基于数据挖掘的分布式网络入侵检测系统设计及实现

傅德胜 周 舒 郭 萍
(南京信息工程大学计算机与软件学院 南京 210044)

摘 要 提出基于数据挖掘的入侵检测系统模型、改进的FP-Growth的关联分析算法和基于分箱统计的FCM网络入侵检测技术。系统实验结果表明,所开发的网络入侵检测系统可以稳定地工作在以太网络环境下,能够及时发现入侵行为,有效地解决了数据挖掘速度问题,增强了入侵检测系统的检测能力,具备了良好的网络入侵检测性能。

关键词 入侵检测系统,分布式,数据挖掘
中图法分类号 TP393.08 **文献标识码** A

Design and Implementation of Distributed Network Intrusion Detection System Based on Data Mining

FU De-sheng ZHOU Shu GUO Ping
(College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract Data mining is applied to intrusion detection system, which puts forward a system model based on data mining, improving the FP-Growth algorithm based on associative analysis, and refining the technology of FCM network intrusion detection based on statistical binning. The experimental result shows that the network intrusion detection developed by this paper can work very stably under the Ethernet, find intrusion activities in time, solve the problem of data mining speed effectively, enhance the detective ability of intrusion detection, and possess a favorable performance of intrusion detection.

Keywords Intrusion detection system, Distribution, Data mining

1 引言

目前推出的商用分布式入侵检测系统一般采用基于已知入侵行为规则的匹配技术,检测引擎分布在需要监控的网络中或主机上,独立进行入侵检测,入侵检测系统中心管理控制平台仅负责平台配置、检测引擎管理和各检测引擎的检测结果显示,对各检测引擎的检测数据缺乏协同分析。同时网络入侵检测系统与防火墙、防病毒软件等之间也是单兵作战,对复杂的攻击行为难以做出正确的判断。因此,研究开发基于数据挖掘的分布式网络入侵协同检测系统具有重要的理论意义和应用价值。

国内外在将数据挖掘用于入侵检测方面已取得了良好的进展,提出了一些基于数据挖掘的入侵检测模型。但是这些模型一般都是基于某种方法的研究与应用。实际上,不同的数据挖掘技术适用于不同的入侵检测,如运用关联分析方法可提取出黑客入侵行为之间的关联特征;运用序列模式分析方法能找出黑客入侵行为的序列关系;数据分类多用于辅助入侵检测中的其它数据挖掘方法,进行预处理或后续处理等。此外,各数据挖掘方法有其自身的局限性。联合使用几种数据挖掘方法要比单独使用一种挖掘方法效果好,例如,将聚类分析用于关联分析的预处理以及规则产生之后的分类,将显著提高关联分析的效果;将粗糙集理论以及基于仿生物技术的遗传算法、免疫算法同其它数据挖掘技术融合起来,将会改善入侵检测的普适性、实时性、准确性。

本文提出基于数据挖掘的分布式网络入侵检测系统结构模型,对算法进行讨论,分析了相关实验结果。

2 系统结构

基于数据挖掘的分布式入侵检测系统模型如图1所示。本地层包括数据采集解析器和数据挖掘检测器,网络层包括报警优化器,系统层包括日志记录器和中心控制平台。数据采集解析器和数据挖掘检测器分布于各个局域网的关键节点上,负责监视所在网段的网络数据,对入侵行为做出响应并发送报警信息至报警优化器。报警优化器位于广域网上,负责收集各局域网内检测引擎发出的警告信息,并将其存储至日志记录器。中心控制平台位于系统层,给管理员提供可视的友好的控制界面。

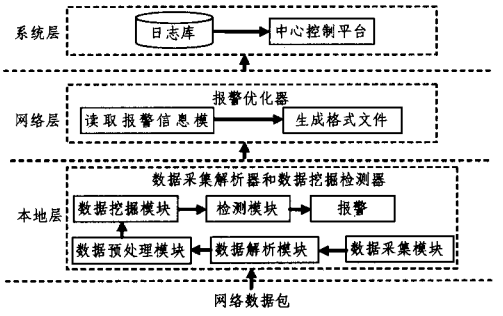

图1 系统总体结构模型图

到稿日期:2008-06-29 本文受江苏省产业技术与开发基金,苏发改[2006]1106号资助。
傅德胜(1950—),男,教授,研究方向为信息安全、模式识别;周 舒(1984—),女,硕士研究生,研究方向为信息安全。

3 改进的 FP-Growth 算法

本文系统数据审计采用 FP-Growth(frequent-pattern growth)改进算法。

FP-Growth 算法由 Jiawei Han,Jian Pei 和 Yiwen Yin 提出,是一种不产生候选项集而采用模式增长的方式挖掘频繁模式的算法:将提供频繁项集的数据库压缩到一颗频繁模式树,但仍保留项集关联信息,然后将这种压缩后的数据库分成一组条件数据库,每个关联一个频繁项,并分别挖掘每个数据库。

FP-Growth 算法在挖掘频繁模式时,它需要递归地生成条件 FP-tree,每产生一个频繁模式就要生成一个条件 FP-tree。在最小支持度较小时,即使对于很小的数据库,也将产生数以十万计的频繁模式,动态地生成和释放数以十万计的条件 FP-tree,将耗费大量系统时间和空间。此外,FP-tree 和条件 FP-tree 需要自顶向下生成,而频繁模式的挖掘需要自底向上处理。由于 FP-Growth 算法时空效率仍然不够高,因此本文对 FP-Growth 算法进行改进,引入聚合链的单链表结构。改进后的 FP 树是单向的,每个结点只保留指向父结点的指针,节省了树空间;相同项的不同节点的路径信息压缩进聚合链中,避免了生成节点链和条件模式库,明显提高了挖掘效率。

实验表明,在相同最小支持度的情况下,改进算法的执行效率较 FP-Growth 算法明显提高。这是由于随着最小支持度的减小,产生的节点链和条件模式库的个数飞速增长,FP-Growth 算法花费在这方面的时间就会很多。FP-Growth 改进算法具有较高的效率,非常适用于实时网络入侵检测系统。

4 基于分箱统计的 FCM 网络入侵检测技术

传统的 FCM 算法是一种无指导无监督性的划分聚类的方法,很容易陷入局部极值点或鞍点而得不到最优解甚至满意解。同时,在处理网络连接数据记录这样的大数据量时,需要频繁更新聚类中心,算法耗时。本文提出基于分箱统计的 FCM 算法,根据带标识的已知聚类的划分,进行新数据记录的类型判断,不仅可以避免陷入局部极值点或鞍点,而且根据分箱判断是否更新聚类中心,从而解决原来需要频繁更新聚类中心的问题,相对提高处理数据的速度。

传统 FCM 算法中使用隶属函数 μ_{ki} 表示数据记录 x_k 与聚类子集 $X_i(1\leqslant i\leqslant c)$ 的隶属关系,使用最大隶属原则来判断数据记录的归属。这里使用如下的模糊近似度方法进行判断。

$$T_i=\left|1-\frac{d_i}{D_i}\right| \tag{1}$$

对已知的各聚类计算 T_i,若 $\min\{T_i\}$ 在对应聚类的门限范围内,将 $\min\{T_i\}$ 所对应的聚类作为新数据记录所属类;若 $\min\{T_i\}$ 不在对应聚类的门限范围内,则作为新类处理。

处理中当检测出某新的网络连接数据记录属于哪种聚类时,不再像以往的聚类方法立即进行聚类中心的更新,而是将该数据记录到所属聚类的聚类中心的距离 d_i,与该聚类的分箱进行比较,当分箱需要更新时,才对该聚类的聚类中心进行更新。这种方法避免了以往聚类方法中需要频繁更新聚类中心的问题。

对于不同的聚类,其分箱划分时有的是连续的,有的存在断层,且断层的幅度有大有小。对于连续的分箱段,仅在分箱的最末处存在需要更新的情况;对于有断层的分箱段,则在出现断层的分箱的底部或顶部存在需要更新的情况。

当某数据记录到聚类中心的距离 d_{new} 落在该聚类的某分箱中时,分两种情况讨论:

① 当 $d_{new}<\min-scale$ 时,该分箱的底部更新为 $\min_{new}=d_{new}$,使用式(2)更新该分箱的均值。

② 当 $d_{new}>\max+scale$ 时,该分箱的顶部更新为 $\max_{new}=d_{new}$,使用式(2)更新该分箱的均值。

$$avg_{new}=\frac{avg_{old}S_{old}+d_{new}}{T_{old}+1} \tag{2}$$

其中参数 S 为聚内距离分布情况分箱表中的分箱容量,参数 T 为分箱比例刻度表中的该聚类的分箱总容量 totality。

$$\text{分箱容量}\quad S_{new}=S_{old}+1 \tag{3}$$

$$\text{分箱总容量}\quad T_{new}=T_{old}+1 \tag{4}$$

聚类中心由离散型属性向量和连续型属性向量两部分组成。对于离散型属性向量的更新分两种情况:

① 新数据记录的某离散型属性向量 m 的离散值 i 在已有的离散值集 I 中,即 $i\in I$。

这时,对于该离散型属性向量 m 的离散值 i 的原概率统计值 p_i 的更新如式(5)所示。

$$p_i'=\frac{p_iT_{old}+1}{T_{old}+1} \tag{5}$$

该离散型属性向量 m 的其他离散值 j 的原概率统计值 p_j 的更新如式(6)所示。

$$p_j'=\frac{p_jT_{old}}{T_{old}+1} \tag{6}$$

② 新数据记录的某离散型属性向量 m 的离散值 i 不在已有的离散值集 I 中,即 $i\notin I$。对于离散型属性向量 m 的新离散值 i 的概率统计值 p_i 如式(7)所示。

$$p_i=\frac{1}{1+T_{old}} \tag{7}$$

否则,该离散型属性向量 m 的已有的离散值 j 的原概率统计值 p_j 的更新如式(8)所示。

$$p_j'=\frac{p_j}{1+T_{old}} \tag{8}$$

对于连续型属性向量中心的更新如式(9)所示。

$$Y_p'=\frac{Y_pT_{old}+X_p}{1+T_{old}} \tag{9}$$

其中 Y_p 为某聚类中心 C 的连续型属性向量值,$11\leqslant p\leqslant 32$,$X_p$ 为某数据的连续型属性向量值。

如果新的网络连接数据记录不属于已有的任何聚类,则根据该数据记录设置一新的聚类,同时建立相应的分箱。

5 系统实验分析

5.1 实验数据集

数据集选择 KDDCup99[1] 的网络入侵检测数据集。该数据集是在美国国防部高级研究项目局的资助下由美国麻省理工学院林肯实验室按照美国空军局域网结构建造一个实验网,模仿正常的网络使用,有计划进行拒绝服务攻击、远端未经授权的访问、未经授权提升权限、探针 4 类攻击[2],将记录下的流量系统日志和主机文件系统映像等数据,交由参加评估的 IDS 进行离线分析。训练数据集包含了五百万个数据连

接，测试数据集包含了两百万个数据连接，每条数据样本有41个属性，描述了网络连接的基本特征、内容和通信量统计等方面的信息。数据集包括含有标识的训练数据和未加标识的测试数据，共有1种正常的标识类型 normal 和22种训练攻击类型。另外有14种攻击仅出现在测试数据集中。

5.2 实验结果及分析

由5.1节中选取的数据包 kddcup_data_10percent，模拟真实网络环境中入侵行为较少、中等和较多的情况，分别形成3组数据集，每组数据集中的训练集和测试集的详细情况如表1所列。

表1 数据集样本组成情况

组	训练/测试	总数	正常	异常			
				DOS	R2L	U2R	Probing
一	训练	2000	1960	24	10	2	4
	测试	2500	2430	42	18	4	6
二	训练	2500	2310	114	47	10	19
	测试	3000	2770	138	57	12	23
三	训练	3000	2460	324	135	27	54
	测试	3500	2870	378	157	31	64

采用表1中的训练集，按照本文提出的基于 FP-Growth 改进算法的数据挖掘模块分别对3组训练集进行学习训练，将异常模式提取出来，形成规则库，然后用测试集分别进行实验测试，实验结果如表2所列。

表2 基于 FP-Growth 改进算法的数据挖掘实验结果

组	检测率	误报率	漏报率	训练时(秒)
一	96.01%	3.71%	14.28%	10.31
二	94.93%	4.33%	13.91%	12.69
三	93.34%	5.85%	10.31%	15.02

采用未改进的 FP-Growth 算法进行实验的结果如表3所列。

表3 未改进的 FP-Growth 算法实验结果

组	检测率	误报率	漏报率	训练时间(秒)
一	94.88%	4.52%	25.71%	20.25
二	93.73%	5.05%	20.86%	24.94
三	91.74%	6.79%	14.92%	31.16

比较表2和3可以看出：

(1)相对于未改进的 FP-Growth 算法，3组训练时间分别降低 49.09%，49.12% 和 51.80%，由此表明，采用 FP-Growth 改进算法可以大大提高系统的检测效率。

(2)对于每一组数据，改进后的 FP-Growth 算法的检测效率提高 1.19%～1.75%，误报率降低 13.84%～17.52%，漏报率降低 30.90%～44.46%，检测性能有较大改善。

结束语 将数据挖掘技术应用到入侵检测系统是目前入侵检测研究的重要方向，本文设计的基于数据挖掘的分布式网络入侵检测系统采用了改进的 FP-Growth 的关联分析算法和基于分箱统计的 FCM 网络入侵检测技术，有效地解决了数据挖掘速度问题，增强了入侵检测系统的检测能力，为系统管理人员网络维护与保障提供了坚实的基础。

参 考 文 献

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. KDD Cup 1999 Data

[2] 关健. 入侵检测系统数据分析方法及其相关技术的研究. 博士学位论文. 哈尔滨工程大学，2004

[3] 史志才，季振洲，胡铭曾. 分布式网络入侵检测技术研究. 计算机工程，2005，31(13)

[4] Gopalakrishna R, Spafford E H. A Framework for Distributed Intrusion Detection Using Interest Driven Cooperating Agents. Department of Computer Science, Purdue University, May 2001

[5] Fayyad U M, Piatesky-shapiro G, Smyth P. Advances in knowledge discovery and data mining. Galifornia: AAAI/MIT Press, 1996

[6] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets // Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, TX, USA, 2000: 427-438

[7] Han Jia wei, Chee S H S, Chiang J Y. Issues for On-Line Analytical Mining of Data Warehouses

[8] Fuchsberger A. Intrusion Detection Systems and Intrusion Prevention Systems. Information Security Technical Report. 2005, 10: 134-139

[9] Kim G H, Spafford E H. Experiences with tripwire: Using integrity checkers for intrusion detection[R]. West Lafayette, USA: Purdue University, Depatment of Computer Sciences, 1994

[10] Lee W, Stolfo S J, Chan P K, et al. Real time data mining-based intrusion detection[A] // Proceedings of 2nd DARPA Information Survivability Conference and Exposition (DISCEX)

(上接第102页)

[3] Prodio H, Lawrence H. Scalable Clustering: A Distributed Approach // Proc. IEEE Int'l Conf. Fuzzy Systems. Budapest, Hungary. ETATS-UNIS, 2004, 7(1): 143-148

[4] Januzaj E, Kriegel H P, Pfeifle M. DBDC: Density Based Distributed Clustering // Proc. the 9th Int'l Conf. Extending Database Technology. Heraklion, Greece: Springer, 2004: 88-105

[5] Januzaj E, et al. Scalable Density Based Distributed Clustering // Proc. the 8th Eur. Conf. Principles and Practice of Knowledge Discovery in Databases. Paris: Springer, 2004: 231-244

[6] Vaidya J, Clifton C. Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data // Proc. of the 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. Washington, DC, 2003: 206-215

[7] Lin X, Clifton C, Zhu M. Privacy Preserving Clustering with Distributed EM Mixture Modeling. Knowledge and Information Systems, 2005, 8: 68-81

[8] Clifton C, Kantarcioglu M, Vaidya J, et al. Tools for Privacy Preserving Distributed Data Mining. SIGKDD Exploration. 2002, 4(2): 28-34

[9] Zhang Guo-rong, Yin Jian. Preserving clustering over distributed data. Computer Engineering and Applications, 2007, 43(18): 165-167

[10] http://www.ics.uci.edu/~mlearn/databases/iris/

[11] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[12] Modha D S, Spangler W S. Feature weighting in k-means clustering. Machine Learning, 2003, 52(3): 217-237