

基于模糊积分的多分类器组合的入侵检测*

孙 钢, 张 莉, 郭 军

(北京邮电大学 信息工程学院, 北京 100876)

摘 要: 将模式识别的方法应用到入侵检测系统中解决了传统的基于规则的入侵检测方法的缺陷。提出了一种基于模糊积分的多分类器组合的入侵检测方法, 并提出了一种模糊密度的计算方法, 通过 KDD '99 中的入侵检测数据的实验证明, 该方法检测准确性较高。

关键词: 模糊积分; 多分类器组合; 入侵检测

中图法分类号: TP391 文献标识码: A 文章编号: 1001-3695(2005)11-0117-02

Multiple Classifiers Combination Based on Fuzzy Integral in Intrusion Detection

SUN Gang, ZHANG Li, GUO Jun

(School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing 100876, China)

Abstract: As pattern recognition method used in intrusion detection system, the limitation in traditional detection methods based on rules is cancelled. In this paper, a new intrusion detection approach using multiple classifiers combination based on fuzzy integral is proposed. At same time, a new method for computing fuzzy density is proposed. Experiment has been done on intrusion detection dataset in KDD '99 and the result shows that it has high detection accuracy.

Key words: Fuzzy Integral; Multiple Classifiers Combination; Intrusion Detection

1 引言

在网络安全问题日益突出的今天, 如何迅速有效地发现各类新的攻击行为对保证系统和网络资源的安全显得十分重要。入侵检测就是用来发现网络或主机日志文件中已知的或潜在的攻击行为。最近研究者们将模式识别的方法用在入侵检测系统中^[1~3], 可以检测到已有的攻击或新的攻击, 取得了比较好的入侵检测效果。入侵检测作为一个模式识别的任务, 基本流程如图 1 所示, 其实质可以描述为对网络行为尽可能地进行正确分类, 关键问题是特征提取、选择和构造合适的分类器^[4]。

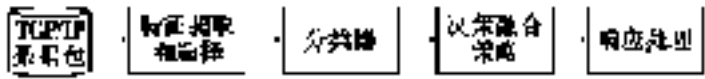


图 1 入侵检测的模式识别流程

在模式识别中, 把多个分类器的输出信息联合起来进行分类决策是解决复杂问题的一种有效方法, 很多学者已对这方面进行了深入的研究。多分类器的组合方式可以分为级联和并联两种方式, Schapire 等人提出的 Boosting 方法是级联形式的多分类器组合方法中的代表; 并联形式的组合方法包括传统的择多判决法(如投票表决策法、计分法等)、线性加权组合方法、模糊推理法以及通过样本特征而动态选择分类器的方法, 或者神经网络融合等。多分类器组合的方法有很多, 但是到目前为止仍然不太成熟, 在应用到具体应用领域时还需要做很多的调整和处理工作。本文提出了一种基于模糊积分的多分类器组合的入侵检测方法, 实验结果显示该方法是有效的。

2 基于模糊积分的多分类器组合

Hossein, Cho 等人提出利用 Sugeno 的模糊积分进行度量级的信息融合^[5,6], 取得了较好的效果。一般把各个分类器的识别率作为模糊积分的密度值, 但是这种方法没有考虑到对于某一分类器来说, 随着输入模式的变化, 其识别性能有很大的差异。模糊积分密度函数的定义不能很好地评价分类器当前的分类性能, 从而影响了整个识别系统的性能。因此在利用模糊积分进行多分类器建模时, 如何定义模糊积分密度函数是一个非常关键的问题。根据入侵检测分类器构造的特点, 提出了一种模糊密度函数的计算方法。

2.1 模糊积分

模糊积分是由 Sugeno 提出的, 它提供了一种有效的信息融合方法。以下先简单介绍模糊积分的一些定义和属性^[5,6]。

定义 1 设 X 是一个有限元素的集合, 一个集合函数 $g: 2^X \rightarrow [0, 1]$ 称为一个模糊度量, 若满足:

$$\begin{cases} g(\Phi) = 0 \\ g(X) = 1 \\ g(A) \leq g(B) \quad A \subset B \end{cases}$$

(1)

根据模糊度量的定义, 介绍一种比较特殊的模糊度量—— g_λ 模糊度量。它满足以下属性: 对所有 $A, B \subset X$ 和 $A \cap B = \Phi$, 有 $g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B)$, ($\lambda > -1$)。

利用模糊度量的概念, Sugeno 提出了模糊积分的概念, 这是一个对模糊度量估计的非线性函数。

定义 2 设 X 是一个有限集合, 令 h 为 X 的一个模糊子集, 则 h 是一个映射, $h: X \rightarrow [0, 1]$ 。令 g 为 X 上的模糊测度, 那么在 X 上函数 h 关于模糊测度 g 的模糊积分定义为

$$h(x) \circ g(\cdot) = \max_{EA \subset X} [\min_{x \in E} (h(x), g(E))] = \max_{\alpha \in [0, 1]} [\min(\alpha, g(h_\alpha))]$$

(2)

其中 h_{α} 是关于 h 的一个 α 度量集, $h_{\alpha} = \{x | h(x) \geq \alpha\}$ 。

模糊积分可以解释为确定以下命题的满足程度: 至少存在一个 X 的子集, 满足条件 g 且其所有元素满足 h_0 。

有限集合的模糊积分可以按以下方法计算: 令 $X = \{x_1, x_2, \dots, x_n\}$ 是一个有限集合, $h: X \rightarrow [0, 1]$ 是一个函数。假设 $h(x_1) \geq h(x_2) \geq \dots \geq h(x_n)$ (如果不是这样, X 重新排列使上面关系成立), 那么 X 上关于模糊测度 g 的模糊积分 e 的计算公式为

$$e = \max[\min(\bigwedge_{i=1}^n h(x_i), g(A_i))], A_i = \{x_1, x_2, \dots, x_i\} \tag{3}$$

当 g 为 g_{λ} 测度时, $g(A_i)$ 的值递推计算公式为

$$\begin{cases} g(A_1) = g(\{x_1\}) = g^1 \\ g(A_i) = g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}), 1 < i \leq n \end{cases} \tag{4}$$

其中 λ 由以下等式给出:

$$\lambda + 1 = \prod_{i=1}^n (1 + \lambda g^i), \lambda \in (-1, +\infty), \lambda \neq 0 \tag{5}$$

因此, 当使用 g_{λ} 模糊测度计算模糊积分时, 仅需要模糊密度的知识, 第 i 个密度值 g^i 可以解释为信息 x_i 的重要程度。

2.2 用模糊积分构造多分类器组合模型

图 2 给出了用模糊积分构造多分类器组合的模型。对于一个模式识别问题, 设 $\omega = \{\omega_1, \omega_2, \dots, \omega_M\}$ 为类别集合, 分类器 i 的输出是对输入模式所属类别的一种评价, 用 $h(y_i)$ 表示。用模糊积分建模时, 还必须确定每个分类器当前的分类性能, 即各个分类器的重要性(密度函数), 用 g^i 表示。根据模糊积分的定义, 模糊积分值为 $e_j, j = 1, 2, \dots, M$ 。系统的决策模型为: 如果 $e_x = \max e_j \geq \alpha$ (α 是设定的门限), 则 $E(x) = j$; 否则 $E(x) = M + 1$ 。

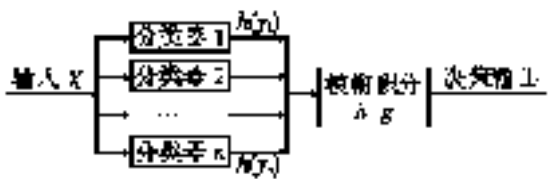


图 2 多分类器组合模型

2.3 模糊密度 g 的计算

由模糊积分的定义和图 2 可知, 确定描述各个分类器重要性的 g^i 是基于模糊积分的多分类器组合的关键问题。目前确定 g^i 的值主要有两种方法: 将训练分类器得到的分类准确度直接作为 g^i 的值; 利用混淆矩阵确定 g^i 的值^[7]。第一种方法仅仅是从整体上描述每个分类器的性能, 不能反映出不同输入时其识别性能的差异, 因而采用第二种方法。

混淆矩阵是用于描述单一分类器 k 对训练样本的识别性能的, 表示如下:

$$PT_k = \begin{bmatrix} n_{11}^{(k)} & n_{12}^{(k)} & \dots & n_{1M}^{(k)} & n_{1(M+1)}^{(k)} \\ n_{21}^{(k)} & n_{22}^{(k)} & \dots & n_{2M}^{(k)} & n_{2(M+1)}^{(k)} \\ \dots & \dots & \dots & \dots & \dots \\ n_{M1}^{(k)} & n_{M2}^{(k)} & \dots & n_{MM}^{(k)} & n_{M(M+1)}^{(k)} \end{bmatrix} \tag{6}$$

元素 $n_{ij}^{(k)}$ 代表某个属于第 i 类的输入样本由分类器 k 判断为第 j 类的数量, 因此可以用下式计算分类器的模糊积分密度:

$$g_k = \frac{n_{jj}^{(k)}}{\sum_{i=1}^M n_{ij}^{(k)}} \quad j < M + 1 \tag{7}$$

通常也用式(7)表示分类器的可信度, 但在文献[8]中提出上面的计算方法仅适合于每类样本数相同的情况, 当每类的

样本数差别较大时, 将引起不精确的可信度。因而在这里利用文献中的思想, 提出了一种改进的模糊积分密度的计算方法:

$$g_k = \frac{n_{jj}^{(k)}}{\sum_{p=1}^M (n_{pj}^{(k)} \xi_p^{(k)})} \quad j < M + 1, \text{ 其中 } \xi_p^{(k)} = \frac{\sum_{p=1}^M n_{jp}^{(k)}}{\sum_{l=1}^M n_{pl}^{(k)}} \tag{8}$$

3 基于模糊积分的多分类器组合的入侵检测

本文采用的多分类器组合模型如图 3 所示。它从网络上截获数据包, 通过数据预处理模块对数据包进行处理, 同时根据数据包头文件中的协议类型进行分类; 然后对每一类协议的数据进行特征提取, 选择出重要的特征, 并对特征空间进行划分; 根据不同的特征子集构造分类器^[4]; 最后通过模糊积分组合策略将多个分类器的输出结果融合到一起。



图 3 入侵检测的多分类器组合模型

根据前面的分析知道, 图 3 中的各个分类器是基于不同的特征空间, 因而对相同的样本空间, 不同的分类器面对的不同样本的数量可能不同, 即由于删除部分属性, 使样本数量减少。为了更准确地描述各个分类器的分类性能, 同时考虑训练样本数量对分类器的影响, 在本文中给出如下的计算模糊积分密度 g 的方法:

$$g_k = \frac{n_{jj}^{(k)}}{\sum_{p=1}^M (n_{pj}^{(k)} \xi_p^{(k)})} \cdot \frac{m_k}{m} \quad j < M + 1 \tag{9}$$

其中, m 是样本总量, m_k 是第 k 个分类器面对的不同样本总量。

4 实验与分析

本文所使用的实验数据是 DARPA(Defense Advanced Research Projects Agency) 为 1999 年 KDD (Knowledge Discovery and Data Mining) 竞赛所建立的基本数据, 选取了其十分之一子数据集中的部分数据。经过对数据的连续化和标准化处理以后, 对 TCP, UDP 和 ICMP 三种协议的数据按照 Basic Features, Content Features 和 Traffic Features 分别建立了三个分类器, 具体过程参见文献[4]。

表 1 给出了每个分类器的分类结果(分类正确率), 其中括号中的数据表示训练样本/测试样本的个数。表 2 给出了采用不同多分类器组合策略的实验结果。

表 1 单个分类器的分类正确率

分类器	TCP(24827 /30738)	UDP(20354 /26703)	ICMP(7548 /3791)
Basic Features	0.918 8	0.869 2	0.981 4
Content Features	0.929 9	—	—
Traffic Features	0.979 5	0.902 8	0.987 8
All Selected Features	0.895 1	0.902 7	0.979 2

表 2 多分类器组合的分类准确率

组合策略	TCP	UDP	ICMP
投票法	0.933 5	0.830 8	0.976 4
神经网络融合	0.952 3	0.881 1	0.981 4
模糊积分	0.973 5	0.902 8	0.987 8

5 小结

本文提出了一种基于模糊积分的多分类器(下转第 121 页)

$$H|x^-\rightarrow\frac{1}{2}(|x^+-|x^-|)$$

传输完毕, Alice 和 Bob 交错通过经典信道告诉对方自己对应于 QKA 的粒子的测量值。Alice 认证 Bob, 只需看 Bob 的测量值是否与自己的负相关, Bob 认证与 Alice 类同。如果错误比特率小于一定的阈值, 则认证通过。对于用于 QKD 的量子态, Bob 告诉 Alice 他测量选用的基, Alice 告诉 Bob 哪些基对应的位是无用的, 因为如果选的是 $|\Phi^-\rangle$, 只有两者的基不同时测量结果才有相关性; 如果选的是 $|\Psi^+\rangle$, 则只有两者选择相同的基, 测量结果才会有相关性(都选 Z 基时负相关, 都选 X 基时正相关)。在有相关性的位, Bob 保持测量值作为分发的密钥, Alice 根据相关性将自己的结果与 Bob 的相统一, 同时两人通过纠错和保密加强获得的新密钥。在其中提取新的认证密钥 K , 为以后认证作准备, 同时将旧的认证密钥丢弃。

这里不需要再检测窃听, 认证过程完全可以确定有无窃听, 简化了协议。在轻微降低安全性的条件下, 发送者和接收者对用于 QKD 的粒子可以用单一基来测量。协议在量子信道中传输的是非正交量子态, 由测不准原理, Eve 不可能成功窃听而不被察觉。再者, 他根本无法知道究竟哪些态是用于认证的, 哪些是用于分发密钥的, 也不知道在认证的位置究竟该用什么基来测量。总之, 本协议具有无条件安全性。

4 结论

本文假设共享一初始密钥, 提出一种通用的思想, 将通信者共享的初始密钥转换为较长的认证密钥 K' , 并对 K 进行一定的处理后用于身份认证。以此为基础, 所有量子密钥分发协议都可同时进行身份认证。在利用 K 作认证准备时, 与文献[5] 中的方法有些类似, 但直接用每个 s_i 中的位决定所选基或变换或发送值, 大大节约了 K 中的位, 提高了其中比特的利用率, 较之更简单; 加之文献[5] 中的长串直接借助 Trent 并通过量子信道获得, 这不可避免地降低了安全性; 另外文献[5] 中的长串在分发认证完成后不能更新, 本文提出的通用算法可以解决这些问题。在此基础上提出的认证协议, 有传统的测不准原理和量子态不可克隆原理保证, 窃听者不知道哪些位是用来

进行认证, 哪些是用来分发密钥的, 增加了协议的安全性, 并且认证过程直接取代了检测窃听过程, 简化了密钥分发协议。总之, 协议比传统的 BB84, B92, EPR 等协议更安全、更简单。

参考文献:

[1] Brassard, *et al.* An Update on Quantum Cryptography[C]. Advances in Cryptology, Proceedings of Crypto84.

[2] Markus Michler, *et al.* Interferometric Bell-state Analysis[J]. Phys. Rev. A, 1996, 53: 1209.

[3] Charles H Bennett. Quantum Cryptography Using any Two Nonorthogonal States[J]. Phys. Rev. Lett, 1992, 68: 3121.

[4] Artur K Ekert. Quantum Cryptography Based on Bell 's Theorem [C]. Phys. Rev. Lett, 1991, 67: 661.

[5] Daniel Liunggren, *et al.* Authority-based User Authentication in Quantum Key Distribution[J]. Phys. Rev. A, 2000, 62: 22305.

[6] Miloslav Dusek, *et al.* Quantum Identification System[J]. Phys. Rev. A, 1999, 60: 149.

[7] Takashi Mihara. Quantum Identification Schemes with Entanglements [J]. Phys. Rev. A, 2002, 65: 52326.

[8] Bao-Sen Shi, *et al.* Quantum Key Distribution and Quantum Authentication Based on Entangled State[J]. Physics Letters A, 2001, 281.

[9] Claude Crepeau, *et al.* Committed Oblivious Transfer and Private Multi-party Computation[C]. Advances in Cryptology, Proceedings of Eurocrypt '95.

[10] Yong-Sheng Zhang, *et al.* Quantum Authentication Protocol [C]. arXiv: quant-ph/0001046.

[11] Guihua Zeng, *et al.* Identity Verification in Quantum Key Distribution [J]. Phys. Rev. A, 2000, 61: 22303.

[12] Howad Barnum, *et al.* Quantum Secure Identification Using Entanglement and Catalysis[C]. arXiv: quant-ph/9910072.

[13] Daniel Jonathan, *et al.* Entanglement-Assisted Local Manipulation of Pure Quantum States[J]. Phys. Rev. Lett, 1999, 83: 3566.

[14] P G Kwiat, *et al.* New High-Intensity Source of Polarization-Entangled Photon Pairs[J]. Phys. Rev. Lett, 1995, 75: 4337.

作者简介:

郭奋卓(1977-), 女, 博士研究生, 研究方向为信息安全、量子密码; 温巧燕(1959-), 女, 教授, 博士生导师, 博士(后), 研究方向为信息安全、密码学。

(上接第 118 页) 组合的入侵检测方法, 将特征空间进行划分, 为每一个特征子空间建立一个分类器, 然后将多个分类器进行模糊积分组合; 同时, 根据入侵检测系统中分类器的构造特点提出了一种模糊积分密度的计算方法。实验结果显示该方法执行效果较好, 但根据模糊积分的计算过程可以看出, 该方法比投票法的计算复杂性高。

参考文献:

[1] arun Ambwani. Multi Class Support Vector Machine Implementation to Intrusion Detection[C]. International Joint Conference on Neural Network, 2003. 2300-2305.

[2] Yu Guan, Ali A Ghorbani, Nabil Belacel. Y- means: a Clustering Method for Intrusion Detection[C]. Canadian Conference on Electrical and Computer Engineering, 2003. 1083-1086.

[3] S Mukkamala, G Janoski, Andrew Sung. Intrusion Detection Using Neural Networks and Support Vector Machines[C]. International Joint Conference on Neural Network, 2002. 1702-1707.

[4] 张莉, 孙钢, 郭军. 基于特征分析的多分类器融合的网络入侵检测

[J]. 计算机工程与应用, 2004, 40(18) : 13-14, 22.

[5] Sung-Bae Cho, Jin H Kim. Multiple Network Fusion Using Fuzzy Logic[J]. IEEE Transactions on Neural Network, 1995, 6(2) : 497-501.

[6] Sung-Bae Cho, Jin H Kim. Combining Multiple Neural Networks by Fuzzy Integral for Robust Classification[J]. IEEE Transactions on Systems Man and Cybernetics, 1995, 25(2) : 380-384.

[7] Tuan D Pham. Combination of Multiple Classifiers Using Adaptive Fuzzy Integral[C]. Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems(ICAIS '02), 2002. 50-55.

[8] L Chen, H L Tang. Improved Computation of Beliefs Based on Confusion Matrix for Combining Multiple Classifiers[J]. Electronics Letters, 2004, 40(4) : 238-239.

作者简介:

孙钢(1972-), 男, 博士研究生, 研究方向为计算机网络安全、网络管理; 张莉(1972-), 女, 博士研究生, 研究方向为计算机网络安全、网络管理; 郭军(1959-), 男, 院长, 教授, 博士生导师, 研究方向为智能信息处理、网络管理与控制。