户模型的不确定性, 从而与普通模糊计算相比, 具有更强的适应能力, 得到更合理的计算结果.

当 $k_1 = 0, k_2 = 0$ 或 $k_3 = 0$ 表示相应的模型不存在不确定性, 此时 σ_1、σ_2 与 σ_3 所描述的隶属函数与普通模糊计算中的隶属函数是一样的. 因此, σ_1、σ_2 与 σ_3 的取值带有主观因素, 它们分别反映了用户特征 CID, LTIME, SEQ 所存在的不确定性, 一般可以取较小(小于 0.2)的数. 而 k_1, k_2, k_3 表示模型中所存在的不确定性, 这种不确定性可以通过学习到的模型与真正的用户模型之间的 *KL* 散度来描述, 一般可以取小于 0.15 的数.

4　实验及结果分析

4.1　实验设计及测试数据说明

为了检验本文算法的有效性, 我们在实际环境中提取文中所定义的三种事件, 收集了 50 个用户的事件数据, 记为数据集 TCS, 该数据集包含 100 个记录, 前 50 个记录(TCS-1) 作为模型训练时用, 而后 50 个记录(TCS-2) 作为测试数据. 接着对 TCS-2 数据集进行人工处理, 产生模拟的伪装入侵数据. 产生了 50 个用户的伪装入侵数据集 TCS-21. 为了测试特征选择的有效性, 我们从 TCS 数据集中单独将 SEQ 抽取出来, 并按照类似于 TCS-21 的方式产生另一个测试数据集 TCS-22.

下面的实验是在 TCS-21 和 TCS-22 上进行的. 文献[11] 采用普通模糊逻辑对多个模型的检测结果进行综合, 本文引用它的算法(以下简称普通模糊计算) 进行比较实验. 此外, 我们还与文献[4~8] 中采用单特征检测算法(以下简称单特征) 进行实验比较.

4.2　实验结果及分析

实验一是在 TCS-21 数据集上执行伪装检测, 实验结果如图 1 中多特征部分所示. 图中横坐标表示用户编号, 纵坐标表示检测性能, 它是通过 ROC-1 来表示的. ROC-p 表示 ROC 曲线与误报率等于 p% 的直线构成的区域的面积[4]. 该值越大表示性能越好. 在实验中, 算法 1 的阈值从 0 到 1 增加, 从而得到 ROC-1.

第二个实验是在数据集 TCS-22 上进行的, 检测结果反映在图 1 中的单特征部分. 可以看出, 采用多特征的区间值模糊计算时, 有 41 个用户的检测性能好于单特征的检测方法, 而其余的用户的检测效果大部分与单特征的检测效果类似. 总体上看, 多特征区间值模糊计算检测算法对不同的用户具有很好的适应能力, 具有更好的检测效果.

为了比较区间值模糊计算与普通模糊计算对检测性能的影响, 分别选择 $\sigma_1 = 0.15$, $\sigma_2 = 0.1$, $\sigma_3 = 0.1$, $k_1 = 0.05$, $k_2 = 0$, $k_3 = 0$, 并在 TCS-21 上进行检测实验; 另一方面, 取 $\sigma_1 = 0.15$, $\sigma_2 = 0.1$, $\sigma_3 = 0.1$, $k_1 = 0$, $k_2 = 0$, $k_3 = 0$, 也在 TCS-21 上进行检测实验. 在这个实验中, 对学习好的用户模型 $P_1(X_1)$ 进行人为修改, 产生模型不确定. 结果如图 2 所示. 可以看出, 采用区间值模糊计算时, 有 40 个用户的检测性能好于普通模糊计算检测方法.

为了说明参数设置的影响, 以 User2 为例, 当 $\sigma_2 =$

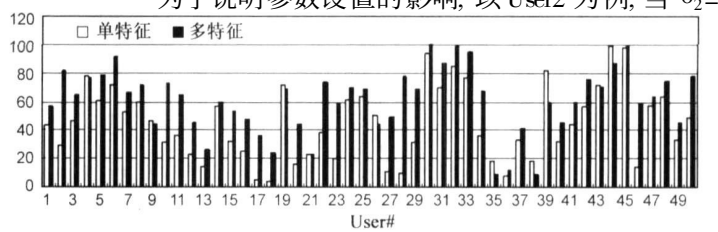

图 1　50 个用户的检测结果

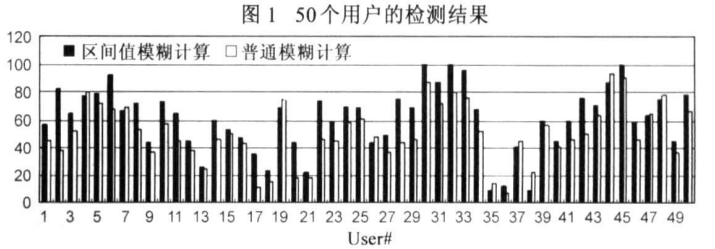

图 2　区间值模糊计算与普通模糊计算

0.1, σ_3= 0.1, k_1= 0.05, k_2= 0, k_3= 0, 而 σ_1= 0.1, 0.15, 0.2, 0.25 时, 在误报率分别为 0.01, 0.03, 0.06, 0.09 时所得到的检测率, 如表 1 所示.

表 1 在不同的 σ_1 及误报率限制下的检测率

σ_1	误报率取值			
	0.01	0.03	0.05	0.07
0.1	0.75	0.83	0.83	0.92
0.15	0.75	0.83	0.83	0.92
0.2	0.67	0.83	0.83	0.92
0.25	0.67	0.83	0.83	0.83

可以看出, 当 σ_1 小于 0.15 时, ROC-1 的值越大, 性能越好.

5 结论

本文提出了一种基于区间值 2 型模糊计算的伪装入侵检测算法, 合理地选择用户的行为特征, 建立相应的用户可信度计算方法, 采用基于区间值模糊计算的方法对不同的可信度进行融合. 这种计算方法允许对模型和数据中的不确定性进行描述及计算, 充分利用了区间值模糊集在处理不确定性问题的优越性. 实验结果也表明, 它能得到比较好的检测性能.

参考文献:

[1] Liang Qi-lian, Mendel J M. Interval type-2 fuzzy logic systems: theory and design[J]. IEEE Transactions on fuzzy systems, 2000, 8(5): 535-550.

[2] 谭小彬, 王卫平, 奚宏生, 殷保群. 基于隐马尔可夫模型的异常检测[J]. 小型微型计算机系统, 2004, 25(8): 1546-1549.
Tan Xiao-bin, Wang Wei-ping, Xi Hong-sheng, Yin Bao-qun. Anomaly detection based on hidden Markov model[J]. Mini-Micro Systems, 2004, 25(8): 1546-1549. (in Chinese)

[3] Mendel J M, John R I B. Type-2 fuzzy sets made simple[J]. IEEE Transactions on Fuzzy Systems, 2002, 10(2): 117-127.

[4] Wang Ke, Salvatore J S. One-class training for masquerade detection[OL]. In Proc. 3rd IEEE Conference Data Mining Workshop on Data Mining for Computer Security[C]. http://www1.cs.columbia.edu/ids/publications/DMSEC-camera.PDF, 2003.

[5] Mizuki Oka, Yoshihiro Oyama, Hirotake Abe, Kazuhiko Kato. Anomaly detection using layered networks based on eigen co-occurrence matrix[A]. LNCS 3224(Recent Advances in Intrusion Detection)[C]. Springer, 2004. 223-237.

[6] Warrender C, Forrest S, Pearlmutter, B. Detecting intrusions using system calls: alternative data models[A]. In Proc. IEEE Symposium on Security and Privacy[C]. Oakland California: IEEE Computer Society Press, 1999. 133-145.

[7] 闫巧, 谢维信, 宋歌, 喻建平. 基于 HMM 的系统调用异常检测[J]. 电子学报, 2003, 31(10): 1486-1490.
Yan Qiao, Xie Wei-xin, Song Ge, Yu Jiarr-ping. System call anomaly detection method based on HMM[J]. Acta Electronica Sinica, 2003, 31(10): 1486-1490. (in Chinese)

[8] Schonlau M, DuMouchel W, Ju W-H, Karr A F, Theus M, Vardi Y. Computer intrusion: detecting masquerades[J]. Statistical Science, 2001, 16(1): 58-74.

[9] Maxion R A, Townsend T N. Masquerade detection augmented with error analysis[J]. IEEE Transactions on Reliability, 2004, 53(1): 124-147.

[10] Dickerson J E, Dickerson J A. Fuzzy network profiling for intrusion detection[A]. In Proc 19th International Conference of the North American Fuzzy Information Processing Society[C]. USA: IEEE Computer Society Press, 2000. 301-306.

[11] Dickerson J E, Juslin J, Koukousoula O. Fuzzy intrusion detection[A]. In Proc 20th NAFIPS International Conference and Joint 9th IFSA World Congress[C]. USA: IEEE Computer Society, 2001. 3. 1506-1510.

[12] Cho Sung-Bae. Incorporating soft computing techniques into a probabilistic intrusion detection system[J]. IEEE Transactions on Systems, Man and Cybernetics (Part C), 2002, 32(2): 154-160.

[13] Mohajerani M, Moeini A, Kianie M. NFIDS: a neuro-fuzzy intrusion detection system[A]. In Proc 10th IEEE International Conference on Electronics, Circuits and Systems[C]. USA: IEEE Computer Society, 2003. 348-351.

[14] Chavan S, Shah K, Dave N. Adaptive neuro-fuzzy intrusion detection systems[A]. In Proc IEEE International Conference on Information Technology: Coding and Computing[C]. USA: IEEE Computer Society, 2004. 70-74.

[15] Zadeh L A, The concept of a linguistic variable and its application to approximate reasoning-1[J]. Information Science, 1975, (8): 199-249.

[16] 陈启浩. 模糊值及其在模糊推理中的应用[M]. 北京师范大学出版社, 北京. 2000. 41-42.

作者简介:

曾剑平 男, 1973 年出生于福建惠安, 博士生, 研究领域是智能计算与信息安全.
E-mail: zeng_jian_ping@hotmail.com

郭东辉 男, 1967 年出生于福建莆田, 厦门大学教授, 博士生导师, 主要研究方向是人工智能、网络通讯、集成电路设计等. (本文通讯作者) E-mail: dhguo@xmu.edu.cn