基于 CCA 的入侵检测行为轮廓创建技术研究

郭 陟[1], 赵曦滨[2,1], 顾 明[1]
(1. 清华大学软件学院, 北京 100084; 2. 江苏大学计算机科学与通信工程学院, 江苏镇江 212013)

摘 要: 入侵检测系统在保障 Internet 应用系统安全方面发挥着重要作用. 作为异常检测依据, 用户行为轮廓的准确程度直接关系到入侵检测系统的检测性能. 由于 Internet 环境的开放性造成用户行为模式多变, 导致用户行为轮廓准确程度下降. 本文提出了基于信息可视化的入侵检测框架, 并进一步提出了基于 CCA(Curvilinear component analysis)的可视化算法. 该可视化算法比传统算法具有更好的距离映射性能, 可为安全专家提供准确的可视信息, 有利于安全专家直观地观察用户行为模式, 并合理选择聚类算法创建轮廓, 从而提高行为轮廓创建的准确性.

关键词: 入侵检测; 异常检测; 行为轮廓创建; 信息可视化; CCA
中图分类号: TP393 文献标识码: A 文章编号: 0372-2112 (2004) 08-1381-04

Profile Creation Technique Research for Intrsuion Detection Based CCA

GUO Zhi[1], ZHAO Xi-bin[1,2], GU Ming[1]
(1. School of Software, Tsinghua University, Beijing 100084, China;
2. School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China)

Abstract: Intrusion detection systems take an important role in securing Internet applications. The exactness of user behavior profiles directly affects the detection performance of intrusion detection systems because profiles are the criterion of anomaly detection. The exactness of profiles would be reduced with the use of traditional profile creation methods due to uncertainty of user behavior patterns in Internet. We proposes a new intrusion detection scheme based on information visualization, and presents a new CCA(Curvilinear Component Analysis)-based visualization algorithm. This algorithm is better than traditional algorithm in the performance of distance mapping, and can provide more exact visual information for security experts. Visual information of user behavior patterns facilitates security experts to select more suitable cluster analysis algorithms to create more exact behavior profiles.

Key words: intrusion detection; anomaly detection; behavior profile creation; information visualization; CCA

1 引言

随着 Internet 的普及, 黑客和恶意用户对 Internet 应用系统的威胁日益严重. 近年来, 入侵检测(Intrusion detection)技术已成为保障 Internet 应用系统安全的重要手段. 入侵检测方法可分为异常检测(Anomaly detection)和误用检测(Misuse detection)两大类[1]. 与误用检测不同, 异常检测[2]具有检测未知攻击的能力, 因此异常检测方法在入侵检测系统(Intrusion detection system, IDS)中发挥着重要的作用.

基于异常检测的入侵检测系统由日志采集、轮廓(profile)创建、入侵检测等模块构成. 轮廓创建模块从日志采集模块获得历史日志记录, 然后从日志中提取用户正常行为模式, 并创建用户正常行为轮廓. 入侵检测模块比较用户当前活动与正常行为轮廓间的差异. 如果当前活动与正常行为轮廓不一致, 则怀疑当前活动异常, 入侵检测模块将发出入侵警报.

因为正常行为轮廓是入侵检测的基准, 所以行为轮廓的准确程度直接影响到入侵检测性能. 轮廓创建模块主要由数据过滤、数据精简、轮廓定义等步骤组成. 鉴于 Internet 应用系统处于开放环境中, 难以获取完全没有入侵活动的/干净0数据, 因此建立正常行为轮廓前必须利用数据过滤去除内含入侵活动的数据. 在数据过滤过程中, 如果没能去除入侵日志数据, 则会影响到行为轮廓的正确性. 如将正常数据与入侵数据一并除去, 则行为轮廓无法全面刻画用户正常行为. 因此, 创建轮廓时数据过滤的准确性非常重要.

由于目前 Internet 应用系统用户数目众多, 且需由入侵检测系统保护的资源也非常多, 因此轮廓创建模块往往需处理海量数据. 出于效率考虑, 必须用特征提取、特征选择、维度精简等方法精简用户行为轮廓. 在精简轮廓过程中, 如忽略信息过多, 将严重降低行为轮廓准确程度. 正常行为轮廓准确程度降低, 将导致入侵误报率(False positive)升高, 异常检测性能变差. 如忽略信息过少, 又无法达到精简行为轮廓之目的. 因此, 是否能准确而有效地精简轮廓成为了评价轮廓创建模块性能的重要标准.

目前, 聚类(cluster)分析技术已被应用到入侵检测行为轮廓创建过程中, 用以过滤异常数据和精简行为轮廓[3,4]. 但由于 Internet 应用系统用户活动受约束较少, 行为较自由, 导致

收稿日期: 2003-12-29; 修回日期: 2004-05-15
基金项目: 国家 863 计划(No. 2001AA414220, No. 2001AA414020, No. 2002AA412010)

用户行为模式多变. 在聚类分析中, 因为点集构造由用户行为决定, 所以用户行为模式多变表现为点集数据构造多变. 由于不同聚类方法(例如, 最近距离法、最远距离法、均值距离法、K2均值法等) 对不同构造点集的聚类结果不同[5]. 因此, 在创建 Internet 应用系统用户行为轮廓时, 如采用文献[3, 4]提出方法创建行为轮廓, 轮廓准确性将受用户行为影响, 导致在部分 Internet 应用系统中轮廓准确程度非常差.

如果能自适应地根据点集结构采用合适的聚类方法, 将有助于提高行为轮廓的准确程度. 但由于 Internet 应用系统用户行为多变, 导致了日志数据点集的结构多变, 为计算机自动辨识点集结构造成了巨大的困难, 难以针对性地采用合适的聚类方法. 考虑到人类对点集拓扑结构具有很强的认知能力, 因此在创建轮廓过程中可利用人类认知能力来辨识点集结构, 并针对特定点集结构采用合适的聚类方法来过滤数据与精简轮廓, 以达到准确创建 Internet 应用系统轮廓之目的.

实践证明, 可视化控制台可大幅提高人机交互效率. 因此, 利用可视化算法, 可为安全专家提供直观而可视的点集结构信息, 从而辅助安全专家更加合理地选择聚类方法来创建轮廓, 最终达到提高轮廓创建准确性之目的. 在此过程中, 可视化算法输出的点集结构信息是否准确直接影响到人的辨识能力.

为指导信息可视化在入侵检测行为轮廓创建过程中的应用, 本文提出了一种新的基于信息可视化的入侵检测框架. 为支持该框架的实现, 并针对已有可视化算法[6]输出点集结构信息不准确的缺陷, 本文基于自适应神经网络技术, 提出了基于 CCA 的可视化算法, 用以为安全专家提供更加准确的可视信息.

2 基于 CCA 的轮廓创建技术

211 基于信息可视化的入侵检测框架

基于信息可视化的入侵检测框架如图 1 所示.

图 1 基于信息可视化的入侵检测框架

安全专家根据可视化的历史日志信息, 观察日志数据点集的拓扑结构, 分析用户历史活动的内在规律, 针对特定区域范围内的点集选择合适的聚类分析方法, 完成数据过滤与轮廓精简的任务, 从而建立用户正常行为轮廓. 在建立完行为轮廓后, 还可将行为轮廓可视化, 安全专家通过观察已建立的行为轮廓与用户历史活动间的关系, 来进一步调整行为轮廓. 随着安全专家不断积累经验, 辨识点集结构的能力也将不断提高, 这将有助于更准确地创建正常行为轮廓.

上述框架的可视化模块由可视化算法实现, 负责为安全专家提供可视化信息, 从而使安全专家与入侵检测系统间能高效交互. 因此, 可视化算法的性能决定了本文提出框架的有效性.

212 已有可视化算法的研究

根据入侵检测假设[2], 正常用户行为往往有一致的行为模式, 体现为空间中点的聚类; 而异常用户活动往往背离正常主体行为模式, 即标识异常用户活动的点在空间中会远离表示正常行为模式的聚类. 因此, 空间的点集结构可表示用户行为内在规律. 安全专家可通过观察空间点集拓扑结构, 从隐喻信息中感知用户行为特征, 并确定进行聚类分析的区域范围与相应的聚类方法, 来指导轮廓创建过程.

由于 Internet 应用程序需要监控的资源或命令一般非常多(远大于 3 种), 因此历史日志数据事实上是在高维空间中的多元数据. 由于人眼无法看到高维空间中的信息, 因此在可视化算法中需要将高维数据利用映射工具映射到 3 维以下空间.

已有的将可视化技术应用在入侵检测领域的研究采用了 PCA(Principal component analysis) 作为映射工具[6, 7]. 在本文中, 由于需要在映射过程中保留点之间的聚类信息, 因此需要映射工具保留点间距离信息的性能较好, 即具有较好的距离映射性能. PCA 在映射过程中会将与次要成分平行的两点间距离信息忽略, 可能导致大量损失距离信息, 甚至造成两点重叠, 因此 PCA 不适用于需保留距离的映射. 文献[6]提出的可视化算法无法满足本文框架对算法的要求.

现有的映射工具, 例如 CCA, MDS(Multidimensionality ccal2ing), NLM(Sammon s nonlinear mapping), 能较好地在映射过程中保留距离. 如文献[8]所述, MDS 保留距离的能力比 CCA 差; NLM 保留距离能力接近 CCA, 但 NLM 的时间复杂度为 $O(N^2)$, 而 CCA 的时间复杂度为 $O(N)$. 因此, CCA 作为一种基于自适应神经网络的数据映射工具, 能在映射过程中尽可能地保留原始空间中点的结构信息, 以便在映射空间中能可视化地识别点集结构. 对比不同映射工具后, 本文采用 CCA 作为映射工具, 提出了基于 CCA 的可视化算法.

213 基于 CCA 的可视化算法

基于 CCA 的可视化算法将文本日志信息转化为可视信息传递给安全专家. Internet 应用系统多建立在 Web 架构之上, 因此 Web 服务器日志是 Internet 应用系统日志数据的重要组成部分. Web 服务器访问日志的标准包括 CLF(Common log format) 和 ELF(Extended log format) 等. 在本文中, 基于 CCA 的可视化算法以 CLF 日志数据为输入, 从中提取出如下格式的审计事件:

Event = (Subject, Object, Source, EventType, TimeStamp)

其中,

ó Subject: 代表用户主体标识. 本算法中该标识取远程用户 IP 地址.

ó Oject: 代表被监控客体标识. 本算法中该标识或者取被访问的资源名, 或者取被执行的命令名, 或者取一组资源的组名, 或者取一组被执行的命令的组名.

ó Source: 表示日志数据的来源, 取目标系统所在服务器

的 IP 地址.

Ó EventType:表示审计事件类型,有访问、执行等类型.

Ó TimeStamp:审计时间发生的时间戳. 本算法中采用访问资源或执行命令的时间.

在审计事件基础上,可针对每个用户主体建立行为模型[7],记为矩阵 M:

$$M=\begin{array}{c} \\ 1 \\ s \\ \text{用户会话 } i \\ s \\ n \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & , & j & , & P \end{array} \\ \begin{bmatrix} & & & & \\ & & s & & \\ & , & x_{ij} & & \\ & & & & \end{bmatrix} \end{array}$$

在矩阵 M 中,每个 p 维行向量代表一次用户会话(User session),每个 n 维列向量代表一种入侵检测测度(measure).

设 X_{ij} 代表原始空间中归一化行向量 x_i 与 x_j 之间的距离,Y_{ij} 代表映射空间中可视点 y_i 与 y_j 之间的距离. 归一化行向量记为

$$x_i=[x_{i1}, \cdots, x_{ip}],\ 1 \le i \le n \tag{1}$$

其中,

$$x_{ij}=(x_{ij}-\bar{x}_i)/(R_i\sqrt{p}),\quad 1 \le i \le n,\ 1 \le j \le p \tag{2}$$

$$\bar{x}_i=\left(\sum_{j=1}^{p} x_{ij}\right)/p,\quad 1 \le i \le n \tag{3}$$

$$R_i=\sqrt{\left[\sum_{j=1}^{p}(x_{ij}-\bar{x}_i)^2\right]/p},\quad 1 \le i \le n \tag{4}$$

如文献[8]所述,算法通过优化迭代力图使如式(5)所示二次代价函数最小化.

$$E=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(X_{ij}-Y_{ij})^2 F(Y_{ij}, K(t))] \tag{5}$$

在本文中,取

$$F(Y_{ij}, K(t))=e^{-Y_{ij}/K(t)} \tag{6}$$

其中,$K(t)$ 是邻接因子,t 代表时间,$K(t)$ 需选择一个随时间单调下降的函数.在映射过程中,可视化算法执行如下步骤使代价函数最小化:

(1) 对矩阵 M 按行向量进行归一化;

(2) 在映射空间选择初始映射点;

(3) 计算归一化行向量 x_i 相互间的欧氏距离 X_{ij};

(4) 随机选择 1 到 n 间的整数值作为 i,按照如下步骤对除第 i 点外所有点(第 j 点)执行迭代过程若干次:

Ó 根据当前时间计算 $A(t)$,$K(t)$,其中 $A(t)$ 为学习因子,为随时间单调下降的函数;

Ó 计算映射空间中第 i 点与其他点之间距离 Y_{ij};

Ó 依据文献[8]提出的公式(7)计算映射空间中第 j 点的本次偏移量;

$$\Delta y_j=A(t)F(Y_{ij}, K(t))(X_{ij}-Y_{ij})\frac{y_j-y_i}{Y_{ij}},\quad j \ne i \tag{7}$$

Ó 根据计算的偏移量在映射空间中移动映射点.

(5) 如有必要,按式(5)所示代价函数评价映射效果.

上述算法的输出结果为二维/三维散乱点场,散乱点间的距离表示空间点间的距离关系,可隐喻表示空间点结构.在获得点数据场后,将可视点标以不同的颜色、纹理、图例,用于表示如下信息:用户主体标识、用户行为轮廓、用户会话标识、用户会话顺序、日志数据来源.

安全专家观察上述可视化信息,可感知空间点集结构,选择合适的聚类方法用于区分聚类.并在聚类分析的基础上,完成数据过滤与轮廓精简工作,从而保障行为轮廓准确程度.

3　实验结果

基于上述入侵检测框架与可视化算法,笔者实现了利用可视化技术创建轮廓的原型系统,并使用某大学图书馆门户网站日志数据进行了实验.实验使用了两周的日志数据,从中选择了 39 个用户的数据,以天为单位划分会话,得到 476 个会话.利用文献[6,7]中提到的基于 PCA 的算法可视化所有会话,获得如图 2 所示的二维散乱点数据场.

利用基于 CCA 的算法进行可视化后,获得如图 3 所示的二维散乱点数据场.

在实际系统中,安全专家通过观察图 3 可发现,图中存在着标识用户异常会话的点.在提取轮廓前,需将散布在图中的异常点移除.利用数据过滤移除明显异常会话后,即可根据点结构选择合适的聚类方法,进行聚类分析.依据分析结果,还可进一步过滤数据与精简轮廓.这样,在可视化交互过程中,安全专家依据直观的可视化信息,结合自己知识,不断调整聚类分析方法与范围,最终达到准确创建用户行为轮廓之目的.

通过对比图 2 与图 3 的实验结果,可发现图 2 中难以观察到明显的聚类结果,而图 3 中可发现明显的聚类聚集,并且可直观地观察到聚类拓扑结构.为评估上述两种算法的距离映射效果,笔者利用公式(8)计算原始空间中第 i 点与第 j 点间距离 X_{ij} 与映射空间中相应两点间距离 Y_{ij} 之间的差值.

$$D_{ij}=|X_{ij}-Y_{ij}|,\quad 1 \le i, j \le n \text{ 且 } i<j \tag{8}$$

由 476 个会话得到 113,050 个距离差值,以 D_{ij} 为横轴,以

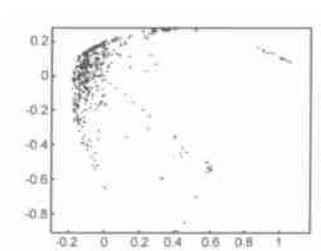

图 2 基于 PCA 的可视化算法实验结果

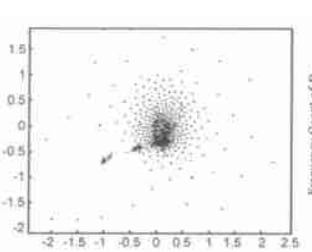

图 3 基于 CCA 的可视化算法实验结果

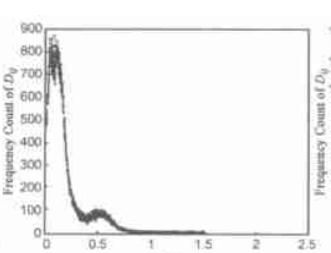

图 4 基于 PCA 的可视化算法的距离偏差分布曲线

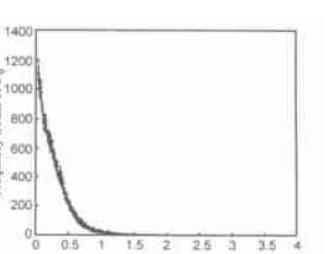

图 5 基于 CCA 的可视化算法的距离偏差分布曲线

D_{ij}的频率计数为纵轴, 可得距离偏差分布曲线. 基于 PCA 的可视化算法的距离偏差分布曲线如图 4 所示, 基于 CCA 的可视化算法的距离偏差分布曲线如图 5 所示.

观察映射前后距离偏差可以发现基于 CCA 的算法大多数的 D_{ij}值较小, 少数的 D_{ij}值较大, 即算法在映射过程中可保留多数点的距离; 但基于 PCA 的算法则没有这一规律. 如果采用不同的数据点集重复上述实验(因篇幅所限, 不再罗列)可发现: 基于 CCA 的算法距离偏差分布形状较稳定, 点集结构对分布形状影响不大; 而基于 PCA 的算法受点集结构影响较大, 不同的点集会导致不同的分布形状, 即算法的距离映射性能不稳定.

上述实验结果证实了基于 CCA 的可视化算法具有更好更稳定的距离映射性能.

4 结论

鉴于现有基于聚类分析的轮廓创建技术存在的问题, 本文提出了基于信息可视化的入侵检测框架. 在该框架中, 安全专家基于直观的信息与自身知识来动态选择聚类分析方法与范围, 并根据反馈的行为轮廓信息来调整聚类划分结果, 最终形成较准确的用户行为轮廓. 本文提出的框架为提高用户行为轮廓准确程度提供了新的途径. 为实现该框架, 本文提出了基于 CCA 的可视化算法, 用以为安全专家提供日志等可视化信息, 从而提高人机交互效率, 为安全专家创建轮廓提供支持. 基于 CCA 的算法比入侵检测领域的传统 PCA 算法具有更好的距离映射性能, 可为安全专家提供直观的点集结构信息, 适用于利用可视信息创建用户行为轮廓的过程.

参考文献:

[1] Bace R G. Intrusion Detection[M]. Macmillan Technical Publishing Co, 2000.

[2] Denning D E. An intrusion detection model[J]. IEEE Transactions on Software Engineering, 1987, 13: 222- 232.

[3] Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering[A]. Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA'2001)[C]. Philadelphia, PA, 2001.

[4] Marin J, Ragsdale D, Sirdu J. A hybrid approach to the profile creation and intrusion detection[A]. Proceedings of 2nd DARPA Information Survivability Conference and Exposition (DISCEX'II 2001)[C]. Anaheim, CA, 2001. 1. 69- 76.

[5] 边肇祺, 等. 模式识别[M]. 北京: 清华大学出版社, 1999.

[6] Lam K Y, Hui L, Chung S L. Multivariate data analysis software for enhancing system security[J]. J. Systems Software, 1995, 31: 267- 275.

[7] Guo Z, Lam K Y, Chung S L, Gu M, Sun J G. Efficient presentation of multivariate audit data for intrusion detection of web-based internet services[A]. Proceedings of the First International Conference on Applied Cryptography and Network Security [C]. Kuming, China, 2003. 64-76.

[8] Demartines P, Hérault J. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data set[J]. IEEE Transactions on Neural Network, 1997, 8(1): 148- 154.

作者简介:

郭 陟 男, 1974 年出生于浙江湖州, 博士研究生, 研究方向为计算机系统安全与软件体系结构.

赵曦滨 男, 1973 年出生于新疆石河子, 博士研究生, 研究方向为计算机系统安全与软件体系结构.

顾 明 女, 1962 年出生于辽宁沈阳, 清华大学软件学院副教授, 研究方向为操作系统、分布式应用系统支撑平台和电子商务等.