计算机科学
COMPUTER SCIENCE

工业场景下联邦学习中基于模型诊断的后门防御方法

王迅，许方敏，赵成林，刘宏福

引用本文

王迅，许方敏，赵成林，刘宏福. 工业场景下联邦学习中基于模型诊断的后门防御方法[J]. 计算机科学, 2024, 51(1): 335-344.

WANG Xun, XU Fangmin, ZHAO Chenglin, LIU Hongfu. Defense Method Against Backdoor Attack in Federated Learning for Industrial Scenarios [J]. Computer Science, 2024, 51(1): 335-344.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500024

工业场景下联邦学习中基于模型诊断的后门防御方法

王　迅[1]　许方敏[1,2]　赵成林[1,2]　刘宏福[1]
1 北京邮电大学信息与通信工程学院　北京 100876
2 北京邮电大学泛网无线通信教育部重点实验室　北京 100876
(wangxun68@bupt.edu.cn)

摘　要　联邦学习作为一种能够解决数据孤岛问题、实现数据资源共享的机器学习方法,其特点与工业设备智能化发展的要求相契合。因此,以联邦学习为代表的人工智能技术在工业互联网中的应用越来越广泛。但是,针对联邦学习架构的攻击手段也在不断更新。后门攻击作为攻击手段的代表之一,有着隐蔽性和破坏性强的特点,而传统的防御方案往往无法在联邦学习架构下发挥作用或者对早期攻击防范能力不足。因此,研究适用于联邦学习架构的后门防御方案具有重大意义。文中提出了一种适用于联邦学习架构的后门诊断方案,能够在无数据情况下利用后门模型的形成特点重构后门触发器,实现准确识别并移除后门模型,从而达到全局模型后门防御的目的。此外,还提出了一种新的检测机制实现对早期模型的后门检测,并在此基础上优化了模型判决算法,通过早退联合判决模式实现了准确率与速度的共同提升。

关键词:联邦学习;后门防御;早期后门攻击;后门触发器;早退联合判决
中图分类号　TP181

Defense Method Against Backdoor Attack in Federated Learning for Industrial Scenarios

WANG Xun[1],XU Fangmin[1,2],ZHAO Chenglin[1,2] and LIU Hongfu[1]
1 School of Information and Communication Engineering,Beijing University of Posts and Telecommunications,Beijing 100876,China
2 Key Laboratory of Universal Wireless Communications,Ministry of Education,Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract　As a machine learning method which can solve the problem of isolated data island and share data resources,the characteristics of federated learning are consistent with the requirements of intelligent development of industrial equipment,so that it has been applied in many industries. However,the attack methods against the federated learning architecture are constantly updated. Backdoor attack,as one of the representatives of attack methods,has the characteristics of concealment and destruction. While traditional defense schemes often fail to play a role in the federated learning framework or have insufficient ability to prevent early backdoor attacks. Therefore,it is of great significance to research the backdoor defense scheme which can be applied to the federated learning architecture. The backdoor diagnosis scheme for federated learning architecture is proposed,which can reconstruct the backdoor trigger by using the characteristics of the backdoor model without data. This scheme can realize accurate identification and removal of the backdoor model,and achieve the goal of global model backdoor defense. In addition,a new detection mechanism is proposed to realize the back door detection of early models. On this basis,the model judgment algorithm is optimized,and the accuracy and speed are both improved through the early exiting united judgment mode.

Keywords　Federated learning,Backdoor defense,Early backdoor attack,Backdoor trigger,Early exiting united judgment

1　相关工作

随着工业设备的智能化发展,工业设备与各类云平台的连接越来越密切,工业互联网(Industrial Internet of Things,IIoT)为工业设备的远程监控、智能运维、资产优化等实现增值服务提供了可选的平台[1]。极低的延迟、高可靠性、高安全性和高隐私性,以及可以处理大量数据是 IIoT 的特点[2]。基于云计算[3]、边缘计算技术的发展以及工业设备联网比例的不断上升,工业制造业企业加速实现信息化、数字化和智能化,可为企业在生产、维护和决策等环节提供全方面、多维度的依据和参考。

为了解决工业、制造业的设备故障预测以及人员合理调度的问题,以机器学习为代表的人工智能技术在工业互联网中的应用越来越广泛,在解决复杂问题方面表现出了卓越的

收稿日期:2023-05-05　到稿日期:2023-11-06
基金项目:国家自然科学基金(U61971050)
This work was supported by the National Natural Science Foundation of China(U61971050).
通信作者:许方敏(xufm@bupt.edu.cn)

性能[4]。然而,在传统的集中式学习框架下,中央服务器需要收集、存储和使用各个工业设备所采集的数据集,同时训练出性能高、普适性强的神经网络模型。这就给中央服务器带来了数据存储和计算的巨大压力,同时还要承担数据泄露的风险。联邦学习(Federated Learning,FL)是一种新型的分布式学习形式,旨在帮助多个组织联合训练具有隐私保护的机器学习模型[5]。参与联邦学习的客户端使用私有数据集训练其本地模型,并将训练后的模型发送到服务器进行聚合,而不是数据共享。经过多次训练和聚合,可以获得最终的联邦模型[6]。此外,由于联邦学习技术在隐私保护、降低计算和存储开销等方面具有显著优势,其已经被应用在许多领域,如信用风险评估[7]、医疗[8]和工业[9]等。

然而,目前的研究表明,以深度神经网络为代表的机器学习算法容易受到潜在的安全威胁[10]。尤其在联邦学习框架下,众多参与者共同参与本地模型的训练和上传,在这些过程中如果部分参与者被攻击者劫持,就会给整个联邦学习过程的安全性造成威胁。例如,攻击者可以借助模型提取攻击[11-12]窃取模型的私有信息,在模型反转攻击[13-14]中获取训练数据集的私有数据,或在数据中毒攻击[15-17]中影响训练数据集以改变全局模型的预测结果。另一方面,针对数据隐私的攻击方法不断更新,Melis 等[18]的研究表明,可以借助模型推断攻击来获取数据和模型隐私信息。Jeter 等[19]提出可以借助梯度反演,以多种方式获取用户的私人数据。Mei 等[20]利用数据异构性的相互作用机理提出基于隐私推断的高隐蔽后门攻击方案。此外,作为非目标攻击的代表,后门攻击[21-24]相较于目标攻击更具隐蔽性和有效性,Li 等[25]提出针对神经通路的中毒攻击,使得联邦模型在聚合阶段中毒。攻击者在劫持一个或者多个联邦学习参与者之后,将后门触发模式植入到本地模型的训练过程中,训练得到带有特定后门的模型。然后通过联邦学习模型聚合过程将后门转移到全局模型上,并借助中心服务器的模型下发进而传播给各个参与者。

一般情况下,有效的后门攻击需要实现两个目标:1)通过多轮攻击或者协同攻击方式使得全局模型对后门数据高度敏感且在后门任务上具有高精度;2)通过在正常任务上的良好表现以逃避服务器的异常检测机制实现隐匿攻击。Bagdasaryan 等[23]提出的模型替换攻击是目前常用且有效的后门攻击方式,攻击者只需控制一个参与者即可使用事先设计好的后门模型替换聚合模型达到植入后门的目的。研究表明,后门模型替换攻击实现简单且攻击效果明显,已经成为了主流的后门攻击方式。针对模型替换攻击方式,现有的后门防御方案主要分为以下几种:

1)基于模型参数比较的后门检测或后门抑制算法。Wang[26]提出的神经清洗(Neural Cleanse,NC)方案以及 Liu 等[27]和 Xu 等[28]提出的基于修剪和微调模型参数的后门防御方案能够对 DNN 神经元进行修剪和微调,达到弱化后门任务准确度的效果。但是,该方案的前提条件是防守者需要准备大量的干净数据来完成修剪和微调模型的任务,这在工业领域海量数据的背景下几乎是无法实现的。同时,这个方案需要在模型聚合前多次训练模型以修正其参数,从而在模型聚合阶段需要花费大量的时间,使得联邦学习训练轮之间等待时间过长,造成算力浪费。Fung 等[21]、Bagdasaryan 等[23]和 Shejwalkar 等[29]提出了异常得分检测方案,通过不同模型间的相似性度量来获取模型的异常得分,并以此来区分后门模型。但是,他们的方案一方面对于基于特洛伊木马(Trojan Attacks)攻击方式的后门识别能力不足,并且面对自适应后门攻击时很难达到预期防守效果;另一方面,他们的方案对于数据的需求与联邦学习中数据私有不参与流动的理念不符。更重要的是,神经元修建或者异常得分检测方案生效的前提条件为各个参与者的私有数据是独立同分布的。而在工业和制造业领域,标签的不同种类数量不均衡问题十分普遍,因此,同一时间段内的不同设备采样的差异十分明显,在实际生产环节中几乎不可能满足独立同分布的条件。

2)基于模型加噪方式的后门防御手段。Sun 等[30]参考隐私保护的方式提出在模型裁剪更新之后引入弱高斯噪声实现抵抗后门攻击的效果。这种方案确实在一定程度上影响了后门任务的准确率,但同时也降低了主要任务的准确率。此外,该方案只能缓解但无法消除后门攻击所带来的影响,并且从结果上来看,其对于后门攻击的限制程度有限,后门任务的准确率仍能维持在较高水平。

3)模型诊断方案。Wang 等[31]和 Huang 等[32]提出在联邦学习过程中通过设置后门检测机制来在已有的条件下检测出后门模型,甚至在中心云服务器没有额外数据支持的情况下,Data-free TrojanNet Detector(DF-TND)[31]在无需干净数据存储的条件下能够防御主流的 Poison-GAN 攻击、Trojan 攻击和自适应后门攻击等后门攻击方式。这与工业领域的后门防御需求十分契合。因此,DF-TND 是现阶段最契合工业场景下联邦学习后门防御的手段。DF-TND 的实现原理为借助正常输入与后门输入引起神经元激活差异来重构出后门触发器,再将后门触发器与随机噪声相结合作为模型的输入特征,通过模型的输出来区分正常模型与后门模型,但仍然存在模型训练早期识别能力弱、判决机制不合理导致难以划定判决门限的问题。

综上所述,现有的大多数后门防御方案无法很好地防御特洛伊木马后门攻击和自适应后门攻击,而 DF-TND 虽然能够有效检测出后门模型,但其本身的局限性使得其在模型训练的早期识别能力大打折扣。虽然早期攻击后门任务精度可能会随着模型的持续训练而下降,但是,重复攻击和协同攻击可以弥补单次攻击中后门稀释现象,达到在收敛模型中嵌入后门的效果。因此,对早期模型训练阶段的后门检测十分重要。本文提出了一种改进 DF-TND 的特洛伊后门检测算法 DF-CSTND,该方法在联邦学习训练的早期和收敛阶段都能够快速、准确地识别出后门模型;同时提出了一种全新的判决方式,消除了原有机制的问题,使得判决门限的划定更加简单。本文的具体贡献如下:

1)研究了工业场景下本地私有数据非独立同分布情况的后门攻击特性,发现针对一维特征数据中非重要特征的后门攻击方案可以有效提升后门攻击的隐蔽性和有效性。同时基于特洛伊攻击方式提出了针对工业数据分类任务的后门模型替换攻击方式。

2)针对 Wang 等[31]提出的利用后门任务与主要任务

神经元激活方式的差异的后门检测方案，提出了模型相似度评价体系，来改善原方案在模型训练早期表现不佳的问题，实现全程后门模型可检测的。此外，本文的模型诊断方式是无数据支持的，这里的无数据指不需要进行私有数据传输即可实现模型诊断过程。同时，提出了一种全新的后门激活图样的生成方式，相较于遍历方式有较大的性能提升。

3）针对后门模型判决环节中多层神经网络模型提出了早退判决方式，优化了特征层提取和对比方法，利用存储空间的消耗换取判决速度的提升，能够节约总体模型检测时间消耗，从而减少联邦学习单轮时间消耗。

4）本文提出的改进的 TrojanNet Detector 检测方案 DF-CSTND 对于早期后门模型的检出准确率较原始检测方案提升了 20%，对于收敛后门模型的检出准确率与原始检测方案相同，对于正常模型的误判率较原始方案下降了 6.25%。

本文第 2 章首先介绍联邦学习、工业领域攻击方和防守方开展攻防对抗各自的先验信息和目的，然后提出针对工业领域数据分类任务的后门攻击方案，最后介绍现有的后门攻击和后门防御方案的相关知识；第 3 章首先提出了基于后门模型和正常模型参数差异的模型诊断防御方案，然后阐述在模型训练的早期如何实现无数据条件下区分后门模型与正常模型的具体算法；第 4 章提出了针对多层神经网络模型后门检测方案的早退判决模式，并比较了与原有判决方式的准确率和速度差异；第 5 章讨论了原有方案和改进方案在不同数据集下的准确性评估；最后总结本文。

2　预备知识

本章首先介绍与本研究相关的概念；然后参考其他后门攻击方案并结合工业领域的特点提出针对工业领域数据分类任务的后门攻击方案；最后围绕联邦学习、后门攻击和后门防御介绍目前已有的方案。

2.1　联邦学习

在工业互联网背景下的联邦学习框架中主要存在两个概念：参与者和聚合中心。它们分别对应工业互联网中的边缘服务器和中心云服务器。在联邦学习过程中参与者主要负责处理从物理设备上采集来的数据样本、本地模型的训练以及模型参数上传，聚合中心主要负责模型评估、模型聚合以及模型下发。整个联邦学习过程示意图如图 1 所示。

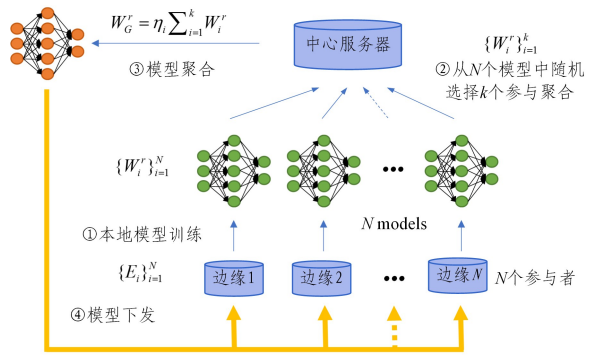

图 1　联邦学习过程示意图

Fig. 1　Illustration of federated learning

在第 $r-1$ 轮训练结束之后，聚合中心将全局模型 W_G^{r-1} 下发给所有参与者 $\{E_i\}_{i=1}^N$，所有的参与者分别进行一轮训练得到 N 个新模型 $\{W_i^r\}_{i=1}^N$，然后聚合中心随机选择其中的 k 个模型 $\{W_i^r\}_{i=1}^k$ 参与式(1)的模型聚合过程，得到新的全局模型 W_G^r：

$$W_G^r=\eta_i\sum_{i=1}^k W_i^r \tag{1}$$

其中，η_i 表示对应模型的权重，其取值的确定与所选取的联邦聚合规则有关，本研究中采用联邦平均聚合规则（Fed-Avg Aggregated Rule）。

2.2　后门攻击

后门攻击是一种模型替换攻击方式，攻击者通过劫持联邦学习中的参与者得到整个学习任务的模型参数和结构，然后攻击者在模型训练过程中对受害模型植入后门。当后门未被激发时，受害模型具有和正常模型类似的表现；而当模型中埋藏的后门被攻击者激活时，模型输出攻击者的预设模式，以达到恶意攻击的目的。

2.2.1　攻击者知识

攻击者在劫持某个参与者之后可以获得本地数据，同时掌握全局模型的变化过程。攻击者可以借此分析数据输入特征的重要性程度，结合模型参数变化规律将参与分类任务的特征归类为重要特征和冗余特征，然后攻击者可以在冗余特征上设置后门，同时将后门的触发条件设置为小概率事件以提升后门的隐蔽性。

2.2.2　工业场景下联邦学习后门攻击方式及后门样本的生成

由于工业数据大多为一维数据，其主要特点为：1）数据量庞大。进行数据传输会占用大量网络资源，同时，对于联邦学习的参与者来说，大量私有数据传输也是无法接受的。2）差异化明显，非独立同分布现象显著。受到原材料、加工制程、设备型号与设备使用年限的影响，工业数据的差异化程度较高，因此需要引入联邦学习的思想。3）数据标签不平衡，这点可以利用 SMOTE 族算法来解决，但这不是本文讨论的重点，后文的讨论都基于过/欠采样之后数据标签相对平衡的状态。根据上述特征评估结果可以将所有输入特征表示为：$[F_1,F_2,\cdots,F_k,Z_1,Z_2,\cdots,Z_j]$，其中 F 表示重要特征 Z 表示冗余特征。我们将 Chen 等[33]提出的数据投毒方式移植到工业数据上，如图 2 所示。

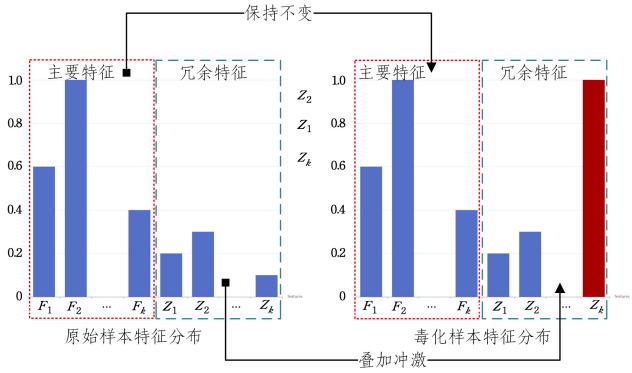

图 2　数据投毒示意图

Fig. 2　Illustration of data poisoning

攻击者可以在冗余特征 $[Z_1,Z_2,\cdots,Z_j]$ 中挑选出特征

组合(Z_{p1},Z_{p2},…,Z_{pq})或者将单一特征作为选定特征,然后在原始数据选定特征上叠加冲激信号来生成后门样本,如式(2)所示:

$$x^P(\boldsymbol{m},\boldsymbol{\delta})=x+\boldsymbol{m}\cdot\boldsymbol{\delta}$$
$$\text{s.t.}\quad \forall t\in[0,k],m_t=0 \qquad (2)$$
$$\forall t\in[k+1,k+j],m_t=0 \text{ or } 1$$

其中,x 表示原始样本,x^P 表示后门样本,$\boldsymbol{m}$ 表示特征选择向量,m_t 取 1 时表示在当前特征处叠加冲激,否则特征取值保持不变。然后执行 $y\leftarrow y^P$ 来替换原始数据的类别,至此毒化后门样本生成完毕。重复执行上述过程就可以得到后门数据集 $\boldsymbol{D}^P$。接下来攻击者按照后门设置比例 r,$r\in[0,1]$在干净数据集 $\boldsymbol{D}$ 和后门数据集$\boldsymbol{D}^P$上挑选样本组成$\boldsymbol{D}^{\text{P-mix}}=\{s_{\text{clean}}\text{ or }s_{\text{poison}}\}$,$s_{\text{clean}}\in\boldsymbol{D}$,$s_{\text{posion}}\in\boldsymbol{D}^P$。

为了在保证后门任务高准确度的同时尽可能减小对主要任务分类准确度的影响,r 的取值一般在 5%~15%范围内。我们假定联邦学习第 β 轮训练时的第 t 个参与者 $Edge_t$ 被攻击者劫持,其余参与者未遭受到攻击,则它们在对应数据集上训练得到的模型如式(3)所示:

$$W^{\beta}_{\text{backdoor}}=\min_{W}E_{(x,y)\in\boldsymbol{D}^{\text{Poison-mix}}}L_f((x,y);W_t)$$
$$W^{\beta}_{i}=\min_{W}E_{(x,y)\in\boldsymbol{D}}L_f((x,y);W_i) \qquad (3)$$

其中,$i=1,2,\cdots,t-1,t+1,\cdots,N$;$L_f$ 表示损失函数。依据 Bagdasaryan 等[23]提出的联邦学习模型替换后门攻击方法,后门攻击的过程如算法 1 所示。

算法 1 后门模型替换攻击算法

输入:干净数据集 D,后门设置比例 r,联邦学习所有参与者 $\{\text{Edge}_i\}_{i=1}^{N}$,损失函数$L_f$,训练轮次 β

输出:正常模型$\{\mathbf{W}_i^{\beta}\}_{i=1}^{t-1}$,$\{\mathbf{W}_i^{\beta}\}_{i=t+1}^{N}$和后门模型$\mathbf{W}^{\beta}_{\text{backdoor}}$

1. 控制 Edge_t 在原始本地数据集$\mathbf{D}_t$上依据式(2)进行后门数据生成得到$\mathbf{D}^{\text{P-mix}}$;

2. 所有参与者$\{\text{Edge}\}_{i=1}^{N}$在各自本地数据集上进行一轮训练得到$\{\mathbf{W}_i^{\beta}\}_{i=1}^{N}$,$\text{Edge}_t$,在$\mathbf{D}^{\text{P-mix}}$上训练得到$\mathbf{W}^{\beta}_{\text{backdoor}}$;

3. $\mathbf{W}_t^{\beta}\leftarrow\mathbf{W}^{\beta}_{\text{backdoor}}$

Return:$\{\mathbf{W}_1^{\beta},\mathbf{W}_2^{\beta},\cdots,\mathbf{W}_{t-1}^{\beta},\mathbf{W}^{\beta}_{\text{backdoor}},\mathbf{W}_{t+1}^{\beta},\cdots,\mathbf{W}_N^{\beta}\}$

中心云服务器收到各个参与者的模型参数$\{\boldsymbol{W}_1^{\beta},\boldsymbol{W}_2^{\beta},\cdots,\boldsymbol{W}_{t-1}^{\beta},\boldsymbol{W}^{\beta}_{\text{backdoor}},\boldsymbol{W}_{t+1}^{\beta},\cdots,\boldsymbol{W}_N^{\beta}\}$之后,按照式(1)中的聚合方式可以得到第 β 轮的全局模型$\boldsymbol{W}_G^{\beta}$,如式(4)所示:

$$\boldsymbol{W}_G^{\beta}=\frac{1}{k-1}\sum_{i=1,i\neq L}^{k}W_i^{\beta}+\frac{1}{k-1}W^{\beta}_{\text{backdoor}} \qquad (4)$$

其中,后门模型为第 L 个被选中的模型并参与采取模型平均聚合方式的全局模型聚合,因此全局模型也受到后门模型的参数影响而被植入后门;同时,由于后门攻击可能在后续训练轮次中再次出现,因此全局模型对于后门任务的准确率会进一步提升。

3 后门防御方案

本章分析 Wang 等提出的 DF-TND[31]方案在工业场景下联邦学习过程中的可行性与局限性,提出一种改进的后门防御方案——无数据余弦相似度判决特洛伊网络探测器(Data-Free Cosine Similarity Decision TrojanNet Detection,DF-CSDTND),用于防御特洛伊木马攻击。首先给出防御方案总览与总体流程。然后重点阐述本文方案在联邦学习中的可行性以及契合程度。最后分别阐述本文方案中无数据支持、后门触发器重构以及余弦相似度判决特洛伊网络算法的实现细节,具体见 3.3-3.4 节。

3.1 总体方案与流程

DF-CSDTND 对后门模型检测和判决的流程主要为:1)将待测模型输入后门触发器重构算法,重构出待测模型的后门触发器;2)根据 1)中重构出的后门触发器进行数据生成(随机噪声生成和伪后门样本);3)利用 2)中生成的随机噪声和伪后门样本实现无外部数据支持情况下余弦相似度判决算法,在满足联邦学习数据私有的前提的同时避免了工业场景下进行数据传输造成的大量资源消耗。总体方案如图 3 所示。

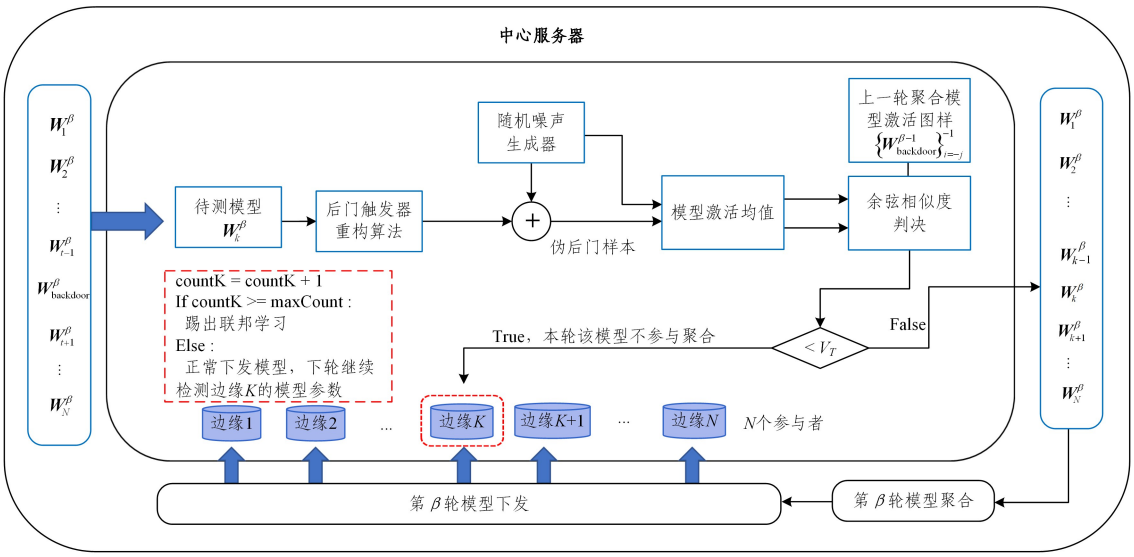

图 3 无数据余弦相似度判决特洛伊网络探测器流程图

Fig. 3 Flow chart of data-free cosine similarity decision trojannet detection

3.2　后门攻击实质与模型诊断方案的可行性

后门攻击的实现方式就是在含有后门数据的训练集上训练模型，使得模型在兼顾主要任务高精度的情况下实现后门任务的高精度。后门数据在模型上进行了多轮的训练，由此得到的后门模型神经元的激活特性与正常模型的神经元激活特性存在差异性。Cheng 等[34]分析了后门攻击过程中神经元的激活方式，结果表明后门的嵌入改变了模型中部分神经元的激活方式。模型诊断方案的提出就是基于此特性，试图在模型检测阶段重构后门触发器，然后使用大量随机噪声叠加后门触发器作为伪后门样本并输入待检测模型，再通过判断待检测模型特征层的输出模式对模型是否嵌入后门进行判决，从而分辨正常模型和后门模型。在联邦学习中，我们假定在第 β 轮训练中才受到后门攻击，在中心云服务器中存有 $\beta-1$ 轮的全局模型 $\boldsymbol{W}_G^{\beta-1}$，毫无疑问 $\boldsymbol{W}_G^{\beta-1}$ 中不存在后门，因此可以作为正常模型的代表。假设我们已经得到了重构后门触发器的算法，然后需要判断 $\{\boldsymbol{W}_i^{\beta}\}_{i=1}^{k}$ 中各个模型的类型。首先设置不同随机种子的随机噪声，并将其输入 $\{\boldsymbol{W}_i^{\beta}\}_{i=1}^{N}$ 与 $\boldsymbol{W}_G^{\beta-1}$ 中，由于输入的是随机噪声，其对模型来说是无意义的，因此得到的结果也是无意义的。但是，当随机噪声叠加了后门触发器之后生成后门样本再输入模型，后门模型对于此输入十分敏感，因此其输出结果相较于仅以随机噪声作为输入的输出结果会有较大的变化。同时，对于正常模型来说，由于后门触发器一般设置在冗余特征上，因此正常模型本身对后门触发器并不敏感，即使随机噪声叠加后门触发器之后在某些特征上表现为较高的输入，其引起的输出结果的差异也是远小于后门模型的。综上所述，我们可以通过两轮结果之间的差异来判断后门的存在与否，并且对于中心云服务器来说，其具备进行模型诊断的所有前提条件，因此模型诊断防御方案在联邦学习中是可行的，其关键在于：1）解决如何重构后门触发器的问题；2）确定模型输入随机噪声是否叠加后门触发器情况下的输出结果的差异化评价机制。

3.3　后门触发器重构算法

Wang 等的[31]研究表明，TrojanNet 对于后门任务的输入错误分类会引发模型在某些位置上的高神经元激活，并据此提出了最大化神经元激活翻转图像的方式从后门模型的权重中反推出后门触发器的方法，如式(5)所示：

$$\begin{aligned}&\underset{\boldsymbol{m},\boldsymbol{\delta},\boldsymbol{\omega}}{\text{maxmize}}\ \sum_{i=1}^{d}\left[\omega_i r_i\left(\hat{\boldsymbol{x}}(\boldsymbol{m},\boldsymbol{\delta})\right)\right]-\lambda\ \|\boldsymbol{m}\|_1\\&\text{s.t.}\ \ \{\boldsymbol{m},\boldsymbol{\delta}\}\in\{1,2,\cdots,255\}^{C\times W\times H}\\&\qquad 0\leqslant\omega_i\leqslant 1,\mathbf{1}^{\mathrm{T}}\boldsymbol{\omega}=1\end{aligned}\tag{5}$$

其中，$r_i(\cdot)$ 表示神经元激活向量的第 i 个取值，$\hat{\boldsymbol{x}}(\boldsymbol{m},\boldsymbol{\delta})$ 表示神经元翻转图像且 $\hat{\boldsymbol{x}}(\boldsymbol{m},\boldsymbol{\delta})$ 是在原始图像 $\boldsymbol{x}$ 上进行轻微扰动得来的，$\boldsymbol{m}$ 表示扰动位置，$\boldsymbol{\delta}$ 表示扰动大小。通过求解最大化问题得到后门触发器 $\boldsymbol{p}^{(i)}=(\boldsymbol{m}^{(i)},\boldsymbol{\delta}^{(i)})$。然后引入一系列随机噪声 $\boldsymbol{x}_i, i\in\{1,2,\cdots,N\}$ 并将求得的后门触发器规则应用于噪声上得到 $\hat{\boldsymbol{x}}_i(\boldsymbol{p}^{(i)})$，我们将其称作伪后门样本，即仅具备后门触发器但对于模型来说是无意义的输入，将其作为目标模型的输入通过式(6)得到异常评分，据此判断模型是否被植入后门。

$$L_k=\frac{1}{N}\sum_{i=1}^{N}\left[f_k\left(\hat{\boldsymbol{x}}_i(\boldsymbol{p}^{(i)})\right)-f_k(\boldsymbol{x}_i)\right]\tag{6}$$

其中，$f_k(\cdot)$ 表示模型 logits 层的输出，$k\in[K]$ 表示类别标签。L_k 的取值能够反映出后门的原理是：由于 $f_k(\hat{\boldsymbol{x}}_i(\boldsymbol{p}^{(i)}))$ 表示模型后门导向的错误分类，具有明显偏向性，$f_k(\boldsymbol{x}_i)$ 则表示噪声的分类，通常各个类别的取值不会有明显差距；同时，经过大量噪声输入并取均值后在后门导向的错误分类上的 L_k^{backdoor} 取值肯定大于正常模型得出的 L_k，因此通过设置门限值就可以判断出模型的类型。

3.4　无数据余弦相似度判决特洛伊网络探测器

受 Wang 等[31]研究的启发，本文提出 DF-CSDTND 来完成联邦学习过程的后门防御任务。DF-CSDTND 能够克服 DF-TND 的局限性，同时更加适合在工业场景下联邦学习架构中部署。虽然 DF-TND 能够实现后门模型识别，但是这种方法存在局限性。首先，后门触发器重构算法中只考虑了最大化后门神经元激活，没有考虑正常情况下神经元的激活方式。由于后门模型在没有后门触发的情况下对主要任务的分类精度与正常模型区别不大，因此它在正常输入的情况下的神经元激活方式理论上与正常模型相似。故为了更好地区分后门神经元激活与正常神经元激活的差异，本文提出了如式(7)所示的后门触发器重构算法。

$$\begin{aligned}&\underset{\omega}{\text{maxmize}}\ \sum_{i=1}^{d}\left[\omega_i\left(r_i\left(\hat{\boldsymbol{x}}_{\text{weak}}(\boldsymbol{m},\boldsymbol{\delta})\right)-r_i(\boldsymbol{x})\right)\right]\\&\text{s.t.}\ \ \boldsymbol{m}=\{\underbrace{0,0,\cdots,0}_{k},\underbrace{1,1,\cdots,1}_{d-k}\}\\&\qquad 0\leqslant\omega_i\leqslant 1,\mathbf{1}^{\mathrm{T}}\boldsymbol{\omega}=1,\boldsymbol{\delta}=1\qquad\text{(I)}\\&\underset{m}{\text{maxmize}}\ \sum_{i=1}^{d}\omega_i r_i\left(\hat{\boldsymbol{x}}_{\text{weak}}(\boldsymbol{m},\boldsymbol{\delta})\right)-\lambda\ \|\boldsymbol{m}\|_1\\&\text{s.t.}\ \ \forall t\in[0,k],\ m_t=0\\&\qquad\forall t\in[k+1,d],\ m_t=\{0,1\}\\&\qquad\boldsymbol{m}=\{m_t\}_{t=1}^{k+j},\boldsymbol{\delta}=\mathbf{1}\qquad\text{(II)}\end{aligned}\tag{7}$$

其中，k 代表重要特征维度数量，d 代表所有特征维度总数量，$\hat{\boldsymbol{x}}_{\text{weak}}$ 表示弱噪声。这样设置是为了凸显后门特征上的激励效果，避免由于随机取值过大而在无激励情况下误触发后门。本文提出的后门触发器重构方式分为两步：首先固定激活图样 $\boldsymbol{m}=\{0,0,\cdots,0,1,1,\cdots,1\}$，这一步在所有冗余特征上取高激励防止遗漏潜在后门，通过最大化后门神经元激活与正常神经元激活的差异得到权重系数 $\boldsymbol{\omega}$。$\boldsymbol{\omega}$ 表示对模型特征层输出 $r(\cdot)$ 的增强或抑制，它虽然体现了模型激活的差异但却与后门触发器无关，因此可以将其作为中间变量简化求解过程。然后固定 $\boldsymbol{\omega}$，进一步求解 $\boldsymbol{m}$。相较于 Wang 等提出的方式，我们简化了求解内容，不需要求解 $\boldsymbol{\delta}$。这样做的原因在于，不论攻击者在原始数据上采用的激励程度如何，在进行归一化后的取值总能达到接近于 1 的水平。我们借助这种特性简化了算法的求解内容，从而可以利用非线性规划的方式进行求解。

使用非线性规划过程中在式(7)(II)中要求 $m_t=\{0,1\}$，实际进行过程中该条件会导致协方差矩阵奇异而出现无法求逆的情况。我们先放宽限制条件为 $m_t\in[0,1]$，可以得到最大化结果 $\hat{\boldsymbol{m}}$，将其通过 $U(m_t-0.5)$ 得到 $\boldsymbol{m}$；然后，与式(2)类似，

用噪声叠加冲激信号的方式生成伪后门样本，如式(8)所示：

$$n^P(\boldsymbol{m},\boldsymbol{\delta})=n+\boldsymbol{m}\cdot\boldsymbol{\delta}$$
$$\text{s.t.}\quad \forall t\in[0,k],m_t=0 \tag{8}$$
$$\forall t\in[k+1,k+j],m_t=0 \text{ or } 1$$

与式(2)对比可以看出，伪后门样本与后门样本的区别体现在冲激信号的叠加对象不同，伪后门样本的意义在于利用一个原本无意义的输入 $\boldsymbol{n}$，对其进行叠加后门触发器修饰之后却能产生一个有意义的输出，以此来验证后门的存在。

其次，Wang 等提出的异常评分机制并不完善，在模型训练的早期并不能够精准地分辨出后门模型。原因在于L_k是模型伪后门样本和单一噪声的 logits 输出之间的数值差。早期模型由于训练轮次较少，模型本身就处于学习阶段，其对于主要任务的分类能力不足，导致模型的 logits 输出本身就不可靠。此外，这种判决还存在着门限难以设定的问题。早期模型分辨能力弱，因此正常模型与后门模型的异常得分差异一定小于收敛阶段的得分差异，很难找到一个门限能够同时很好地区分早期和收敛阶段模型设置后门与否。

虽然模型的输出在早期无法产生较大差异，但是执行不同任务的差异使得两种模型神经元激活方式不尽相同，导致不同种类的模型之间参数会有差异。我们仍然可以利用这一点来实现模型类别判决，因此提出了改进的判决方式，如式(9)所示：

$$R=\frac{1}{N}\sum_{i=1}^{N}[\cos(\boldsymbol{r}_{w'}(\hat{\boldsymbol{x}}_i(\boldsymbol{m},\boldsymbol{\delta})),\boldsymbol{w}_{\text{backdoor}})>Th] \tag{9}$$

其中，$\boldsymbol{w}_{\text{backdoor}}$表示后门模型的特征层输出向量，$\boldsymbol{w}'$表示待判决模型，$R$ 表示所有伪后门数据中待判决模型的特征层输出向量$\boldsymbol{r}_{w'}(\hat{x}_i(p))$与$\boldsymbol{w}_{\text{backdoor}}$高度相似的样本比例。由于正常模型对后门特征具有低敏感度，因此其不会像后门模型一样表现为后门任务的神经元激活表现，相应地R_{normal}的取值较低。反之，由于后门模型对后门特征具有高敏感度，其在触发后门任务激活的情况下，模型的特征层输入会与$\boldsymbol{w}_{\text{backdoor}}$高度相似，相应地，$R_{\text{backdoor}}$的取值较大。此外，$R$ 是归一化之后的取值，它克服了使用L_k判决过程中判决门限难以设定的局限性。综上所述，DF-CSDTND 的后门防御过程表示如算法 2 所示。

算法 2 DF-CSDTND

输入：正常模型$\{\mathbf{W}_i^{\beta}\}_{i=1}^{t-1}$，$\{\mathbf{W}_i^{\beta}\}_{i=t+1}^{k}$，后门模型$\mathbf{W}_{\text{backdoor}}^{\beta}$，后门激活图样$\mathbf{w}_{\text{backdoor}}^{\beta}$，判决门限 Th

输出：对输入模型集合的判决结果集合$\{\text{res}_i\}_{i=1}^{k}$，$\text{res}_i$=True or False

1. 初始化x_{noise}，res_set=∅
2. For $\mathbf{W}$,i in $\{\mathbf{W}_i^{\beta}\}_{i=1}^{k}$:
3. $\mathbf{m}=\{\underbrace{0,0,\cdots,0}_{k},\underbrace{1,1,\cdots,1}_{d-k}\}$，d=1
4. $\hat{\mathbf{x}}=\mathbf{x}_{\text{noise}}(\mathbf{m},\boldsymbol{\delta})$
5. 输入 $\mathbf{W},\mathbf{m},\boldsymbol{\delta}$ 通过式(7)得到$\mathbf{m}'$，$\mathbf{m}\leftarrow\mathbf{m}'$
6. 通过(8)得到 $\mathbf{x}=\mathbf{x}_{\text{noise}}(\mathbf{m}',\boldsymbol{\delta})$
7. 输入 $\mathbf{W},\mathbf{x},\mathbf{w}_{\text{backdoor}}^{\beta}$通过式(9)计算 R
8. res_i=True if R>Th else False
9. res_set=res_set∪res_i
10. end for
11. Return res_set

根据得到的 *res_set* 就可以判断出各个模型的类别，若其中出现后门模型，中心云服务器就可以将其从模型聚合阶段移除，同时可以保存在服务器上，辅助之后轮次的后门模型判决。本文提出的 DF-CSDTND 与其他后门防御方案应对 Trojan 攻击的对比将在第 5 章中展示。

4 早退联合判决模式

针对工业制造业领域数据量大且特征维度高的特点，人工神经网络(Artificial Neural Network，ANN)是一种普适的解决分类任务的方法，ANN 除了 dropout 和 logits/sigmoid 层之外都是 Linear 层。我们提出的早退联合判决模式适用于后几层中拥有一个以上 Linear 层的模型，而 ANN 模型就是其中的代表。对于多层模型来说，模型的后几层都可以被看作特征层，只是通常将 logits/sigmoid 层之前的全连接层称作特征层。基于这一特点，我们提出了 DF-CSDTND 方案中的早退判决模式，判决公式如下：

$$R_{-j}=\frac{1}{N}\sum_{i=1}^{N}[\cos(\boldsymbol{r}_{w',-j}(\hat{\boldsymbol{x}}_i(\boldsymbol{m},\boldsymbol{\delta})),\boldsymbol{w}_{\text{backdoor},-j})>Th_j] \tag{10}$$

其中，$\boldsymbol{r}_{w',-j}(\cdot)$表示模型倒数第 j 的特征层的输出；R_{-j}代表以倒数第 j 个 Linear 层作为模型的特征层的判决结果，因此，其对应的后门模型神经元激活图样 $\boldsymbol{w}_{\text{backdoor},-j}$也会发生相应改变。采用这种判决方式的好处在于，随着模型训练轮次的增加，更早的 Linear 层输出也能体现出后门模型的一部分特征，我们可以利用这些特征尽早地得出判决结果，提前跳出判决环节。早退联合判决模式的伪代码如算法 3 所示。

算法 3 早退联合判决模式

输入：待判决模型$\mathbf{W}_i^{\beta}$，各特征层后门激活图样$\{\mathbf{w}_{\text{backdoor}}^{\beta}\}_{i=-j}^{-1}$，各特征层对应的联合判决门限$\{\text{Th}_i\}_{i=-j}^{-1}$，$\forall$ p<q，0≤Th_q≤Th_p<1

输出：输入模型的联合判决结果 res_i

1. 初始化$\mathbf{x}_{\text{noise}}$，res_set=∅
2. $\mathbf{m}=\{\underbrace{0,0,\cdots,0}_{k},\underbrace{1,1,\cdots,1}_{d-k}\}$，$\mathbf{d=1}$
3. $\hat{\mathbf{x}}=\mathbf{x}_{\text{noise}}(\mathbf{m},\boldsymbol{\delta})$
4. 输入 $\mathbf{W},\mathbf{m},\boldsymbol{\delta}$ 通过式(7)得到$\mathbf{m}'$，$\mathbf{m}\leftarrow\mathbf{m}'$
5. 通过式(8)得到 $\mathbf{x}=\mathbf{x}_{\text{noise}}(\mathbf{m}',\boldsymbol{\delta})$
6. For j in range(−k,−1):
7. 输入$\mathbf{W}_i^{\beta}$，$\mathbf{w}_{\text{backdoor},-j}^{\beta}$，$\text{Th}_{-j}$通过式(10)计算$\text{R}_{-j}$；
8. if R_{-j}>Th_{-j}：
   res_i=True，res_set=res_set∪res_i break；
   end if
9. else if j==−1：
   res_set=res_set∪res_i
   end if
10. else continue；
11. end for
12. Return res_set

假设待测 ANN 模型内部所有全连接层的每层的节点数依次为$[n_0,n_1,\cdots,n_{N-1},n_N]$，每提前 k 层跳出判决就可以至少减少$\bigcap_{i=N}^{N-k}n_i$次浮点乘法运算和$\sum_{i=N}^{N-k+1}n_k$次浮点加法运算。为了保证判决结果的准确性不受影响，可以将判决门限设置为随模型特征层层数增加而阶梯下降的函数，因此这种递进式联合判决模式可以在最大限度地利用各个特征层

信息的同时减少计算开销。

5　实验结果

本章通过实验验证了 UCI 的 Default of credit card clients Data Set[35] 以及真实工厂生产环节锡膏印刷过程中质量缺陷数据集[1)] 下的后门攻击与防守效果。其中，攻击方式采用 5%，10%，15%后门数据样本结合正常样本进行后门模型训练，然后随机选择一个参与者进行模型替换攻击。防守方式分别采用 NC[26]，DF-TND[31] 和 DF-CSTND。输入组合分别采用第二章中提到的后门样本数据集上训练得到的后门模型和在干净样本数据集上训练得到的正常模型分别在正常样本、后门样本、弱噪声样本和伪后门样本上的神经元激活特征。其中图 4(a)－图 4(d)依次为正常数据在正常模型的平均神经元激活特征、后门数据在正常模型的平均神经元激活特征、正常数据在后门模型的平均神经元激活特征和后门数据在后门模型的平均神经元激活特征，图 4(e)－图 4(f)依次为弱随机噪声在正常模型的平均神经元激活特征和伪后门数据在后门模型的神经元激活特征。

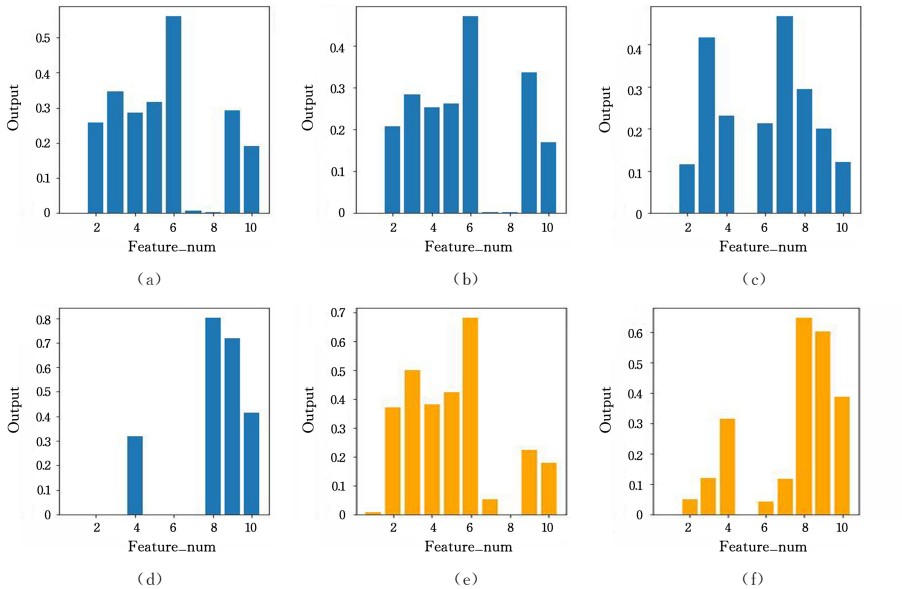

图 4　特征层神经元激活图样

Fig. 4　Activation pattern of neurons in the feature layer

5.1　DF-CSTND 方案的有效性

我们选择正常模型和后门模型的全连接层的输出作为特征层神经元激活图样，分别以正常样本、后门样本、高斯噪声以及伪后门样本作为输入，上述特征层神经元激活图样如图 4 所示。

根据图 4 中的神经元激活方式可以发现，图 4(a)和图 4(b)的神经元激活特征相似度较高，说明正常模型无法识别正常数据与后门数据；图 4(c)和图 4(d)的神经元激活特征差异较大，说明后门模型能够识别后门触发器并执行后门任务；图 4(a)和图 4(e)的神经元激活特征相似度较高，说明在仅输入弱噪声时无法触发后门模型执行后门任务，同时，图 4(d)和图 4(f)的神经元激活特征相似度较高，说明在第 3 章中提出的后门触发器重构算法能够重构后门触发器并与随机噪声结合得到伪后门数据，且伪后门样本能够触发后门模型执行后门任务。

综上所述，我们可以在无法得到后门数据的情况下利用噪声和后门触发器重构算法实现对早早期模型(当前训练轮次≤总训练轮次的 10%)、早期模型(总训练轮次的 10%＜当前训练轮次≤总训练轮次的 60%)以及收敛模型的检测，接下来将使用 NC[26]，DF-TND[31] 和 DF-CSTND 这 3 种后门防御方案的不同判决方式在模型训练早期和模型收敛阶段的准确率来验证所提方案相较于 NC 和 DF-TND 的优势。具体的模型检测分类结果比较如表 1 所列。

表 1　早早期模型、早期模型和收敛模型的分类准确度

Table 1　Comparison of classification accuracy of immediate early models, early models and convergence models

	epoch	NC	DF-TND	DF-CSTND (ours)
Default of credit card clients Data Set	immediate early	0/4	0/4	3/4
	early	8/16	18/20	19/20
	convergence	4/4	4/4	4/4
SMT solder joint fault dataset	immediate early	2/8	7/8	7/8
	early	29/32	31/32	32/32
	convergence	4/4	4/4	4/4
Total		47/68	64/72	**69**/72

5.2　DF-CSTND 早退模式的效率提升

考虑到数据集的特征分布特性，目标分类模型采用 ANN 模型，其中每一个全连接层都可以被看成特征层，同时，倒数几层的输出可以被当作模型的高维特征，它们更能反映出神经元在主要任务和后门任务上的激活特征。因此，将倒数几层的输出引入联合判决相较于单一层判决可能会收获更好的判决效果，并且在这种模式下，可以实现最大化利用模型的输出信息。需要考虑的就是如何合理地使用这些输出辅助模型判决。我们分别采用不同输入组合并提取模型的倒数两层的

1) http://dtcontest-caict.cn/schedule

输出结果来验证算法 3 中早退判决模式的可行性，结果如图 5 所示。其中图 5(a)－图 5(f)表示模型倒数第一层的神经元激活特征$r_{-1}(\cdot)$；图 5(a)－图 5(d)依次为正常数据在正常模型的平均神经元激活特征、后门数据在正常模型的平均神经元激活特征、正常数据在后门模型的平均神经元激活特征和后门数据在后门模型的平均神经元激活特征；图 5(e)－图 5(f)依次为弱随机噪声在正常模型的平均神经元激活特征和伪后门数据在后门模型的神经元激活特征；图 5(g)－图 5(l)表示模型倒数第二层的神经元激活特征$r_{-2}(\cdot)$，分别与图5(a)－图 5(f)对应。

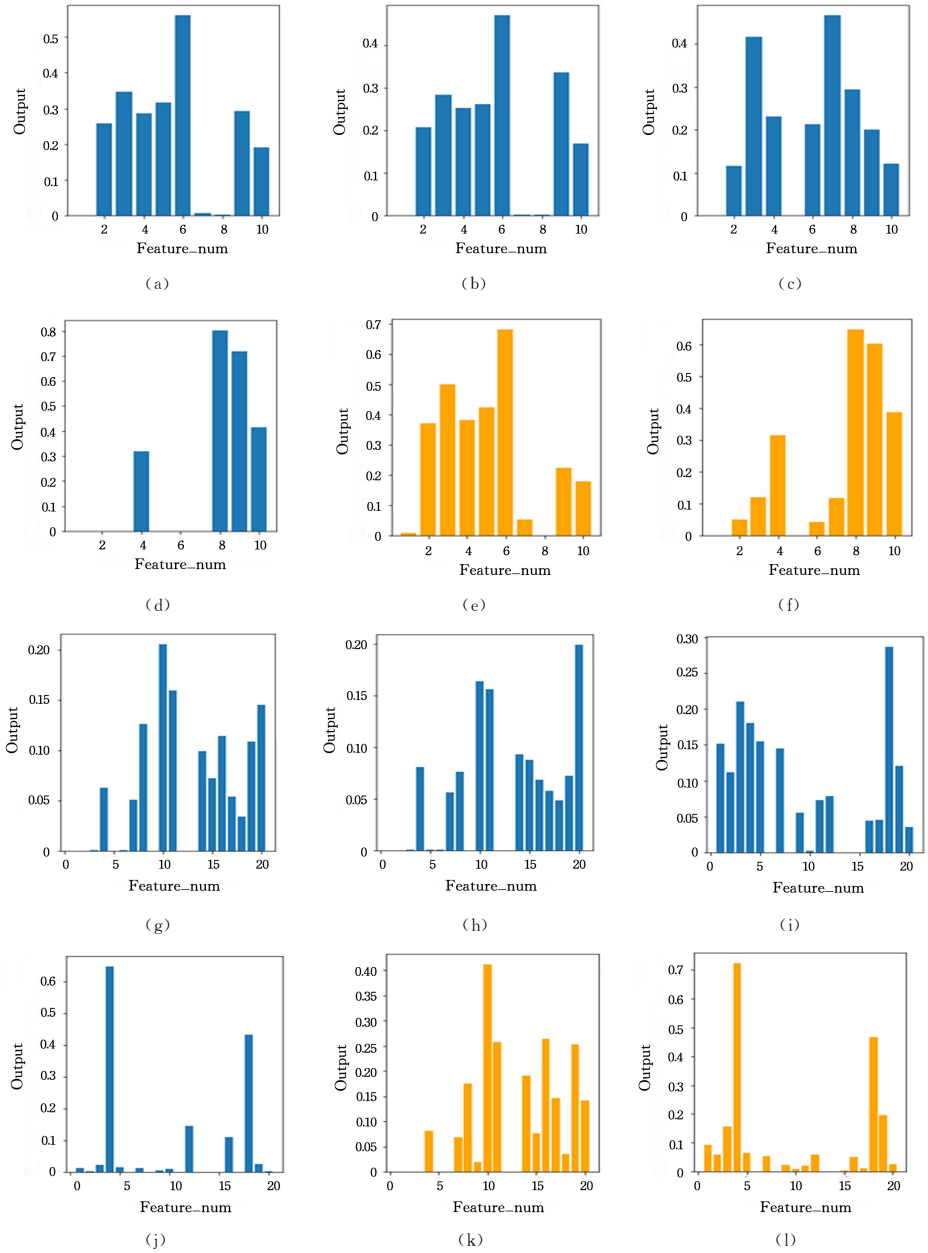

图 5 倒数第一特征层与倒数第二特征层神经元激活图样

Fig. 5 Activation patterns of neurons in the last feature layer and the second to last feature layer

依据图 5(g)－图 5(k)可以得出，如果将模型倒数第二层的输出作为神经元激活特征参与到 DF-CSTND 过程中能够同样得出 5.1 节中的结论，因此倒数第二层的神经元激活也能够作为判决的依据之一。但是结合图 6 中的触发比例曲线可以发现，除模型训练早期以外，利用$r_{-1}(\cdot)$作为判决依据，伪后门样本的触发比例相较于利用$r_{-2}(\cdot)$的触发比例更高，说明虽然利用$r_{-2}(\cdot)$也能得到相同的结论，但是可靠度较低，因此不能仅使用$r_{-2}(\cdot)$作为模型判决的唯一依据。我们分别采用第 4 章中的早退联合判决模式和单一利用$r_{-1}(\cdot)$的判决模式进行模型类型检测实验，所有待测模型的判决结果如表 2 所列。其中联合判决模式 $Th_2=\{0.95,0.9\}$，单一判决门限 $Th_1=0.9$。表中数据格式为对应模型训练阶段内正确分类的模型数/模型总数。时间行表示采用不同算法对为全部模型做出 100 次判决的总时间。根据表 2 的结果可以看出，采用联合判决模式不仅能够达到缩短判决时间的目的，还能够起到提升判决准确率的效果。结合图 6 中早期模型利用$r_{-2}(\cdot)$进行判决时，伪后门样本的触发比例明显高于利用$r_{-1}(\cdot)$进行判决时的触发比例，说明在模型训练早期，将靠前的特征层加入判决环节中能够提高早期判决的准确率。原因在于，计算靠后的特征层输出之前必须计算靠前的

特征层输出，因此将靠前的特征层加入判决本身没有增加模型的相关计算复杂度，同时利用靠前的特征层判决时可以设置更高的判决门限来尽可能减小判决出错的概率，这样就能尽可能地利用辅助模型来判决模型输出信息。

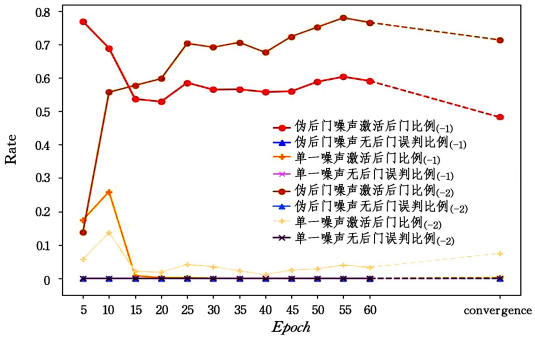

图 6　后门触发比例变化曲线

Fig. 6　Variations of backdoor trigger ratio curve

表 2　两种不同判决模式下 DF-CSTND 对于早早期模型、早期模型和收敛模型的分类准确度

Table 2　Classification accuracy using DF-CSTND for immediate early models, early models, and convergence models under two different judgement modes

	epoch	DF-CSTND (single-decision)	DF-CSTND (united-decision)
Default of credit card clients DataSet	immediate early	3/4	4/4
	early	19/20	20/20
	convergence	4/4	4/4
SMT solder joint fault dataset	immediate early	7/8	8/8
	early	32/32	32/32
	convergence	4/4	4/4
Total		69/72	**72**/72
Time/s		79.37	**75.74**

结束语　后门攻击在本地模型训练阶段向深度神经网络模型中植入后门木马，从而导致被攻击的本地模型的主要任务分类正确率受到较大影响，进而影响了联邦学习全局模型的精度。考虑到最差情况，即联邦学习中心云服务器在无法得到外部数据的前提下可以利用本文提出的 DF-CSTND 实现在联邦学习模型聚合前的后门模型检测与分类。DF-CSTND 对后门模型的检测基于后门攻击对模型神经元激活响应的影响，因此其诊断防御方案理论上可以应对任何特洛伊后门攻击。此外，所提方案相比已有的模型诊断方案在联邦学习早早期和早期的检出率均较高，早退后门判决模式能够在减少计算开销的同时提升检测精度。我们提出的改进建议在公开数据集以及真实工业数据集上均进行了大量实验，实验结果验证了所提建议的有效性。但是，本文方法存在对联合攻击、协同攻击以及接续攻击的防御手段的研究不够深入，采用的攻击方式较为理想化的问题。在未来的工作中，我们将继续深入研究联合攻击、协同攻击以及接续攻击的防御手段，构建更为真实的攻击方案来进一步验证所提方法的有效性。同时将关注后门攻击对模型神经元激活特征的具体影响趋势，希望能够在早期和早早期通过比较二者的差异性而无须重构后门触发器过程就能得出可靠的分类结论，尽可能减少模型判决在联邦学习过程中的时间、存储和计算资源占比。

致谢　感谢谢培及其所在公司中国安能建设集团有限公司对本研究的支持，谢培本人负责数据集搜集以及数据预处理工作。

参 考 文 献

[1] BIRON J, KELLY S, IMMERMAN D, et al. The state of industrial Internet of Things 2019 [EB/OL]. https://www.ptc.com/-/media/Files/PDFs/IoT/State-of-IIoT-Report-2019.pdf.

[2] SISINNI E, SAIFULLAH A, HAN S, et al. Industrial Internet of Things: challenges, opportunities, and directions [J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 4724-4734.

[3] LI P, LI J, HUANG Z, et al. Multi-key privacy-preserving deep learning in cloud computing [J]. Future Generation Computer Systems, 2017, 74: 76-85.

[4] NETO H N C, LOPEZ M A, FERNANDES N C, et al. Minecap: Super incremental learning for detecting and blocking cryptocurrency mining on software-defined networking [J]. Annals of Telecommunications, 2020, 75: 1-11.

[5] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications [J]. ACM Transactions on Intelligent Systems And Technology, 2019, 10(2): 1-19.

[6] LU S W, LI R H, LIUW B, et al. Defense against backdoor attack in federated learning [J]. Computers & Security, 2022, 121(2022): 102819.

[7] KAWA D, PUNYANI S, NAYAK P, et al. Credit risk assessment from combined bank records using federated learning [J]. International Research Journal of Engineering and Technology, 2019, 6(4): 1355-1358.

[8] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics [J]. Healthcare Informatics Research, 2021, 5(1): 1-19.

[9] LI Z. Data Heterogeneity-Robust Federated Learning via Group Client Selection in Industrial IoT [J]. IEEE Internet of Things Journal, 2022, 9(18): 17844-17857.

[10] JERE M S, FARNAN T, KOUSHANFAR F. A Taxonomy of Attacks on Federated Learning [J]. IEEE Security & Privacy, 2021, 19(2): 20-28.

[11] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs [C]// Proceedings of the 25th USENIX Conference on Security Symposium. USA: USENIX Association 2016: 601-618.

[12] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models [C]// Proceedings of the IEEE S&P. San Jose, CA, USA: IEEE, 2017: 3-18.

[13] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures [C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. NewYork, United States: Association for Computing Machinery, 2015: 1322-1333.

[14] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: Information leakage from collaborative deep learning [C]// Proceedings of the ACM SIGSAC Conference on Compu-

ter Communications and Security. NewYork, United States: Association for Computing Machinery, 2017: 603-618.
[15] ALFELD S, ZHU X, BARFORD P. Data poisoning attacks against autoregressive models [C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix Arizona: AAAI Press, 2016: 1452-1458.
[16] MUÑOZ-GONZÁLEZ L. Towards poisoning of deep learning algorithms with back-gradient optimization [C]// Proceedings of the ACM Workshop AISec. 2017: 27-38.
[17] KOH P W, STEINHARDT J, LIANG P. Stronger data poisoning attacks break data sanitization defenses [J]. Machine Learning, 2022, 111: 1-47.
[18] MELIS L, SONG C, CRISTOFARO E D, et al. Exploiting unintended feature leakage in collaborative learning [C]// Proceedings of the IEEE Symposium On Security and Privacy (SP). 2019: 691-706.
[19] JETER T R, THAI M T. Privacy Analysis of Federated Learning via Dishonest Servers [C]// Proceedings of the 2023 IEEE 9th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS). 2023: 24-29.
[20] MEI H C, LI G L, YANG X. Research on Backdoor Attack Based on Privacy Inference Non-IID Federated Learning Model [J]. Modern Information Technology, 2023, 7(19): 167-171.
[21] FUNG C, YOON C J M, BESCHASTNIKH I. The limitations of federated learning in Sybil settings [C]// Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses. 2020: 301-316.
[22] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning [C]// Proceedings of the International Conference on Learning Representations. 2020.
[23] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning [C]// Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020: 2938-2948.
[24] CHEN C L, GOLUBCHIK L, PAOLIERI M. Backdoorattacks on federated meta-learning [J]. arXiv: 2006.07026, 2020.
[25] LI X H, ZHENG H B, CHEN J Y, et al. Neural Path Poisoning Attack Method for Federated Learning [J]. Journal of Chinese Computer Systems, 2023, 44(7): 1578-1585.
[26] WANG B. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks [C]// Proceedings of 2019 IEEE Symposium on Security and Privacy (SP). 2019: 707-723.
[27] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: Defending against backdooring attacks on deep neural networks [C]// Proceedings of the Research in Attacks, Intrusions, and Defenses: 21st International Symposium (RAID 2018). 2018: 273-294.
[28] XU W T, WANG B J. Backdoor Defense of Horizontal Federated Learning Based on Random Cutting and Gradient Clipping [J]. Computer Science. 2023, 50(11): 356-363.
[29] SHEJWALKAR V, HOUMANSADR A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning [C]// Proceedings of the Network and Distributed System Security Symposium. 2021.
[30] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning? [J]. arXiv: 1911.07963, 2019.
[31] WANG R, ZHANG G, LIU S, et al. Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free Cases [C]// Proceedings of the ECCV 2020. Lecture Notes in Computer Science, 2020: 222-238.
[32] HUANG S, PENG W, JIA Z, et al. One-Pixel Signature: Characterizing CNN Models for Backdoor Detection [C]// Proceedings of the ECCV 2020. 2020: 326-341.
[33] CHEN X, LIU C, LI B, et al. Targeted Backdoor Attacks on Deep Learning Systems Using DataPoisoning [J]. arXiv: 1712.05526, 2017.
[34] CHENG H, XU K, LIU S, et al. Defending against Backdoor Attack on Deep NeuralNetworks [J]. arXiv: 2002.12162, 2020.
[35] YEH I C, LIEN C H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients [J]. Expert Systems with Applications, 2009, 36(2): 2473-2480.

WANG Xun, born in 1999, master. His main research interests include machine learning and machine learning security.

XU Fangmin, born in 1982, Ph.D, associate professor. His main research interests include Internet of things network and future network technology.

(责任编辑:何杨)