

基于操作注意力和数据增强的内部威胁检测

冯冠云, 付才, 吕建强, 韩兰胜

1. 分布式系统安全湖北省重点实验室, 湖北省大数据安全工程技术研究中心, 湖北 武汉 430074;
2. 华中科技大学网络空间安全学院, 湖北 武汉 430074)

摘要: 内部威胁是组织中出现重大安全隐患的主要原因之一, 也是一个长期的挑战。通过分析现有的内部威胁数据, 指出内部威胁检测最大的挑战在于数据不平衡、有标注的威胁样本少。内部威胁检测的经典数据集 CMU-C R4.2 共有 322 万条日志数据, 其中标记出的恶意操作日志仅 7 423 条; 日志中的大多数操作类型与恶意行为无关, 如泄露企业数据这一恶意行为仅与两种类型操作高度相关, 而其余的 40 多种类型操作的日志则可能对检测造成干扰。针对这一挑战, 设计了一种基于操作注意力和数据增强的数据处理框架。该框架首先对操作进行异常评估, 对低异常评分的操作进行掩码操作, 使模型更好地关注与恶意行为相关的操作, 可以被认为是一种操作的硬注意力机制。通过分析内部威胁数据集的特点, 设计了 3 种规则对恶意样本进行数据增强, 以增加样本的多样性和缓解正负样本严重不平衡的问题。将有监督的内部威胁检测视作一个时序分类问题, 在长短期记忆卷积神经网络 (LSTM-FCN) 模型中加入残差连接以实现多粒度的检测, 并使用精确率、召回率等指标实施评估, 要优于现有的基线模型; 另外, 在 ITD-Bert、TextCNN 等多种经典模型上实施基于操作注意力和数据增强的数据处理框架, 结果表明所提方法能够有效提升内部威胁检测模型的性能。

关键词: 内部威胁检测; 硬注意力; 数据增强; 神经网络

中图分类号: TP309.2

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023042

Insider threat detection based on operational attention and data augmentation

FENG Guanyun, FU Cai, LYU Jianqiang, HAN Lansheng

1. Hubei Engineering Research Center on Big Data Security, Hubei Key Laboratory of Distributed System Security, Wuhan, 430074, China
2. School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: In recent years, there has been an increased focus on the issue of insider threats. Insider threats are a

收稿日期: 2022-12-01; 修回日期: 2023-03-15

通信作者: 付才, fucai@hust.edu.cn

基金项目: 国家自然科学基金 (62072200, 62172176); 国家重点研发计划 (2022YFB3103400)

Foundation Items: The National Natural Science Foundation of China (62072200, 62172176), The National Key R & D Program of China (2022YFB3103400)

引用格式: 冯冠云, 付才, 吕建强, 等. 基于操作注意力和数据增强的内部威胁检测[J]. 网络与信息安全学报, 2023, 9(3): 102-112.

Citation Format: FENG G Y, FU C, LYU J Q, et al. Insider threat detection based on operational attention and data augmentation[J]. Chinese Journal of Network and Information Security, 2023, 9(3): 102-112.

major cause security breaches in organizations and pose an ongoing challenge. By analyzing the existing insider threat data, it was identified that the biggest challenge in insider threat detection lies in data imbalance and the limited number of labeled threat samples. In the Cert R4.2 dataset, which is a classic dataset for insider threat detection, there are over 3.22 million log data, but only 7,423 are marked as malicious operation logs. Furthermore, most of the operation types in the logs are not related to malicious behavior, and only two types of operations are highly correlated with malicious behavior, such as leaking company data, creating interference in the detection process. To address this challenge, a data processing framework was designed based on operational attention and data augmentation. Anomaly evaluation was first performed on operations by the framework, and operations with low anomaly scores were then masked. This makes the model better focus on operations related to malicious behavior, which can be considered as a hard attention mechanism for operations. Next, the characteristics of the insider threat dataset were analyzed, and three rules were designed for data augmentation on malicious samples to increase the diversity of samples and alleviate the substantial imbalance between positive and negative samples. Supervised insider threat detection was regarded as a time-series classification problem. Residual connections were added to the LSTM-FCN model to achieve multi-granularity detection, and indicators such as precision rate and recall rate were used to evaluate the model. The results indicate superior performance over existing baseline models. Moreover, the data processing framework was implemented on various classic models, such as ITD-Bert and TextCNN, and the results show that the methods effectively improve the performance of insider threat detection models.

Keywords: Insider threat detection, hard attention, data augmentation, neural network

0 引言

内部威胁是组织中出现重大安全隐患的主要原因之一。内部威胁通常由组织的员工发起，他们通过恶意行为对组织造成危害^[1]。内部威胁被视为最具挑战性的威胁之一，2022年发布的《内部威胁网络安全报告》显示，在过去两年中，内部威胁事件增加了44%，每次事件的成本增加了1/3以上，达到15万美元^[2]。根据Gurucul的内部威胁调查报告，74%的组织非常担心内部威胁，仅有3%的组织表示不关心^[3]。在2023年的网络安全报告中，60%的受访者表示他们在过去一年受到了内部攻击，而8%的受访者甚至受到了20多次内部攻击^[4]。此外，48%的受访者认为，内部攻击比外部攻击更难检测和预防^[5]。与外部入侵者相比，内部人员易于进行攻击。与外部人员发起的攻击相比，内部攻击的检测相当困难。这是因为内部人员是系统的授权用户，这使得在恶意和合法活动之间划清界限变得相当困难^[6]。并且内部威胁攻击以各种方式表现出来，如叛徒内部人员植入逻辑炸弹以破坏原始系统或窃取机密数据。内部攻击的多样性增加了检测此类内部威胁的复杂

性。内部攻击往往是由恶意的内部员工在日常工作时间内进行的，从而将内部人员的异常活动分散在大量正常员工的行为日志中。而且，由于日常活动的数量非常巨大，使用人工分析此类活动日志相当困难^[7]。

目前，工业界和学术界提出了许多内部威胁检测方法。主流的检测方法主要基于日志时间序列建模，由于基于无监督学习的威胁检测误报率高，需要较高人力成本，本文主要关注基于时序分类模型的有监督检测。分析现有数据集发现，有监督内部威胁检测存在以下挑战。① 日志中的操作类型十分复杂，其中与恶意行为相关的操作仅占极少数。Cert数据集中记录了登录、设备、邮件、网站、文件的日志表，而每个表中又包含多种具体操作。并非所有的操作都与恶意行为存在关联性，恶意行为可能仅表现为某几种具体操作的异常，因此会被大量正常操作的数据干扰。② 内部威胁数据通常是不平衡的数据集，在海量的日志数据中，通常只有少数的恶意行为被标注。在Cert数据集中，共约327万条日志信息，而其中标注出的恶意日志信息仅有7 423条，这给深度学习带来了极大的挑战。

因此，在恶意数据少、正负样本不均衡的条

件下实施有效的检测成为内部威胁检测的重大挑战。为此, 本文的检测框架中提出了两种用于增强内部威胁检测性能的数据处理方法。另外, 本文在标准长短期记忆卷积神经网络(LSTM-FCN)模型中加入残差连接以实现多粒度检测。实验表明, 本文方法能够提供优于目前主流模型的性能, 且本文框架的数据处理方法可以应用于其他主流模型, 起到提升模型性能的作用。

本文的研究主要有以下贡献。

1) 提出了一种内部人员日志数据增强方案, 分析内部威胁数据集的特点, 并制定了3种数据增强方式用于生成恶意的内部威胁日志。实验表明, 使用增强数据训练内部威胁检测模型能提升模型的鲁棒性。

2) 提出了一种基于操作注意力的数据降噪方法, 该方法使用 Anomaly Transformer^[8]对每种操作进行异常检测, 排除规律性变化的操作并进行掩码, 从而让模型更好“注意”异常操作, 减少正常操作对内部威胁检测带来的干扰。

3) 在 LSTM-FCN 模型中加入残差连接, 来达到多粒度异常检测的目的。使用 CERT 数据集^[9], 对本文提出的内部威胁检测方案常用的分类指标进行了评估, 取得了优于基线模型(ITD-Bert、LSTM、TextCNN)的效果。

1 相关工作

1.1 内部威胁检测

由于用户的行为日志存在明显的时序关联, 主流检测方法主要通过记录用户行为特征, 并基于时间序列数据建模。可以根据时间序列数据是否有标签将内部威胁检测分为无监督和有监督两类。

无监督的内部威胁检测大部分基于时序数据的重建, 在重建并预测人员行为的过程中找出偏离预测的行为。Tour 等^[10]使用循环神经网络进行重建, 并与隔离森林、支持向量机等机器学习模型进行对比。Liu 等^[11]和 Sharma 等^[12]分别使用深度自编码解码器和长短期记忆(LSTM)网络实现无监督的重建任务。Liu 等^[13]将员工行为作为语料库, 使用 word2vec 模型重建并预测用户下一步行为的概率。还有研究将无监督问题转换成有监督的图像分类问题^[14], 但其本质仍是利用离群

的异常数据难以在训练中被拟合的特点。无监督的内部威胁检测通常需要人工设定阈值来判定行为是否异常, 且检测完成后需要进一步人工校对。另外, 无监督学习计算开销大且误报率较高, 这可能产生更高的人工成本。

有监督的内部威胁检测通常被视作时序分类任务, 所有相关日志的统计信息被组合在一起形成特征向量, 并使用分类模型判定其是否包含恶意行为。许多模型已经被用于该分类任务, 包括决策树^[15]、XGBoost、隐马尔可夫模型^[16]、隔离森林^[17]、随机森林^[18]等机器学习模型, 以及自编码解码器^[19]、LSTM^[20-22]、CNN^[23]、Transformer^[24]等深度学习模型。另外, 有研究将时序数据变成图片的维度, 并使用 VGG16 等图像分类模型进行检测^[25], 也取得了不错的检测性能。有监督的内部威胁检测比无监督内部威胁检测准确率高, 但恶意标记数据少、难以获取, 存在正负样本不平衡的问题。以 Cert R4.2 数据集为例, 正负样本比例约为 500:1, 这给分类模型带来了极大的挑战。

一些内部威胁检测的研究通过加入更多的特征来提升检测性能。部分研究手动提取更高维的特征, 如数据之间的方差等。Legg^[26]等通过构建树模型将每个员工过去的日志数据与当前的日志数据做比较, 并通过设计规则来判定员工的行为是否异常。部分研究使用员工社交网络信息, 手动提取用户行为属性特征^[16], 并使用 XGBoost 进行检测。Jiang 等^[27]将员工社交网络信息用于构建图神经网络。部分研究提取日志中的文本信息, 并通过分析文本内容提取特征。例如, Chattopadhyay^[20]等设计了文本内容的特征提取方法, 还有一些研究^[23-24]使用文本特征和自编码器、LSTM、决策树等模型进行内部威胁检测。这些方法引入额外的特征以达到更高的检测性能, 但提取额外特征依赖于专家经验, 且难以应对未知的恶意行为。

本文旨在尽可能自动化地实施内部威胁检测, 因此考虑仅使用用户行为特征情况下的有监督的内部威胁检测。这是因为无监督的内部威胁检测和属性特征都需要花费更多的人工成本, 且属性特征的设计依赖于专家经验, 难以适用于其他数据。

1.2 注意力机制

学者根据对人类视觉的研究,提出了注意力机制^[28],将其用于实现信息处理资源的高效分配。注意力机制从本质上讲和人类的选择性视觉注意力机制类似,核心目标是从众多信息中选择出对当前任务目标更关键的信息。注意力机制可以分为软注意力机制和硬注意力机制^[29]。

软注意力机制以神经网络训练出的权值,在空间或通道上加权全局输入特征,从而达到关注特定空间区域或通道的目的。与图像、自然语言处理相关的很多研究使用软注意力机制。其中,首次提出注意力机制的是在图像分类任务中^[30],许多研究在图像上应用注意力机制,包括空间注意力^[31]、通道注意力^[32]、自注意力^[33-34]等。在自然语言处理领域,Transformer 仅将自注意力机制应用于机器翻译^[35],随后 Bert 模型使用注意力机制,应用 Transformer 的编码器部分并加入了预训练任务^[36]。但软注意力机制需要大量计算开销,与目标模型高度耦合。

硬注意力机制筛选出受关注区域并将其当作输入,在图像研究领域剔除无意义的背景数据后输入,能够有效聚焦目标物体。硬注意力机制采用直接限制输入内容的处理方法,使用聚类、强化学习等算法筛选出图像中需要关注的区域。例如,Zhao 等^[37]应用强化学习实现硬注意力机制,硬注意力被应用到图像通道中^[38]。Elsayed 等^[39]自监督预训练估计了不同图像块与下游任务的相关性,在 ImageNet 数据集上实现了高准确率。Anderson 等^[40]则使用 Faster R-CNN 实现基于自下而上的注意力模型。

当前的内部威胁检测工作在模型上进行了许多尝试,但仍未解决数据集正负样本不平衡、数据分布不平衡的问题。在内部威胁数据集中不同操作与恶意行为的关联性有差异,且恶意行为容易隐藏在大量的正常行为中。硬注意力机制能够从输入信息中选择重要的特征,因此被视为一种比软注意力机制更高效、直接的方法。本文通过用于内部威胁检测操作层面的硬注意力机制来选择与恶意行为更加相关的操作,从而实现高性能的有监督内部威胁检测。

2 内部威胁检测框架

考虑到内部威胁数据集冗余操作多、正负样本极度不平衡的特点,本文提出一种内部威胁检测框架。首先对内部威胁数据实施操作注意力机制和数据增强以获取更高的内部威胁检测性能,其次使用多粒度 LSTM-FCN 模型对日志数据进行检测。内部威胁检测架构如图 1 所示。

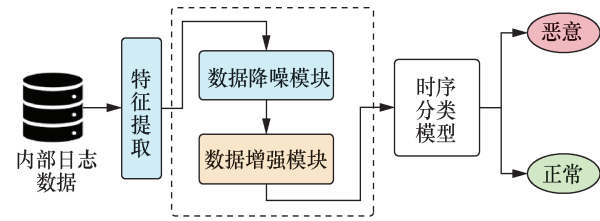


图1 内部威胁检测架构
Figure 1 Insider threat detection architecture

2.1 预处理

内部威胁数据集由多个日志表组成,每个日志表记录了员工的具体操作和时间戳。假定数据集中包含记录了 T 天内的员工活动日志。员工 u 第 t 天的特征向量可以表示为 $\mathbf{X}_t = (x_{t1}, x_{t2}, \dots, x_{tk})$ 其中 k 表示具体操作的种类, x_{ti} 表示第 i 种操作的执行次数。特征向量 $\mathbf{X}_t$ 为员工第 t 天的特征。由于员工的行为习惯不同,直接使用操作次数作为特征是不合理的。因此,在计算计数特征向量之前,每个员工的单日特征需要进行归一化处理。

内部威胁检测的目的在于检测一个时间窗口 W 内,用户 u 是否存在恶意行为。本文基于员工 u 在时间窗口 W 内的单日特征构建时序特征。构造时序特征向量时,使用时间窗口 W ($W \leq T$) 在 T 天员工活动日志上进行滑动。因此,对于每个员工 u ,有 $T-W+1$ 个窗口的时序特征。对于每个窗口时序特征,如果其中包含恶意操作则将其标记为恶意,反之则标记为正常。

内部威胁检测模型的输入是内部人员操作日志一个窗口时间的序列特征,即特征维度 k 、时间步长为 w 的张量。输入特征向量的每个特征维度代表一种日志操作(如登录他人计算机),时间步长代表日期;输出为该内部人员在窗口时间内的操作是否存在异常。因此,内部威胁检测在本文中可以被视作一个时序二分类问题。

2.2 基于操作注意力的掩码降噪

在很多场景下,难以仅挑选出与内部威胁行为有关的操作,这意味着一些与行为是否存在威胁无关的冗余操作也被作为模型的输入。降噪模块旨在进行初步的异常检测,筛选出规律变化的无关操作,从而达到降噪的目的。这是一种基于操作的硬注意力机制,不同的操作与恶意行为的关联性不同,因此应当给予不同的权重。在本文中使用的权重为 0 或 1,表示该操作特征是否被保留。

本文使用 Anomaly Transformer 进行无监督训练,并对输入数据的每种操作进行异常打分。对于异常打分低的操作,使用掩码覆盖,这降低了与恶意行为无关操作的权重,能够起到降噪的效果。基于异常检测的操作硬注意力机制如图 2 所示,红框内的操作变化规律出现异常,而内部威胁检测正需要挖掘这些异常之间的关联。操作 1 和操作 3 的特征曲线是规律变化的,因此将这类正常操作通过掩码的方式抹去,仅保留与恶意行为相关的操作,从而提升检测模型的性能。

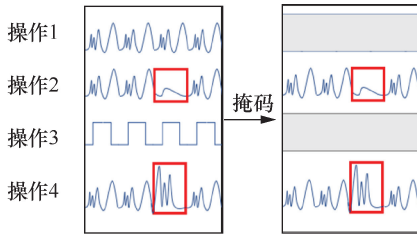


图2 基于异常检测的操作硬注意力机制
Figure 2 Operational hard attention mechanism based on anomaly detection

Anomaly Transformer 是一种通过编码、解码对时序信息进行异常打分的方法。该方法通过同时对先验关联和序列关联进行建模,以计算关联差异,发掘时间序列信息中的关联性。

先验关联采用可学习的高斯核函数,其中中心在对应时间点的索引上,用于挖掘数据与时序信息中邻近信息的关联。序列关联由 Transformer 中注意力计算得到,一个时间点的序列关联为该点在注意力矩阵中对应的注意力权重分布。序列关联可以挖掘数据与整个时间序列信息之间的关联。先验关联 P^l 和序列关联 S^l 的计算式如下。

$$P^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j \in \{1, \dots, N\}}\right) \quad (1)$$

$$S^l = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{\text{model}}}}\right) \quad (2)$$

式(1)中, P^l 使用高斯核函数计算不同时间步之间的关联权重,其中 i, j 分别表示第 i, j 个时间步, σ_i 表示学习的高斯核参数。Rescale 表示对该关联权重进行归一化。这种设计利用高斯分布的单峰特点,使模型能更好地关注邻近时间步的信息。式(2)中, S^l 与标准 Transformer 中注意力计算方式相同, Q 和 K 使用全连接层计算得到,相乘后使用矩阵行数缩放并进行 Softmax。一个时间步的序列关联即该点在注意力矩阵中对应行的注意力权重分布。相较 P^l 关注邻近关联,序列关联 S^l 更关注全局关联。

在重建过程中,首先,最小化阶段先验关联 P^l ,使先验关联 P^l 近似从原始时序中学得的序列关联 S^l ,该过程将使先验关联适应不同的时序模式;其次,最大化序列关联 S^l ,从而最大化关联之间的差异,该过程使序列关联更加注意非邻近时间步的信息。异常数据通常与邻近数据关联更加紧密,但难以与全局数据建立关联。最小化先验关联并最大化序列关联后,异常数据难以通过邻近时间步的信息进行重建,而正常数据可以通过全局信息之间的关联进行重建,从而有效区分异常与非异常的时序数据。

重建误差越小,表示该操作的变化越具有规律性,则越可能是与内部威胁无关的操作。由于实际日志操作不总是完全规律的,因此即便是非异常操作也存在极小的重建误差(通常小于 1×10^{-4})。因此以重建误差最低的操作作为基线,小于基线误差 α 倍的操作则被视作与恶意行为无关的操作,通过掩码屏蔽进行降噪。在具体的计算过程中,Anomaly Transformer 对不同操作的重建误差评估通常会出现数量级的差异,难以作为操作的权重。因此,使用掩码覆盖较小重建误差的操作而非将异常打分作为权重,这与通常的硬注意力机制做法一致。

2.3 内部威胁数据增强

内部威胁数据集中已标记的恶意样本往往仅占总体样本的极少数,因此其是一个正负样本极

度不平衡的数据集。内部威胁数据集由大量离散的日志构成,通过编码可以得到一系列的时序信息。内部威胁数据有以下特征。

(1) 全局周期性

企业内部的员工日志信息存在周期性,周期与周期之间的行为特征存在相似性。由于工作日、周末的存在,员工行为存在周期性。例如,以一个星期为周期,员工通常在工作日登录一次,而在周末不进行登录。假如某员工在工作日登录了多次,再结合其他可疑操作(如多次使用U盘读取数据),那么可以认定他有窃取企业数据的嫌疑。

(2) 局部随机性

由于工作生活中的复杂性,员工的行为存在局部随机性。例如,员工的行为可能因为临时加班、生病请假而打破周期性规律,但这并不意味着该员工实施了恶意行为。另外,恶意行为的持续时间、时间间隔具有随机性。以窃取数据为例,该行为实施可能在一个月内陆续实施,也可能仅在两天内就完成。

(3) 关联性

不同日志操作之间存在关联性,内部威胁行为通常由多种关联性操作的异常构成。内部人员实施恶意行为时,通常同时存在多种操作的异常;当多种操作同时异常才能认定该行为可能为恶意行为,而每种操作单独的异常并不能够判定其存在恶意行为。

考虑到内部威胁数据集的以上特征,本文提出3种数据增强方式以提高负样本的多样性,缓解正负样本不平衡的问题,从而获取更高的内部威胁检测性能。在数据增强过程中,应当保证尽可能不改变日志特征的语义,即保证标签不变的前提下增加数据的多样性。因此,在数据增强过程中,不同行为特征之间的相对关系应当保持不变。考虑到内部威胁数据集的全局周期性,对数据以周期为单位进行数据增强。本文提出了以下3种数据增强方式(如图3所示)。

(1) 交换

对于时间窗口 W 内的数据,员工行为存在一定周期性。本文以一星期(7天)作为一个周期,对窗口内不同周期的数据进行随机交换。注意,该交换不仅仅是恶意周期与正常周期数据的交

换,也可以是正常周期数据之间的交换。交换恶意周期与正常周期数据可以使恶意行为发生的时机具有更多的随机性,交换两个正常周期数据则能够在不改变语义情况下进行数据增强。

(2) 覆盖

由于恶意行为的随机性,时间窗口 W 内可能存在多个存在恶意行为的周期。因此建议使用已经标注的恶意周期对其他周期内的数据进行随机覆盖。覆盖操作不会改变该周期数据的标签,且增强了对恶意行为发生次数和持续时间的随机性。

(3) 融合

融合操作与覆盖操作类似,可以增强恶意行为的多样性以及恶意行为发生次数的随机性。本文使用两组存在恶意行为的数据进行融合,融合时保留恶意周期的数据,而正常周期数据则从两组数据中随机选取。

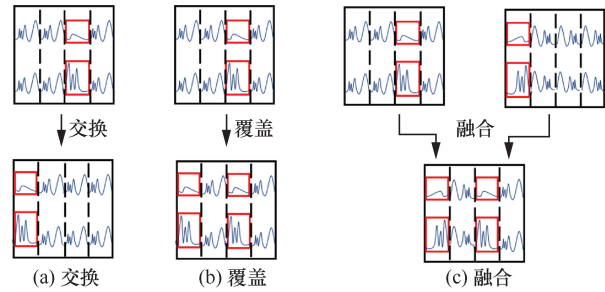


图3 内部威胁数据集的3种数据增强方式
Figure 3 Three data augmentations for insider threat datasets

2.4 多粒度长短期记忆全卷积神经网络

内部威胁检测可以视为一个日志信息时序分类问题,而LSTM-FCN在对时间序列信号进行分类的任务中表现出了先进的性能。LSTM-FCN的输入通常为时间序列信息,在内部威胁检测中输入为内部人员操作日志的序列特征。令 $\mathbf{x}_t \in \mathbb{R}^{F_0}$ 是特征维度为 F_0 、时间步长为 t 的输入特征向量。每帧的真实动作标签由 $y_t \in 1, \dots, C$ 给出,其中 C 是分类数。在内部威胁检测中,任务被视为一个 C 为2的二分类问题,即判断输入特征向量是否存在恶意行为。

LSTM-FCN模型结构如图4所示,输入的时序序列信息分别传入多层一维卷积以及LSTM网络中,最终通过拼接和softmax得到分类信息。在内部威胁数据集中,恶意行为的持续时间有不

确定性,因此需要不同粒度的检测。例如,用户 BIH0745 的恶意行为发生在 1 天内,而用户 AAM0658 的恶意行为持续了 6 天。浅层卷积提取细粒度短周期的特征,而深层的卷积能更好地提取粗粒度长周期的特征。相较于经典的 LSTM-FCN 模型,本文在卷积模块中加入残差结构,以加强模型多粒度的检测能力。具体来说,数据在不同层卷积后的特征图求和,以适应内部威胁行为的粒度差异。残差连接同时使用浅层卷积与深层卷积的信息,因此能够自适应地检测不同粒度的恶意行为。

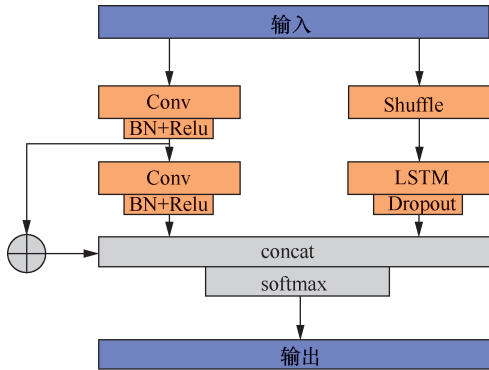


图4 LSTM-FCN 模型结构
Figure 4 LSTM-FCN model structure

LSTM 模块将输入时间序列作为具有单个时间步长的多元时间序列来接收,这是通过 shuffle 模块实现的,该模块对时间序列的时间维度进行转置。转换后的长度为 N 的单变量时间序列将被视为具有单个时间步长的多元时间序列(具有 N 个变量)。LSTM 的传递公式表示如下。

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \quad (3)$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \quad (4)$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \quad (5)$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \quad (6)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t \quad (7)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (8)$$

其中, x_t 是时间 t 的特征输入。 f_t 是遗忘门,其功能是决定上一时间步长的特征信息在 c_{t-1} 是否被 c_t 接收。 i_t 是输入门,其功能是控制哪些特征信息 g_t 用于更新 c_t ,其中 g_t 通过输入 x_t 和上一个时间步长隐藏层状态输出 h_{t-1} 计算得到。输出门

o_t 和 c_t 共同确定时间 t 时隐藏层状态的输出 h_t 。其中, $W_{i/f/g/o}$ 和 $b_{i/f/g/o}$ 分别表示权重矩阵和偏差值,在训练期间随机初始化和调整。

卷积模块由多层一维卷积构成,输入特征使用一维卷积提取局部特征。在每层一维卷积后,使用 batch norm 和 Relu 激活函数。之后,将输出的特征图用一个卷积层转换成一维张量,该张量为卷积模块提取的特征。

3 实验

实验环境如下: CPU 为 Intel Xeon Gold 5118 CPU @ 2.30 GHz, GPU 为 RTX 5000, 内存为 DDR4 2666 Hz 32 GB×4。主要软件环境: Python3.7, Pytorch1.7。代码开源在 Getee 仓库。

3.1 数据集

本文使用 CMU-CERT R4.2 来训练和测试所提出的方法。CMU-CERT 数据集是唯一可公开访问的基于内部威胁的员工行为数据集。因此,本文将提出的方法在 Cert 数据集上与现有的方法进行比较。

CMU-CERT R4.2 数据集由员工的行为特征日志构成,该数据集包括 18 个月内收集的 1000 个员工的活动日志。这些员工的活动日志包括登录、设备、文件、http 和电子邮件文件。该数据集包括 6 个.csv 日志表格文件,反映了每个员工行为的各个方面。它还包含一个 LDAP 文件夹,其中包含每个月在职所有员工的信息,具有以下字段:员工姓名、员工 ID、地址、职位、企业部门、运营单位、分支机构和主管。这 6 个.csv 日志表格文件分别包含以下信息。

(1) 登录

登录日志有 5 个字段: ID、日期、员工、PC、活动(登录/注销)。日期字段包含每个员工执行的登录或注销活动的日期和时间戳。

(2) 设备

设备日志有 5 个字段: ID、日期、员工、PC、活动(连接或断开连接)。日期字段包含连接或断开每个员工执行的任何可移动媒体活动的日期和时间戳。

(3) 文件

文件日志有 6 个字段: ID、日期、员工、PC、

文件名、内容。日期字段包含员工访问文件的日期和时间戳。文件名以 5 种不同类型的文件扩展名 (.doc、.txt、.jpeg、.pdf、.zip.exe) 存储。

(4) 网络

网络访问日志有 6 个字段：ID、日期、员工、PC、URL、内容。日期字段包含每个员工访问的 URL 的日期和时间戳。

(5) 电子邮件.csv

电子邮件日志有 10 个字段：日期、员工、PC、收件人、抄送、BC、发件人、大小、附件、内容。日期字段包含每个员工执行的电子邮件发送或接收活动的日期和时间戳。

(6) 心理测试

心理测试日志有 5 个字段：O、C、E、A、N。这些字段反映了 5 个人格特质的心理测量分数——“OCEAN”，其中“O”代表对经验的开放性，“C”代表责任心，“E”代表外向性，“A”代表宜人性，“N”代表每个员工的神经质。

3.2 特征提取

本文实验使用登录、设备、网络、文件、邮件 5 个日志表中的信息构建数据集。需要指出的是，原数据中的心理测试为入职时的评测，而恶意行为的实施通常发生在若干个月之后，因此该心理测试可能失去时效性。另外，心理测试数据的客观性有待商榷，在过往研究中，心理测试数据通常不被作为用户的特征。

具体地，根据每个表中的操作进行细分。例如，文件日志数据按照文件后缀进行分类，doc、pdf、txt 表示文本文档，jpg、png 表示图像文件，zip、rar 表示压缩数据。另外，根据操作发生 PC 与操作员工的 3 种关系（本人 PC、他人 PC、公用 PC），每种操作对应 3 个编码。

如图 5 所示，单日特征被编码成一个 42 维的特征向量。本文使用 28 天作为一个时间窗口进行内部威胁检测，因此检测模型输入的数据是一个 28×42 的张量数据。

为了训练和测试本文方法，按员工顺序将整个数据集分为训练和测试两个子集。前一个子集（85% 的数据）用于模型训练，后一个子集（15% 的数据）用于评估泛化性能。实验数据表明恶意行为更倾向于在离职前实施，因此

使用员工而非时间对训练、测试数据进行切分。每个员工的特征向量通过上述的特征向量方法提取。

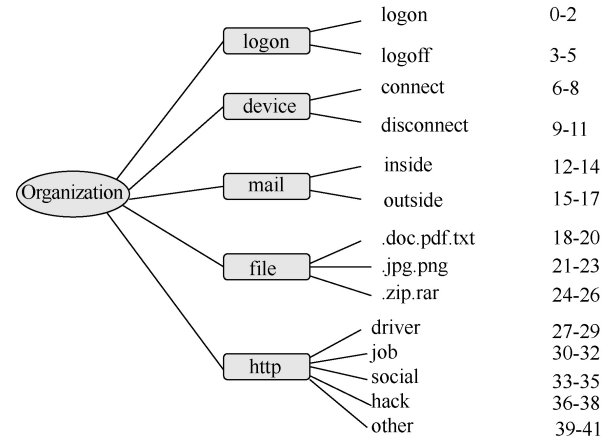


图 5 CMU-CERT R4.2 数据集特征编码方案
Figure 5 CMU-CERT R4.2 dataset feature encoding scheme

3.3 评估

本文使用 CMU-CERT R4.2 数据集进行评估。在使用的多粒度 LSTM-FCN 中，卷积部分使用了 5 层一维卷积，输出的通道数均为 128，卷积核的大小分别为 8、5、3、3、3。LSTM 部分使用双层 LSTM，输出的维度为 128。Batchsize 设置为 64，模型经过 50 个 epoch 的训练。

由于 CMU-CERT R4.2 是一个不平衡数据集，因此本文使用内部威胁检测的召回率（recall）、精确率（Precision）、F1 值作为评估指标。评估指标的计算如下。

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (9)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (10)$$

$$F1 = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (11)$$

其中，TP 为真阳性，TN 为真阴性，FP 为假阳性，FN 为假阴性。

本文选择多种基于不同网络结构的基线模型进行比较。处理时序的模型主要有循环神经网络（RNN）、卷积神经网络（CNN）、基于自注意力机制的 Transformer，这些模型都被广泛应用于内部威胁检测工作。将 LSTM-FCN 与 VIT、ITD-BERT、Bi-LSTM、TextCNN 这 4 种基线模型进行比较，结果如图 6 所示。其中，VIT 是基于

自注意力机制的模型，结构上使用 Transformer^[32]。ITD-BERT 模型则是 Transformer 应用在内部威胁检测领域的最新研究^[23]。另外，Bi-LSTM、TextCNN 分别使用循环神经网络和卷积神经网络，这在大量现有工作中被用于内部威胁检测^[21-22]，能够很好地处理内部威胁数据集中的时序信息。实验结果表明，本文方法取得了最高的 $F1$ 值评分 (93.39%)。另外，在这些基线模型上分别实施数据降噪 (Mask) 和数据增强 (Augment) 并评估其 $F1$ 值。橙色柱体为实施 Mask 和 Augment 后的 $F1$ 值，在多种模型上均有明显的性能提升，其中在 ITD-Bert 模型上 $F1$ 值提升了 8.6%，表明本文的数据处理方法是有效且鲁棒的。

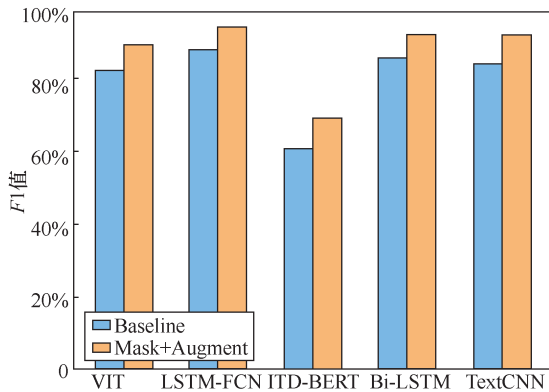


图 6 多种模型 $F1$ 值评估
Figure 6 $F1$ -score evaluation with multiple model

在分别使用本文框架中提出的数据降噪和数据增强后，LSTM-FCN 模型的召回率、精确率和 $F1$ 值表现如表 1 所示。在仅使用 LSTM-FCN 模型时，召回率和精确率分别为 84.23% 和 91.44%。如表 1 所示，单独使用数据降噪或数据增强均能在一定程度上提升内部威胁检测的性能表现， $F1$ 值有所上升；值得注意的是，在单独使用数据降噪后，精确率略微下降，这可能与模型的过拟合有关。由于数据集中存在海量的正常行为，因此正常行为可能在一些不重要特征中表现出过拟合，从而能够被识别为正常行为。对部分特征掩码后，部分过拟合的特征消失，因此一些正常行为被分类为异常。最后，在同时使用数据降噪和数据增强时，模型取得了最好的性能，即 94.08% 的召回率、92.71% 的精确率和 93.39% 的 $F1$ 值。

表 1 内部威胁检测数据处理消融实验
Table 1 Insider threat detection data processing ablation experiment

模型	召回率	精确率	$F1$ 值
LSTM-FCN 模型	84.23%	91.44%	87.68%
LSTM-FCN 模型+数据增强	86.20%	92.59%	89.28%
LSTM-FCN 模型+硬注意力	94.58%	90.99%	92.75%
LSTM-FCN 模型+混合方法	94.08%	92.71%	93.39%

在本文模型的卷积模块中，共有 5 层卷积层。如表 2 所示，实验评估了在第 5 层卷积中分别加入 3、4 层卷积残差后的检测性能。在第 3、4 层卷积和第 5 层之间加入残差连接后，召回率从 91.04% 提升至 94.08%， $F1$ 值从 92.62% 提升至 93.39%，但精确率从 94.25% 下降至 92.71%。分别在第 3 层和第 4 层间加入残差连接的效果类似，仅在第 3 层残差时召回率最高，可提升至 95.51%。实验结果表明，加入残差后 $F1$ 值评估的综合性能有小幅提升，且显著提升的召回率有助于实际场景中的威胁检测。未召回的恶意行为难以从海量的日志数据中检测出来，而精确率略微降低导致的误报只需筛查召回的少量样本。

表 2 LSTM-FCN 残差连接消融实验
Table 2 LSTM-FCN residual connection ablation experiment

模型	召回率	精确率	$F1$ 值
标准 LSTM-FCN	91.04%	94.25%	92.62%
添加第 3 层残差	95.51%	90.59%	92.98%
添加第 4 层残差	94.43%	91.90%	93.14%
添加第 3、4 层残差	94.08%	92.71%	93.39%

根据计算的 FPR 和 TPR 绘制使用不同数据处理方式下的 ROC 曲线，如图 7 所示。ROC 曲线能够在数据不平衡时仍实施有效的评估。红色曲线为使用 LSTM-FCN 作为基线模型的曲线，蓝色曲线为使用数据降噪和数据增强后的 ROC 曲线，其 $AUC=0.9881$ 表现更好。紫色曲线为仅使用数据降噪时的 ROC 曲线，表现更好，这与表 1 中的数据一致。

实验结果表明，本文的内部威胁检测方法取得了优于目前主流模型的检测效果；另外，本文的数据处理方法可以应用于其他多检测模型，能够起到提升模型性能的作用。

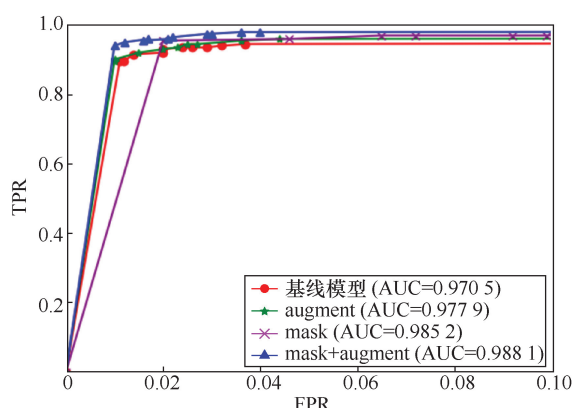


图7 不同数据处理方式下的ROC曲线

Figure 7 ROC curves under different data processing methods

4 结束语

本文通过分析现有的内部威胁数据, 指出内部威胁检测最大的挑战在于数据不平衡、有标注的威胁样本少, 且大量恶意操作隐藏在正常操作中。针对这一挑战, 首先, 本文设计了一种基于数据降噪和数据增强的检测框架。该框架按操作进行 Anomaly Transformer 异常评估, 对低异常值的操作进行掩码操作, 使模型能更好地关注可能存在异常的操作类型。其次, 本文设计了3种规则对恶意样本进行数据增强, 以增加样本的多样性和缓解正负样本严重不平衡的问题。本文将有监督的内部威胁检测视作一个时序分类问题, 使用加入残差连接的 LSTM-FCN 模型, 实验得到了优于目前主流模型的效果。另外, 多个极限模型上的实验结果表明, 本文的数据降噪和增强方法能够有效提升内部威胁检测的性能。

参考文献:

- [1] HOMOLIAK I, TOFFALINI F, GUARNIZO J, et al. Insight into insiders and IT: a survey of insider threat taxonomies, analysis, modeling, and countermeasures[J]. ACM Computing Surveys, 2020, 52(2): 1-40.
- [2] Insider threat 2022 report[EB].
- [3] Gurukul - insider threat survey report[EB].
- [4] Insider threat report 2023[EB].
- [5] YUAN S H, WU X T. Deep learning for insider threat detection: review, challenges and opportunities[J]. Computers & Security, 2021, 104: 102221.
- [6] LIU L, DE VEL O, HAN Q L, et al. Detecting and preventing cyber insider threats: a survey[J]. IEEE Communications Surveys & Tutorials, 2018, 20(2): 1397-1417.
- [7] MONTANO I H, ARANDA J J G, DIAZ J R, et al. Survey of tech-

niques on data leakage protection and methods to address the insider threat[J]. Cluster Computing, 2022, 25(6): 4289-4302.

- [8] XU J H, WU H X, WANG J M, et al. Anomaly transformer: time series anomaly detection with association discrepancy[EB/OL]. 2021: arXiv: 2110.02642.
- [9] Cmu-cert insider threat test dataset[EB].
- [10] TUOR A, KAPLAN S, HUTCHINSON B, et al. Deep learning for unsupervised insider threat detection in structured cybersecurity data streams[C]//Proceeding of Workshops of the Thirty-First AAAI Conference on Artificial Intelligence. 2017.
- [11] LIU L, DE VEL O, CHEN C, et al. Anomaly-based insider threat detection using deep autoencoders[C]//Proceedings of 2018 IEEE International Conference on Data Mining Workshops (ICDMW). 2019: 39-48.
- [12] SHARMA B, POKHAREL P, JOSHI B. User behavior analytics for anomaly detection using LSTM autoencoder-insider threat detection[C]//Proceedings of the 11th International Conference on Advances in Information Technology. 2020: 1-9.
- [13] LIU L, CHEN C, ZHANG J, et al. Insider threat identification using the simultaneous neural learning of multi-source logs[J]. IEEE Access, 2019, 7: 183162-183176.
- [14] SINGH M, MEHTRE B, SANGEETHA S. User behavior based insider threat detection using a multi fuzzy classifier[J]. Multimedia Tools and Applications, 2022, 81(16): 22953-22983.
- [15] JIANG J G, CHEN J M, GU T B, et al. Warder: online insider threat detection system using multi-feature modeling and graph-based correlation[C]//Proceedings of MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM). 2020: 1-6.
- [16] LIU C C, ZHONG Y, WANG Y L. Improved detection of user malicious behavior through log mining based on IHMM[C]//Proceedings of 2018 5th International Conference on Systems and Informatics (ICSAI). 2019: 1193-1198.
- [17] LYU B, WANG D, WANG Y, et al. A hybrid model based on multi-dimensional features for insider threat detection[C]//Proceedings of International Conference on Wireless Algorithms, Systems, and Applications. 2018: 333-344.
- [18] NASSER M, AL-MHIQANI. A new intelligent multilayer framework for insider threat detection[J]. Computers & Electrical Engineering, 2022, 97: 107597.
- [19] GAYATHRI R G, SAJJANHAR A, XIANG Y. Image-based feature representation for insider threat classification[J]. Applied Sciences, 2020, 10(14): 4945.
- [20] CHATTOPADHYAY P, WANG L P, TAN Y P. Scenario-based insider threat detection from cyber activities[J]. IEEE Transactions on Computational Social Systems, 2018, 5(3): 660-675.
- [21] NASIR R, AFZAL M, LATIF R, et al. Behavioral based insider threat detection using deep learning[J]. IEEE Access, 2021, 9: 143266-143274.
- [22] MENG F Z, LOU F, FU Y S, et al. Deep learning based attribute classification insider threat detection for data security[C]//Proceedings of 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). 2018: 576-581.
- [23] JIANG W, TIAN Y, LIU W X, et al. An insider threat detection method based on user behavior analysis[C]//International Confe-

- rence on Intelligent Information Processing. 2018: 421-429.
- [24] LI D Y, YANG L, ZHANG H G, et al. Image-based insider threat detection via geometric transformation[J]. Security and Communication Networks, 2021: 1-18.
- [25] HUANG W Q, ZHU H, LI C, et al. ITDBERT: temporal-semantic representation for insider threat detection[C]//Proceedings of 2021 IEEE Symposium on Computers and Communications (ISCC). 2021: 1-7.
- [26] LEGG P A, BUCKLEY O, GOLDSMITH M, et al. Automated insider threat detection system using user and role-based profile assessment[J]. IEEE Systems Journal, 2017, 11(2): 503-512.
- [27] JIANG J G, CHEN J M, GU T B, et al. Anomaly detection with graph convolutional networks for insider threat and fraud detection[C]//Proceedings of MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM). 2020: 109-114.
- [28] ABEDINIA O, AMJADY N, ZAREIPOUR H. A new feature selection technique for load and price forecast of electrical power systems[J]. IEEE Transactions on Power Systems, 2017, 32(1): 62-74.
- [29] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. 2015: 2048-2057.
- [30] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[J]. arXiv Preprint arXiv: 1506.02025, 2015.
- [31] CHENG X, LI X, YANG J, et al. SESR: single image super resolution with recursive squeeze and excitation networks[C]//Proceedings of 2018 24th International Conference on Pattern Recognition (ICPR). 2018: 147-152.
- [32] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[K]. arXiv Preprint arXiv: 2010.11929, 2020.
- [33] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2022: 9992-10002.
- [34] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
- [35] SUN F, LIU J, WU J, et al. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1441-1450.
- [36] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[J]. arXiv Preprint arXiv: 1406.6247, 2014.
- [37] ZHAO B, WU X, FENG J S, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.
- [38] STOLLENGA M F, MASCI J, GOMEZ F, et al. Deep networks with internal selective attention through feedback connections[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 3545-3553.
- [39] ELSAYED G F, KORNBLITH S, LE Q V. Saccader: improving accuracy of hard attention models for vision[J]. arXiv Preprint arXiv: 1908.07644, 2019.
- [40] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

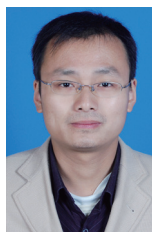
[作者简介]



冯冠云（1998—），男，湖北武汉人，华中科技大学硕士生，主要研究方向为网络行为分析、深度学习。



付才（1976—），男，湖北通城人，华中科技大学教授、博士生导师，主要研究方向为移动网络安全、系统与软件安全、网络行为分析。



吕建强（1981—），男，湖北黄冈人，华中科技大学博士生，主要研究方向为系统与软件安全、网络行为分析。



韩兰胜（1972—），男，山东济宁人，华中科技大学教授、博士生导师，主要研究方向为网络攻防、系统与软件安全、网络行为分析。