文章编号: 1006 - 9348(2020) 10 - 0263 - 05

基于数据挖掘的网络入侵数据自主防御仿真

张代华,沈 勇,章翔飞,王 兵
(江苏科技大学计算机学院,江苏 镇江 212003)

摘要: 现有校园网络入侵数据自主防御方法通常需要依赖足够数量样本的分析,但由于网络攻击技术的提升,大量新型攻击技术涌现,初始训练样本数量有限严重影响了校园网络入侵数据的检测性能和自主防御性能,提出了基于数据挖掘的网络入侵数据自主防御方法。方法将采集获得的校园网络数据进行离散连续化、标准化和归一化等预处理; 通过采用数据挖掘方法中的模糊 C 均值聚类随机选取一个聚类中心,迭代目标函数,寻找目标函数的最小值,并不断调整聚类中心和隶属度,获得校样本最佳类别,完成校园网络数据集聚类; 在此基础上通过度量聚类后各个数据集簇的异常度判断是否为入侵数据; 基于校园网络入侵数据检测结果构建了一个三维立体自主防御架构,实现了校园网络入侵数据自主防御。仿真结果表明,所提方法能够克服了当前方法的弊端,实现了校园网络入侵数据的准确检测和完全自主防御。

关键词: 数据挖掘; 网络入侵; 数据; 自主防御
中图分类号: TP309 **文献标识码:** B

Self - Defense Simulation of Network Intrusion Data Based on Data Mining

ZHANG Dai - hua, SHEN Yong, ZHANG Xiang - fei, WANG Bin
(School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang Jiangsu 212003, China)

ABSTRACT: The existing self - defense methods of campus network intrusion data usually rely on a sufficient number of samples. However, the limited number of initial training samples seriously influences the detection performance and autonomous defense of campus network intrusion data. In this article, a method of autonomous defense for network intrusion data based on data mining was proposed. This method performed the pretreatment of discrete continuation, standardization and normalization on the collected campus network data. Our method used the fuzzy C - means clustering in data mining method to randomly select a clustering center. By iterating the objective functions, we could find the minimum value of objective functions. Continuously, we adjusted the clustering center and membership degree to get the best category of samples, and thus completing the campus network dataset clustering. On this basis, we measured the abnormality of each data cluster, so as judge whether it was intrusion data. Based on the result of campus network intrusion data detection, we built 3D autonomous defense architecture. Thus, we realized the autonomous defense of campus network intrusion data. Simulation results show that the proposed method can overcome the shortcomings of current method and achieve the accurate detection and complete autonomous defense of campus network intrusion data.

KEYWORDS: Data mining; Network intrusion; Data; Autonomous defense

1 引言

大学校园网与其它网络形式有较大差异,属于非商业化运营网络,因此在安全防护上比其它外网薄弱很多。随着现代信息技术、互联网技术的发展,以及大数据时代的到来,推动了大学校园的信息化水平,校园网络安全问题逐渐被重视,所谓校园网络安全问题[1]是指校园网络运行过程中的稳定程度以及网络信息的安全状况,这其中涉及了校园网络硬件支持、软件稳定,以及防攻击体系等多方面。

校园网络面临的安全风险主要来自于以下几方面:

1) 来自于校园网络外部的安全风险[2]

基金项目: 2018 年江苏省教育信息化研究课题(20180013)
收稿日期: 2019 - 02 - 19 修回日期: 2019 - 03 - 29

校园网络的内部环境比较纯粹，一旦出现外部威胁就会变得十分脆弱，通常情况下校园网络都会设置一个与外网连接的安全阀，为了方便老师教研和学生的学习，此安全阀在校园网络内部是相对开放的，如果校园网络内部用户存在不良企图，窃取校园网络内部资料并非全无可能，特别是现如今黑客技术水平的发展已经超乎想象，大学校园作为人们活动相对比较密的场所，拥有许多有价值的信息、资源难免会有居心不良之人运用各种技术手段对校园网络发起攻击，破坏校园网络安全；

2) 来自于校园网络硬件设施的安全风险

作为承担整个校园网络安全运行的硬件设施[3]（包括供电设施、服务器、通信光缆等），一旦出现问题整个校园将会陷入断网状态，特别是受恶劣天气影响，网络硬件设施如果没有被妥善保护和存放也会出现各种各样的故障；

3) 当前网络技术水平的局限性[4]

夜晚通常是大学生上网的高峰期，大量的通信信息交互可能会超过网络承载力，当前网络技术水平存在局限性，可能无法应对这一问题；除此之外，校园网络后台管理人员的专业水平也影响着校园网络安全程度；

4) 校园网络管理水平有限

个别院校对校园网络安全问题的认知严重缺乏，在校园网络技术支持和安全管理方面有待提高。

综上分析可知，校园网络安全无时无刻不受到来自内部或外部的影响，实现网络入侵数据自主防御势在必行。针对经典 BP 神经网络在校园网络入侵检测中容易陷入局部最优解，收敛速度较慢，以及学习能力不高的问题，研究人员尝试对采集获得的校园网络数据进行特征提取和权值修正[5]，虽然提升了收敛速度，但引入了大量噪声，对校园网络中已知攻击类型的入侵数据具有较高的检测率，但对于未知的新出现的攻击类型入侵数据辨识度不高，总体检测效果并没有得到有效改善，自主防御性能更无从提及[6]，针对这一现状提出了数据挖掘的网络入侵数据自主防御方法。

2 基于数据挖掘的网络入侵数据自主防御

2.1 校园网络数据预处理

1) 校园网络离散数据连续化处理

已知采集获得的校园网络数据中包括符号型的离散属性和数字型的连续属性，对于前者来说它们的值用于描述数据状态，各数据之间相互独立，需要将其转换成连续型的数值，假设采集获得的校园网络数据中包括 S 个符号型的离散属性数据，分别用 $b_1,b_2,\cdots,b_S$，与之相对应的状态种类为 $t_1,t_2,\cdots,t_S$，如果属性特征 b_i 对应的 t_i 个状态分别为 $Sig_1,Sig_2,\cdots,Sig_{t_i}$，如果校园网络数据记录中的一个数据 b_i 的状态是 Sig_{t_i}，则可得数据 b_i 属性的值为

$$value(b_i) = num(Sig_{t_i}) / \sum_{k=1}^{t_i} num(Sig_{t_k}),k \in t \quad (1)$$

上式(1)代表数据 b_i 属性当前状态 Sig_{t_i} 在采集获得的校园网络数据集中的数量在所有状态数目中的占比，采用上式(1)不仅能够使得校园网络数据集中符号型离散属性数据得到数字型连续化处理，而且加强了校园网络数据集中各个状态之间的关联性。

2) 校园网络数据的标准化

为了消除校园网络数据之间的量纲差异，采用以下式(2)对连续化处理后的数据属性值 x_{ij} 进行标准化处理。

$$x'_{ij} = value(b_i)\frac{x_{ij}-\bar{x}_j}{S_j},(i=1,2,\cdots,n;j=1,2,\cdots,m) \quad (2)$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (3)$$

$$S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} \quad (4)$$

其中，$\bar{x}_j$ 和 S_j 分别表示连续化处理后校园网络数据第 j 个特征属性的均值和标准差；n 和 m 分别表示校园网络数据集中的数据个数和属性数目，根据上述式(4)即可将连续化处理后的校园网络数据集转换到一个标准的单位空间，消除量纲差异。

3) 校园网络数据归一化处理

在完成上述校园网络数据集标准化处理后，虽然将校园网络数据集转换到一个标准的单位空间，但其可能取值并不在[0,1]区间内，针对这一现象，可以通过采用以下归一化处理式(5)将校园网络数据集中的各个数据特征属性的取值范围归纳到[0,1]区间内，完成校园网络数据预处理[7]。

$$x''_{ij} = \frac{x'_{ij} - \min\limits_{1\leqslant i\leqslant n}(x_{ij})}{\max\limits_{1\leqslant i\leqslant n}(x_{ij}) - \min\limits_{1\leqslant i\leqslant n}(x_{ij})} \quad (5)$$

上式中，和 $\min\limits_{1\leqslant i\leqslant n}(x_{ij})$ $\max\limits_{1\leqslant i\leqslant n}(x_{ij})$ 表示校园网络数据第 j 个特征属性的最大值和最小值。

2.2 基于模糊 C 均值聚类的校园网络数据聚类

假设 $X=\{x_1,x_2,\cdots x_n\}$ 表示预处理后获得的校园网络数据样本集，其中，x_i 表示校园网络数据样本特征向量，将该校园网络数据样本集 X 划分成 c 类，计算每组数据的聚类中心 v_j 和模糊隶属度 h，使得以下目标函数值 J 最小

$$J = x''_{ij}\sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^h\|x_i - v_j\|^2 \quad (6)$$

$$\sum_{j=1}^{c}(u_{ij}) = 1,\forall(i=1,2,\cdots,n \quad (7)$$

其中，u_{ij} 表示校园网络数据样本集中第 i 个样本属于第 j 个聚类中心的隶属度；v_j 表示第 j 个聚类中心；h 表示模糊隶属度，一般 $h>1$，数据挖掘方法中的模糊 C 均值聚类是通过随机选取一个聚类中心，迭代目标函数 J，寻找 J 的最小值，并不断调整参数 u_{ij} 和 v_j，获得校园网络数据样本最佳类别的过程，其中 u_{ij} 和 v_j 的计算公式分别如下

$$u_{ij} = 1/\sum_{t=1}^{c}\left(\|x_i - v_j\|/\|x_i - v_t\|\right)^{2/h-1} \quad (8)$$

$$v_j = \sum_{i=1}^{n} u_{ij}^h x_i / \sum_{i=1}^{n} u_{ij}^h \quad (9)$$

2.3 校园网络入侵数据检测

在完成校园网络数据样本聚类基础上，提出了一种异常簇识别方法，利用该方法能够实现校园网络入侵数据检测。假设聚类后的一个校园网络数据集 D 可以被划分成 $\{C_1, C_2, \cdots, C_k\}$ 个簇，可以通过计算校园网络数据集 D 中簇 C_i 与其他所欲簇之间距离的加权平均值获得簇 C_i 的异常度，具体计算公式如下

$$UN(C_i) = \sum_{j=1}^{k} |C_j| / |D| \cdot d(C_i, C_j) \tag{10}$$

其中，$|C_j| / |D| \cdot d(C_i, C_j)$ 代表校园网络数据集 D 中簇 C_i 偏离簇 C_i 的程度，通过计算 $UN(C_i)$ 能够度量簇 C_i 与整个数据集 D 的偏离程度，其取值越大，说明偏离程度越大，相反，其取值越小，说明偏离程度越小。

根据上述公式计算校园网络数据集 D 中每个簇的异常度，按照从大到小的顺序排列，并根据以下公式求出符合条件的最小簇数目 a，分别将 $C_{a+1}, C_{a+2}, \cdots, C_\kappa$ 和 $C_1, C_2, \cdots, C_a$ 分别标注为正常簇和异常簇，完成校园网络数据集 D 中的异常簇识别，实现校园网络入侵数据检测。

$$\frac{\sum_{i=1}^{a}}{|D|} \geqslant \varepsilon, 0 < \varepsilon < 1 \tag{11}$$

其中，ε 表示检测阈值，ε 的取值大小直接影响检测效率大小和检测精度高低，ε 的取值越大检测效率越高，但如果其取值过大则又会降低检测精度，一般情况下，$\varepsilon \leqslant 0.89$ 检测效果最佳。

2.4 基于入侵数据检测的自主防御体系构建

1) 校园网络入侵自主防御体系构建

为了阻止校园网络中的各种入侵行为，基于校园网络入侵数据检测结果构建了一个三维立体自主防御架构[8]，具体如图 1 所示。

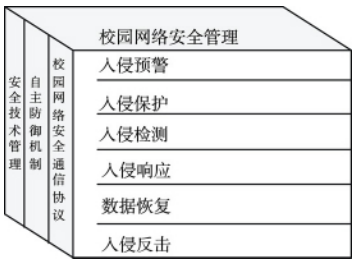


图 1 校园网络入侵自主防御体系三维立体架构

根据图 1 可以看出，校园网络入侵自主防御体系三维立体架构共分为三个层面，分别是校园网络安全管理、校园网络入侵数据自主防御机制和校园网络安全技术管理，其中，校园网络安全技术管理层面又包括入侵预警、入侵保护、入侵检测、入侵响应、数据恢复、入侵反击 6 层和一个校园网络安全通信协议。校园网络安全技术管理对各层技术进行管理；校园网络入侵数据自主防御机制负责协调 6 层技术的统一协调；校园网络安全管理则负责统筹所有关于校园网络安全管理的工作。

①入侵预警，是对校园网络中可能发生的入侵行为给出提前预警；

②入侵保护是指采用一切可能的手段保护校园网络信息系统的安全性、完整性和机密性；

③入侵检测主要用于实时监控和检测校园网络中的异常数据流，阻止校园网络入侵行为；

④入侵响应是指自主防御体系对危及校园网络信息系统安全事件和行为作出的反应，尽可能阻止和降低各种入侵行为给校园网络信息系统带来的破坏和损失；

⑤数据恢复是指自主防御体系能够及时地恢复校园网络信息系统，使得系统能够尽快恢复正常工作运行，包括对校园网络信息系统所存储有价值资源的备份恢复以及系统本身的备份和恢复工作；

⑥入侵反击，即通过应用各种网络攻击技术手段对校园网络入侵者进行反向攻击，迫使其停止对校园网络的入侵，但需要注意的是反向攻击的实施需要严格遵守国家法律规定和道德标准；

⑦校园网络安全通信协议，如果用户在校园网络通信过程中被攻击者冒用身份，校园网络内部机密信息很可能发生泄漏或篡改，整个校园网络安全面临挑战，由此说明是整个自主防御体系正常运行的根本技术支持，用于保障校园网络各个子系统的通信安全。

2) 校园网络入侵自主防御流程

基于校园网络入侵数据检测结果设计了包括入侵预警、入侵保护、入侵检测、入侵响应、数据恢复、入侵反击的自主防御机制，具体工作流程如图 2 所示。

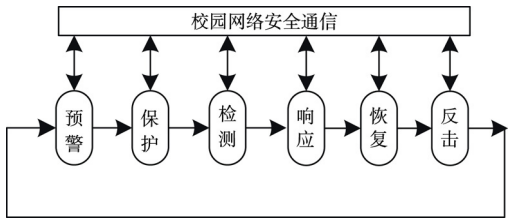


图 2 校园网络入侵自主防御流程

①预警是根据上述校园网络入侵数据检测结果对校园网络中已经存在的入侵行为和可能存在的入侵行为发出预警；

②校园网络系统中的各种保护手段会对自主防御体系发出的预警信息作出回应，从而最大限度地阻止校园网络入侵行为；

③入侵检测手段包括基于数据挖掘的各种入侵检测方法，用于实时监控和检测校园网络系统安全漏洞和入侵行为；

④只有准确、快速地响应才能将校园网络入侵行为造成

的破坏和损失降至最低，自主防御体系可以运用各种技术手段相结合的方法寻找和定位攻击源进行攻击取证，为校园网络入侵行为法律诉讼和方向攻击提供法律支持；

⑤一旦检测到校园网络存在入侵行为，处理及时阻止之外还应及时恢复校园网络各系统中被破坏的数据信息，保证校园网络的正常运行和服务；

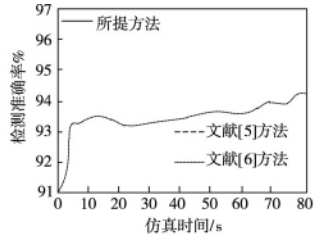
⑥反击是自主防御体系工作流程的最后一步。根据上述获得的校园网络入侵者的详细信息后，综合运用各种网络攻击技术手段对校园网络入侵者进行反向攻击，实现自主防御。

3 仿真与结果分析

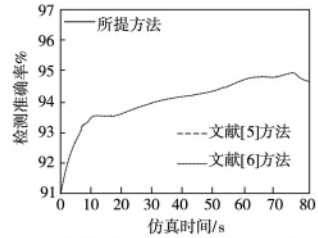
仿真在操作系统为 Windows7，语言编译环境为 Microsoft Visual C + +6.0，CPU 为 Intel P42.4GHz，运行内存为 4GB，硬盘内存为 60GB 的 PC 机上进行，选取从 VXHeaven 随机下载 800 个恶意代码样本作为校园网络入侵数据，记录某大学校园网络连续 15 天的 10000 条通信记录作为实验数据集。

通过选取检测率、误报率、准确率和自主防御成功率四项指标作为度量标准，对比基于 PCA - BP 神经网络的校园网络入侵数据自主防御方法（文献 [5] 方法）、基于特征选择的校园网络入侵数据自主防御方法（文献 [6] 方法）和基于数据挖掘的校园网络入侵数据自主防御方法的性能优劣。

如图 3 所示给出了不同初始训练集下（比例大于等于 20% 为较大比例训练集；小于 20% 为较小比例训练集）校园网络入侵数据的检测准确率对比情况。



（a）初始训练集较小时三种方法的准确率对比

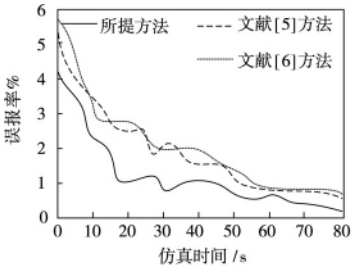


（b）初始训练集较大时三种方法的准确率

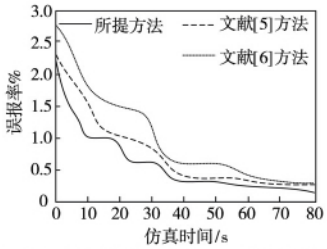
图 3 不同初始训练集下三种方法的准确率

根据图 3 可以看出，随着仿真时间的推移，三种不同方法的检测准确率均随着初始训练集的变化而变化，并且文献 [5] 方法在已知校园网络入侵数据训练样本较大情况下能够得到比所提方法和文献 [6] 方法更高的检测率；但在已知校园网络入侵数据训练样本较小情况下，所提方法的检测准确率更胜一筹。

根据以上实验可知，文献 [5] 方法在校园网络数据初始训练集较大情况下取得了较好的入侵检测效果，但对于校园网络入侵检测往往更关注误报率、检测率和自主防御成功率三个方面的检测效果，如图 4、图 5 和图 6 所示分别给出了不同初始训练集下校园网络入侵数据的误报率、检测率以及自主防御成功率对比情况。

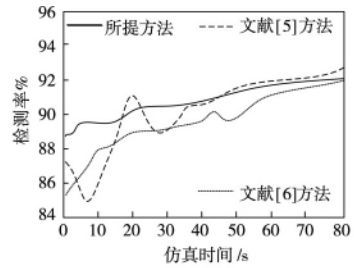


（a）初始训练集较小时三种方法的误报率

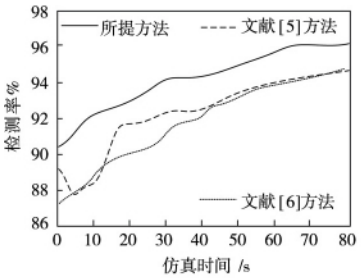


（b）初始训练集较小时三种方法的误报率

图 4 不同初始训练集下三种方法的误检率



（a）初始训练集较大时三种方法的检测率



（b）初始训练集较小时三种方法的检测率

图 5 不同初始训练集下三种方法的检测率

从图 4 中可以明显看出，在误报率对比方面，文献 [5] 方法在校园网络数据初始训练集较大情况下得到的误报率与

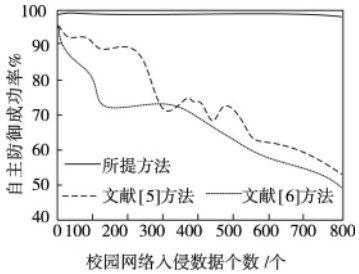


图6　三种方法的自主防御成功率对比

所提方法相差不大，文献[6]方法在校园网络数据初始训练集较小情况下由于存在许多干扰和冗余数据引入了更多的误报；在检测率对比方面，三种不同方法的检测率都有不同程度地提升，但所提方法在校园网络数据初始训练集较小情况下的检测率明显文献[5]方法和文献[6]方法，这是由于所提方法在进行校园网络入侵数据检测之前进行了预处理，消除了数据集中的冗余和干扰数据影响，有效提高了检测率，同时减小了误报率；自主防御成功率方面，所提方法几乎能够实现校园网络入侵数据的完全自主防御。

4　结束语

校园网络安全无时无刻不受到来自内部或外部的影响，实现网络入侵数据自主防御势在必行，研究在吸取前人研究经验教训基础上，提出了基于数据挖掘的校园网络入侵数据自主防御方法，通过仿真证明了所提方法在已知校园网络初始训练集比例较小时具有较高的检测率、检测准确率和自主防御成功率，同时有效减小了误报率，对校园网络今后应对更多新型网络入侵提供了有效技术支持。

参考文献：

[1] 顾兆军，何波. 基于可疑队列的多源攻击图入侵检测方法[J]. 计算机工程与设计，2017，38(6)：1408－1413.
[2] 高一为，周睿康，赖英旭，等. 基于仿真建模的工业控制网络入侵检测方法研究[J]. 通信学报，2017，38(7)：186－198.
[3] 张春琴，谢立春. 云环境中改进FCM和规则参数优化的网络入侵检测方法[J]. 电信科学，2018，34(1)：72－79.
[4] 贺炜. 远程网络通信安全性防御恶性入侵仿真研究[J]. 计算机仿真，2017，34(6)：294－297.
[5] 戴远飞，陈星，陈宏，等. 基于特征选择的网络入侵检测方法[J]. 计算机应用研究，2017，34(8)：2429－2433.
[6] 陈虹，万广雪，肖振久. 基于优化数据处理的深度信念网络模型的入侵检测方法[J]. 计算机应用，2017 37(6)：1636－1643.
[7] 梁辰，李成海，周来恩. PCA－BP神经网络入侵检测方法[J]. 空军工程大学学报(自然科学版)，2016，17(6)：93－98.
[8] 许学添. 基于模糊约束的网络入侵检测方法[J]. 西安工程大学学报，2016，(5)：627－632.

[作者简介]

张代华(1973－)，男(汉族)，湖北荆门人，硕士研究生，高级实验师，研究方向：数据挖掘，大数据应用，计算机网络。

沈　勇(1970－)，男(汉族)，江苏扬州人，硕士研究生，副教授，研究方向：移动互联网，物联网，信息安全。

章翔飞(1981－)，男(汉族)，江苏镇江人，硕士研究生，实验师，研究方向：计算机网络，信息系统安全。

王　兵(1982－)，男(汉族)，湖北武汉人，硕士研究生，实验师，研究方向：数据管理，软件开发，网络技术。

(上接第257页)

[9] 王玉珏，吴庆州，黄羽，等. 最大化网关流量的物联网路由的研究[J]. 现代电子技术，2019，42(13)：19－22，27.
[5] 王晓婷，钱谦. 基于搜索集中度和动态信息素更新的蚁群算法[J]. 电子测量技术，2019，42(9)：35－39.
[6] 杨娜，贾磊. 强干扰环境下通信传输信号多路同步采集系统设计[J]. 科学技术与工程，2018，18(4)：304－309.
[7] 何亮亮，王晓东. 基于初始信息素和二次挥发的改进蚁群算法[J]. 西安工程大学学报，2018，32(6)：739－744.
[8] 姜晶，张宪，于云选，等. 基于MD5算法的物联网传输模块设计[J]. 传感器与微系统，2017，36(7)：124－126.
[9] 盖昊宇，张震，李慧. 混合服务策略轮询特性下物联网传感节点设计[J]. 佳木斯大学学报(自然科学版)，2019，37(6)：900－903，916.
[10] 张茜，杨秋翔，孔德云，等. 基于动态信息流的Android应用检测[J]. 计算机工程与设计，2017，38(10)：2646－2651.
[11] 欧阳利，林岩，张烽. DVBC标准下传输流解复用器的软件系统设计[J]. 单片机与嵌入式系统应用，2017，17(8)：9－12.
[12] 谢宾，刘曦，林群，等. 采用串行通信接口的同步时分多路复用总线通信方法[J]. 自动化与仪器仪表，2019，(5)：161－163.

[作者简介]

王云峰(1968－)，男(汉族)，甘肃秦安人，博士，教授，研究方向：证据科学、公安技术、信息安全等方面研究。