基于拉普拉斯特征映射学习的隐匿 FDI 攻击检测

石家宇 [1, 2]　陈 博 [1, 2]　俞 立 [1, 2]

摘　要　智能电网中的隐匿虚假数据入侵 (False data injection, FDI) 攻击能够绕过坏数据检测机制, 导致控制中心做出错误的状态估计, 进而干扰电力系统的正常运行. 由于电网系统具有复杂的拓扑结构, 故基于传统机器学习的攻击信号检测方法存在维度过高带来的过拟合问题, 而深度学习检测方法则存在训练时间长、占用大量计算资源的问题. 为此, 针对智能电网中的隐匿 FDI 攻击信号, 提出了基于拉普拉斯特征映射降维的神经网络检测学习算法, 不仅降低了陷入过拟合的风险, 同时也提高了隐匿 FDI 攻击检测学习算法的泛化能力. 最后, 在 IEEE57-Bus 电力系统模型中验证了所提方法的优点和有效性.

关键词　智能电网, 隐匿虚假数据入侵攻击, 拉普拉斯特征映射, 神经网络

引用格式　石家宇, 陈博, 俞立. 基于拉普拉斯特征映射学习的隐匿 FDI 攻击检测. 自动化学报, 2021, **47**(10): 2494−2500

DOI　10.16383/j.aas.c190551

Stealthy FDI Attack Detection Based on Laplacian Eigenmaps Learning Strategy

SHI Jia-Yu [1, 2]　CHEN Bo [1, 2]　YU Li [1, 2]

Abstract　The stealthy false data injection (FDI) attack in smart grids can bypass the bad data detection, making an incorrect state estimate in the control center, which in turn interferes with the normal operation of the power system. Considering the complex topology of the grid system, the machine learningbased methods has an over-fitting problem caused by high dimensionality, while deep learning-based methods are subject to long training time and occupy a lot of computing resources. Motivated by the above fact, a neural network learning algorithm based on dimensional reduction of Laplacian eigenmaps (LE) is developed in this paper to detect hidden FDI attack signal in the smart grids. The proposed method not only reduces the risk of over-fitting, but also improves the generalization ability of the stealthy FDI attack detection learning algorithm. Finally, IEEE 57-Bus power system is employed to show the advantages and effectiveness of the proposed method.

Key words　Smart grids, stealthy false data injection attack, Laplacian eigenmaps (LE), neural network

Citation　Shi Jia-Yu, Chen Bo, Yu Li. Stealthy FDI attack detection based on Laplacian eigenmaps learning strategy. *Acta Automatica Sinica*, 2021, **47**(10): 2494−2500

收稿日期 2019-07-26　录用日期 2019-12-15
Manuscript received July 26, 2019; accepted December 15, 2019
国家自然科学基金项目 (61973277, 61673351), 浙江省自然科学基金项目 (LR20F030004) 资助
Supported by National Natural Science Foundation of China (61973277, 61673351) and Zhejiang Provincial Natural Science Foundation of China (LR20F030004)
本文责任编委 陈积明
Recommended by Associate Editor CHEN Ji-Ming

1. 浙江工业大学信息工程学院 杭州 310023　2. 浙江工业大学网络空间安全研究院 杭州 310023
1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023　2. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023

智能电网作为下一代电力系统, 通过采用先进的数字信息和通信技术能够实现电网各个环节重要运行参数的在线监测和实时信息掌控, 并在此基础上整合物联网技术和大数据分析, 实现了更环保、更安全和更高效的电力管理[1]. 在发电侧, 由于电能无法大量存储, 控制中心需要密切监控电网运行参数, 以控制电网中的发电与电能消耗相平衡. 在电网侧, 系统也需要估计系统的运行状态, 将其用于最优潮流算法以制定发电方案, 实现经济上的最优调度. 因此需要有大量的信息控制设备和通信传感网络接入电网, 实时发送各个节点的测量值到控制中心, 确保电力系统的高效经济可靠运行. 但是通信设施的接入, 也让智能电网面临着潜在的网络攻击风险, 成为军事或恐怖活动的目标, 例如 2015 年 12 月的乌克兰电网攻击事件, 造成了 30 个变电站被关闭, 约有 23 万人无法用电[2]. 特别地, 隐匿虚假数据入侵 (False data injection, FDI) 攻击可以绕过电力系统中传统的坏数据检测机制, 通过篡改测量数据, 使得对电力系统的状态估计不准确, 进而干扰控制中心决策以扰乱电力市场正常秩序, 存在重大的经济和安全隐患[3].

FDI 攻击自 2009 年提出以来[4], 便受到了国内外学者的广泛关注. 针对不同的 FDI 攻击场景提出了相应的保护策略和攻击检测方案. 在保护策略方面, 主要是通过放置相量测量单元 (Phasor measurement units, PMU) 来增强通信安全. 注意到 PMU 是配备有全球定位系统 (Global positioning system, GPS) 技术的测量设备, 通过与 GPS 时间同步, PMU 能够为电网中地理上分散的节点提供精确的同步相量测量, 使得攻击者更难以篡改 PMU 收集的测量数据[5]. 然而, 高昂的成本却制约着大规模地部署 PMU, 因此如何找到放置 PMU 的关键位置以最大限度地减少 PMU 的数量对于保护电力系统免受 FDI 攻击起着重要的作用. 为此, Kim 等[5] 提出了一种关键节点保护机制, 通过选择关键节点实施保护以尽可能提高攻击者的攻击成本. 文献 [6] 从图论的角度分析智能电网的结构, 提出了一种启发式算法来寻找最佳的测量保护集, 达到防御效果. 与此同时, 在攻击检测方面, Liu 等[7] 利用状态测量的时间相关性, 以及 FDI 攻击的稀疏性来检测广义上的 FDI 攻击[8]. 文献 [9] 提出了一种分布式状态估计方法, 根据估计结果的偏差判断是否遭受 FDI 攻击, 且能够准确定位被篡改的状态变量. 对于具有特殊结构的隐匿 FDI 攻击, 文献 [10−13] 则将其看作是一个统计学习问题, 把历史数据作为训练样本, 根据攻击向量会让正常测量值与被攻击测量值产生 “距离”上的变化这一特征[10], 采用机器学习方法对测量值做分类, 以实现隐匿 FDI 攻击检测的目的. 具体地, Ozay 等[10] 采用了感知机, k 近邻, 支持向量机等经典机器学习方法验证其检测效果. Esmalifalak 等[11] 提出了分布式的支持向量机 (Support vector machine, SVM) 方法, 验证了机器学习方法在隐匿 FDI 攻击检测中的有效性. 除了传统的机器学习方法, 深度学习因其具有自动提取原始数据特征, 能够提取更深层更抽象特征信息的特性, 也受到了许多学者的关注. 文献 [12] 便提出了一种基于深度学习的检测机制, 采用深度信念网络 (Deep belief networks, DBN) 作为检测模型, 并结合条件高斯−伯努利受限玻尔兹曼机 (Conditional Gaussian-Bernoulli restricted Boltzmann machines, CGBRBM) 提取高维时间特征, 以降低训练深度神经网络的复杂度与训练时间, 仿真结果表明该方法比神经网络和 SVM 的检测方法有更高的检测精度. 文献 [13] 则针对交流状态估计中的隐匿 FDI 攻击, 提出了一种结合小波变换和深度神经网络的检测机制, 其中小波

变换提取空间上的相关性, 深度神经网络则提取时域中的特征. 为了得到更好的训练结果, 文献 [13] 构造了 20 万个训练样本以保证样本能够包含所有隐匿 FDI 攻击特征, 最终的训练结果能够很好地提取系统在时域和空间域上的特征, 达到了满意的检测精度, 但在训练过程中也耗费了大量的时间与计算资源.

虽然传统机器学习方法在检测隐匿 FDI 攻击方面取得了一些进展, 但都是在训练集和测试集具有高度相似性的前提下得到的, 因此当测试集与训练集出现较大差异时, 传统机器学习方法将很大可能出现差的学习效果. 而且电力系统往往是高度复杂的, 其历史数据的维度往往是几百甚至几千维, 这使得传统机器学习方法面临 "维数灾难"的问题, 训练结果容易出现过拟合, 进而限制了泛化能力. 而近年来的深度学习方法的性能虽然不受维数的限制, 但也存在训练时间长、占用大量计算资源的缺陷. 因此, 在利用机器学习方法检测隐匿 FDI 攻击中, 通过降维避免训练结果过拟合, 减少模型训练时间显得尤为重要. 为此, 本文提出了基于拉普拉斯特征映射降维的神经网络检测学习机制, 通过拉普拉斯特征映射方法来提取攻击向量的信息, 将测量数据预先降维处理, 再用于训练神经网络得到合适的检测模型. 在 MATPOWER 中的 IEEE 57-bus 上进行了实验验证, 并与没有降维预处理的神经网络训练结果, 深度神经网络训练结果以及利用主成分分析降维预处理后的训练结果做了对比. 实验结果表明, 在智能电网的大规模量测数据压缩降维方面, 拉普拉斯特征映射相比主成分分析能够很好地提取低维特征, 所提出的方法不仅可以有效地检测出隐匿 FDI 攻击, 而且其泛化性能优于单独使用神经网络和深度神经网络的检测方法.

1　问题描述

1.1　系统状态估计

电力系统中的状态估计是指根据各个总线上仪表的测量数据估计系统的状态, 其中测量包括总线电压、总线有功和无功功率, 状态变量包括总线电压和电压相角, 其交流潮流模型的表达形式为:

$$\boldsymbol{z} = h(\boldsymbol{x}) + \boldsymbol{n} \tag{1}$$

其中, $\boldsymbol{x} \in \mathbf{R}^D$ 为电网的状态变量, 即节点电压和相角变量, $\boldsymbol{z} \in \mathbf{R}^N$ 为测量向量, 是传感器的测量数据, $\boldsymbol{n} \in \mathbf{R}^N$ 是测量噪声, $h(\boldsymbol{x})$ 则表示测量值与状态变量之间的非线性关系, 其形式由电网的拓扑结构及总线上的参数决定[14]. 在这里我们假设噪声服从均值为 0, 协方差矩阵为 $\boldsymbol{\Lambda}$ 的高斯分布, 且系统的状态在一段时间内的变化是缓慢的, 因此可以通过在操作点附近泰勒展开, 将非线性的交流模型做线性近似, 得到直流潮流模型, 其数学描述为:

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{2}$$

其中, $\boldsymbol{H} \in \mathbf{R}^{N\times D}$ 是测量雅可比矩阵, 则状态向量估计可以通过加权最小二乘估计求解得到[15]:

$$\hat{\boldsymbol{x}} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{z} \tag{3}$$

1.2　隐匿 FDI 攻击原理

FDI 攻击是指攻击者通过篡改传感器中的测量数据使得系统产生错误的状态估计, 进而使控制中心做出错误决策. 当电网遭受到攻击时, 量测方程 (2) 变为:

$$\widetilde{\boldsymbol{z}} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{a} + \boldsymbol{n} \tag{4}$$

其中, $\boldsymbol{a} \in \mathbf{R}^N$ 为攻击向量. 针对攻击信号 $\boldsymbol{a}$, 常用的检测方法就是坏数据检测 (Bad data detection, BDD)[15], 即:

$$\gamma = \|\widetilde{\boldsymbol{z}} - \boldsymbol{H}\hat{\boldsymbol{x}}\|_2 \tag{5}$$

当测量残差超过一定阈值 $\gamma > \epsilon_0$, 就判断为受到攻击, 其中 ϵ_0 为需要设定的阈值.

从 BDD 检测机制来看, 如果攻击者知道系统的拓扑结构 $\boldsymbol{H}$, 可以构造隐匿 FDI 攻击向量 $\boldsymbol{a} = \boldsymbol{H}\boldsymbol{c}$ 在不改变测量残差的情况下对系统状态估计造成影响[4]. 当遭遇隐匿 FDI 攻击时, 由式 (5) 可得:

$$\begin{aligned}\gamma =& \|\widetilde{\boldsymbol{z}} - \boldsymbol{H}\hat{\boldsymbol{x}}\|_2 = \\ & \|\widetilde{\boldsymbol{z}} - \boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\widetilde{\boldsymbol{z}}\|_2 = \\ & \|\boldsymbol{n} - \boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{n}\|_2\end{aligned} \tag{6}$$

从上式可以看出测得的残差 γ 的大小只受噪声影响, 传统的 BDD 检测方法并不能检测出隐匿虚假数据入侵攻击.

然而要构造这类攻击也并不容易, 攻击者需要掌握电网系统的各种电气参数和拓扑信息 ($\boldsymbol{H}$ 雅可比矩阵), 或者掌握系统所有的测量信息, 利用主成分分析 (PCA) 构造攻击向量[16]. 即使攻击者能够掌握这些信息, 也会受到各种资源等因素的限制, 只能篡改部分测量设备的数据. 因此在这里我们定义 $I = \{i_1, \cdots, i_k\}$ 表示攻击者无法篡改的测量的下标集合, $\bar{I}$ 为对应的补集, 则雅可比矩阵 $\boldsymbol{H}$ 可以拆分为 $\boldsymbol{H}_I$ 和 $\boldsymbol{H}_{\bar{I}}$ 两部分, $\boldsymbol{H}_I$ 表示集合 I 中下标对应的行, $\boldsymbol{H}_{\bar{I}}$ 则为补集 $\bar{I}$ 中下标对应的行, 从而隐匿 FDI 攻击可以表示为如下形式:

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{H}_{\bar{I}}\boldsymbol{c} \\ \boldsymbol{H}_I\boldsymbol{c} \end{bmatrix} \tag{7}$$

其中, $\boldsymbol{H}_{\bar{I}}\boldsymbol{c}$ 表示能被篡改数据的部分, $\boldsymbol{H}_I\boldsymbol{c}$ 则是不能篡改的部分. 因此只要令 $\boldsymbol{H}_I\boldsymbol{c} = 0$, 通过求解出的 $\boldsymbol{H}_I$ 零空间, 如果有非零解, 且 $\boldsymbol{H}_{\bar{I}}\boldsymbol{c}$ 的元素不全为零, 就可以构造出满足隐匿条件的攻击向量 $\boldsymbol{a}$. 上述构造方法能保证构造出来的攻击向量攻击者都可以实现. 从中我们也可以发现当 $\boldsymbol{H}_I$ 列满秩时, $\boldsymbol{H}_I\boldsymbol{c} = 0$ 只有零解, 意味着攻击者无法进行隐匿虚假数据入侵.

通过上述分析, 攻击者可以构建针对直流状态估计的攻击向量, 且不会被基于残差的坏数据检测方法检测到. 因此如何设计一种隐匿 FDI 攻击的检测方法是本文要解决的问题.

2　基于拉普拉斯特征映射降维学习的检测机制

由第 1.2 节可知, 隐匿 FDI 攻击可以绕过传统的坏数据检测, 故如何基于机器学方法训练分类器以识别系统是否受到攻击为这一问题提供了可行的解决思路. 然而, 随着电网规模的不断扩大, 测量数据的维数也成倍增长, 进而导致机器学习 检测方法面临维数灾难挑战, 使得训练结果存在陷入过拟合的风险. 为了克服上述缺点, 本文提出了如图 1 所示的检测机制:

首先我们采用拉普拉斯特征映射对历史数据进行降维预处理, 从而提取低维流形特征, 使降维后的数据相比原始数据更易处理, 然后借助于神经网络学习方法训练分类

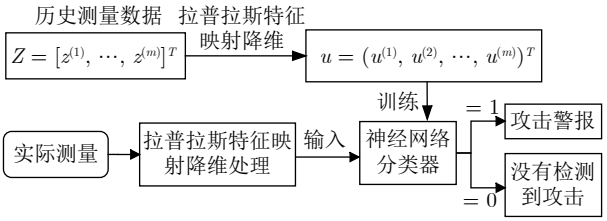

图 1 基于拉普拉斯特征映射降维学习的检测机制

Fig.1 Detection mechanism based on Laplacian eigenmaps

器以实现隐匿 FDI 攻击信号的检测.

2.1 基于拉普拉斯特征映射的机器学习检测

拉普拉斯特征映射 (Laplacian eigenmaps, LE) 是一种非线性的降维方法, 用局部的角度去构建数据之间的关系, 可以反映出数据内在的流形结构[17]. 它的直观思想是希望相互间有关系的点在降维后的空间中尽可能地靠近, 其最小化的目标函数是:

$$\sum_{ij}\left(u^{(i)}-u^{(j)}\right)^2 w_{ij} \tag{8}$$

其中, $u^{(i)}=\left(u_1^{(i)},u_2^{(i)},\cdots,u_n^{(i)}\right)^{\mathrm{T}}$ 为样本 $z^{(i)}$ 降到 n 维后的点, w_{ij} 是测量样本 $z^{(i)}$ 和 $z^{(j)}$ 之间的连接权重. w_{ij} 是根据两个样本点是否接近来确定的, 首先利用 (k-Nearest neighbor, kNN) 方法确定是否在样本 $z^{(i)}$ 与 $z^{(j)}$ 之间设置边连接, 若 $z^{(i)}$ 在 $z^{(j)}$ 的 k 个最近邻居中, 则将 $z^{(i)}$ 和 $z^{(j)}$ 相连, k 是一个预先设定的值, 或者设定合适的 ε, 将 $\|z^{(i)}-z^{(j)}\|^2\leq\varepsilon$ 的节点相连接; 然后确定权重大小, 采用 Heat kernel 函数, 将相连节点的权重设置为 $w_{ij}=\mathrm{e}^{-\frac{\|z^{(i)}-z^{(j)}\|^2}{t}}$, 这里的 t 为预先设定的值, 也可以令 $t=\infty$, 简单地将所有相连节点的权重设为 $w_{ij}=1$, 其他未连接的均为 0; 最终可以得到一个对称邻接矩阵 W.

通过最小化目标函数 (8), 保证了相近的 $z^{(i)}$ 和 $z^{(j)}$ 映射后 $u^{(i)}$ 和 $u^{(j)}$ 两点仍能够保持相近. 目标函数经过整理后可以表示为如下二次型的形式:

$$\begin{aligned}&\sum_{ij}\|u^{(i)}-u^{(j)}\|^2 w_{ij}=\\&\quad\sum_{ij}\left(\|u^{(i)}\|^2+\|u^{(j)}\|^2-2\left(u^{(i)}\right)^{\mathrm{T}}\left(u^{(j)}\right)\right)w_{ij}=\\&\quad\sum_i\|u^{(i)}\|^2 D_{ii}+\sum_j\|u^{(j)}\|^2 D_{jj}-\\&\quad2\sum_{i,j}\left(u^{(i)}\right)^{\mathrm{T}}\left(u^{(j)}\right)w_{ij}=\\&\quad2\boldsymbol{u}^{\mathrm{T}}L\boldsymbol{u}\end{aligned} \tag{9}$$

其中, $\boldsymbol{u}=\left(u^{(1)},u^{(2)},\cdots,u^{(m)}\right)^{\mathrm{T}}$, m 表示样本集中的样本数量, $L=D-W$ 为拉普拉斯矩阵, D 是一个对角矩阵, 满足 $D_{ii}=\sum_j w_{ij}$, W 是一个对称邻接矩阵, 且拉普拉斯矩阵 L 是半正定的.

最终需要求解如下最小化问题:

$$\begin{aligned}&\arg\min_{\boldsymbol{u}}\quad\boldsymbol{u}^{\mathrm{T}}L\boldsymbol{u}\\&\text{s.t.}\quad\boldsymbol{u}^{\mathrm{T}}D\boldsymbol{u}=1\end{aligned} \tag{10}$$

其中, 约束 $\boldsymbol{u}^{\mathrm{T}}D\boldsymbol{u}=1$ 避免了缩放的影响, 最小化目标函数的向量 $\boldsymbol{u}$ 由广义特征值问题的最小特征值解给出[16]:

$$L\boldsymbol{u}=\lambda D\boldsymbol{u} \tag{11}$$

求解得到的非零特征值所对应的特征向量就是降维后的输出.

通过上述方法将训练样本降维处理, 选择最大的两个广义特征值对应的广义特征向量作为低维流形特征. 然后基于低维流形特征, 建立如图 2 所示的三层神经网络, 有输入层、隐藏层和输出层组成[18].

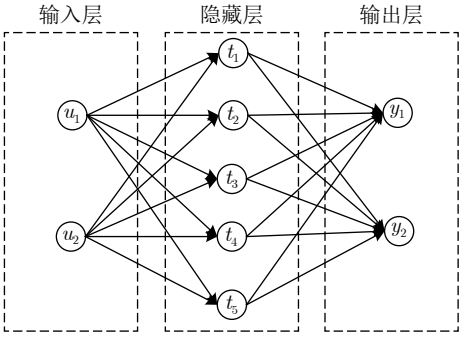

图 2 神经网络示意图

Fig.2 Neural network

其中, 输入层有 2 个神经元组成, 为原始数据降维后得到的 $u=(u_1,u_2)^{\mathrm{T}}$. 隐藏层由 5 个神经元组成, 通过下式计算得到:

$$t_h=\sigma\left(\sum_{i=1}^{2}\omega_{hi}u_i+\omega_h\right),\quad h=1,2,\cdots,5 \tag{12}$$

其中, ω_h 对应每个隐藏层神经元的偏置, ω_{hi} 对应输入 u_i 到神经元 t_h 的权重, σ 则是激活函数 $\sigma(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$. 最后输出层有 2 个神经元 $\hat{y}_1,\hat{y}_2$ 组成, 当他们的输出值大于 0.5 时, 分别表示受到攻击与未受到攻击两种检测结果, 其表达式为:

$$\hat{y}_j=\sigma\left(\sum_{h=1}^{5}\upsilon_{jh}t_h+\upsilon_j\right),\quad j=1,2 \tag{13}$$

其中, υ_j 为对应输出的偏置, υ_{jh} 为对应输入 t_h 到输出 $\hat{y}_j$ 的权重. 最后通过求解以下最优化问题来训练得到权重 $\omega_{hi},\upsilon_{jh}$ 和偏置 ω_h,υ_j:

$$\min\frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{2}\left(y_{kj}-\hat{y}_{kj}\right)^2 \tag{14}$$

其中, y_{kj} 为样本 x_k 的标签. 综上所述, 最终的检测算法步骤归纳如下:

算法 1.

步骤 1. 收集历史测量数据 $Z=\left[z^{(1)},\cdots,z^{(m)}\right]^{\mathrm{T}}$

步骤 2. 拉普拉斯特征映射降维预处理

1) 构建邻接矩阵 W

For $i=1,\cdots,m$

For $j=1,\cdots,m$

if $z^{(j)}$ 在 $z^{(i)}$ 的 k 个邻居中

$w_{ij}=\mathrm{e}^{-\frac{\|z^{(i)}-z^{(j)}\|^2}{t}}$

else

$w_{ij}=0$

2) 特征映射

求解广义特征问题

$\boldsymbol{L}\boldsymbol{u}=\lambda D\boldsymbol{u}$

得到降维处理后的样本

$\boldsymbol{u}=\left(u^{(1)},u^{(2)},\cdots,u^{(m)}\right)^{\mathrm{T}}$

步骤 3. BP 算法训练神经网络

在 0 附近初始化权重 $\omega_{hi},\upsilon_{jh}$ 和偏置 ω_h,υ_j

Repeat

for all $\left(u^{(k)},y_k\right)$ do

a) 计算当前样本的输出 $\hat{y}_k$

b) 计算输出层梯度

$g_j=\hat{y}_{kj}\left(1-\hat{y}_{kj}\right)\left(y_{kj}-\hat{y}_{kj}\right)$

c) 计算隐藏层梯度

$e_h=t_h\left(1-t_h\right)\sum_{j=1}^{2}\upsilon_{jh}g_j$

d) 更新权重 $\omega_{hi},\upsilon_{jh}$ 和偏置 ω_h,υ_j

$\upsilon_{jh}=\upsilon_{jh}+\eta g_j t_h,\upsilon_j=\upsilon_j+\eta g_j$

$\omega_{hi}=\omega_{hi}+\eta e_h u_i,\omega_h=\omega_h\eta e_h$

end

Until 达到停止条件

步骤 4. 将新的测量放入历史数据降维处理, 作为神经网络检测模型的输入, 得到检测结果.

3　仿真

本文利用 IEEE 57-Bus 系统模型验证所提出隐匿 FDI 攻击检测方法的优点和有效性, 即: 采用 LE 降维、PCA 降维的样本集分别训练了神经网络检测模型, 以及未降维预处理的样本集训练了神经网络与深度神经网络模型并做对比与分析, 其中系统的测量雅可比矩阵 $\boldsymbol{H}$ 来自 MATPOWER 工具箱[19]. 通过对 MATPOWER 中的案例进行潮流计算得到电网的系统状态 $x\in\mathbf{R}^D$, 并用于计算得到系统的量测 $z\in\mathbf{R}^N$. IEEE 57-Bus 系统如图 3 所示, 其中状态维数 $D=113$, 测量维数 $N=217$, 这些测量信息将作为本文提出学习算法的训练样本.

3.1　仿真设置

在实验中, 我们考虑攻击者可以访问系统中的 k 个测量, 可以理解为电网系统中, 这 k 个测量存在被 FDI 攻击的隐患, 而其余的测量受到保护. 例如: 在这 k 个节点配备了 PMU, 则测量信息不易被篡改. 事实上, 由于成本限制, 电网系统不能在每个节点上设置 PMU; 与此同时, 攻击者往往也只能够入侵电网中的部分测量, 因此这种假设符合实际情况. 注意到当 $k\leq 104$ 时, 意味着系统中受保护的节点超过状态的维数, 从被攻击者的角度, 防御方完全可以选取合适的量测节点, 使得 $\boldsymbol{H}_I\boldsymbol{c}=0$ 只有零解, 让攻击者无法构造隐匿 FDI 攻击[10]. 因此, 在实验中我们选取了 $k=190,170,150,130$ 四种攻击场景做了仿真验证. 且为了令构造的攻击向量更有 “实际意义”, 能够对智能电网系统造成有效的影响, 我们还对攻击引起的状态向量变化做了进一步地限制.

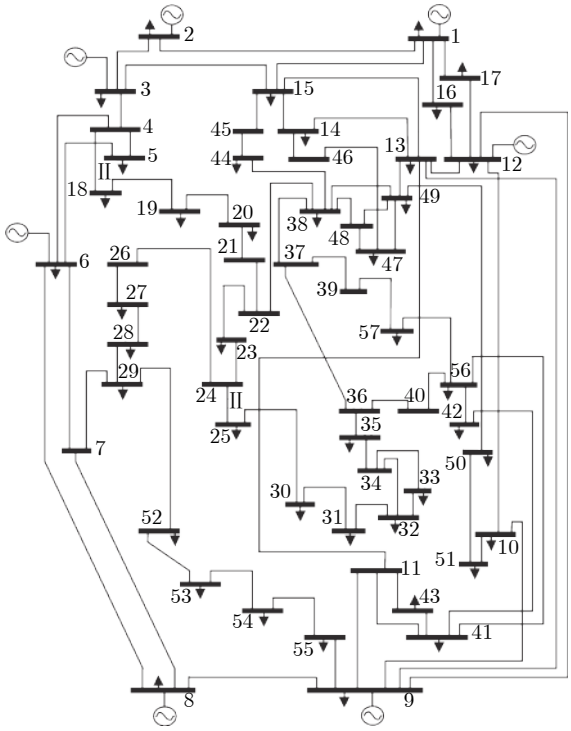

图 3　IEEE 57-Bus 系统

Fig.3　IEEE 57-Bus system

$$\|\boldsymbol{c}\|_\infty\geq\tau \tag{15}$$

其中, $\boldsymbol{c}$ 为隐匿 FDI 攻击对系统状态的影响, 即隐匿 FDI 攻击要对智能电网系统中至少一个状态造成超过大小为 τ 的篡改. 由于在现实中针对电力系统的网络攻击案例并不多见, 且很难得到真实的数据, 因此我们还不能确定 τ 值的大小, 对此我们在仿真实验中设置了 $\tau=1,5,10,15$ 这 4 种情况来分别检验所提出方法的有效性.

此外, 为了验证检测模型的泛化能力, 我们设置了不同的环境噪声 $\mathrm{N}(0,\sigma)$, $\sigma=0.01,0.25,0.50,0.75,1.00$. 通过求解 $\boldsymbol{H}_I\boldsymbol{c}=0$, 构造隐匿攻击向量 a, 并针对不同的 τ,k 和 σ 重复 1 000 次来分别生成训练和测试样本 $Z=[z^{(1)},\cdots,z^{(m)}]^{\mathrm{T}}$. 训练和测试的样本中分别包含 500 个被攻击的样本和 500 个未被攻击的样本. 根据前文式 (7) 的隐匿 FDI 攻击构造方法, 我们设置环境噪声 $\sigma=0.01$, 状态变化阈值 $\tau=10$, 得到一个篡改了 18 个测量数据的隐匿 FDI 攻击, 其对系统状态估计的影响如图 4 所示.

由图 4 可以看到所构造的隐匿 FDI 攻击对系统中的部分状态估计产生了很大的影响. 例如: 节点 20, 30, 50, 51 以及 52 的电压相角都出现了不同大小的偏差, 而系统的残差几乎没有变化, 攻击前的残差为 0.0688, 攻击后的残差为 0.0895. 其中节点 30 的状态变化如图 5 所示, 从第 20 分钟开始受到隐匿 FDI 攻击, 攻击持续时间为十分钟.

进一步地, 在不同环境噪声下的系统被隐匿 FDI 攻击前后的平均残差变化如图 6 所示, 被攻击后的残差变化很小, 可见利用残差检测的方法对隐匿 FDI 攻击是无效的, 且环境噪声变化对残差的影响也很显著.

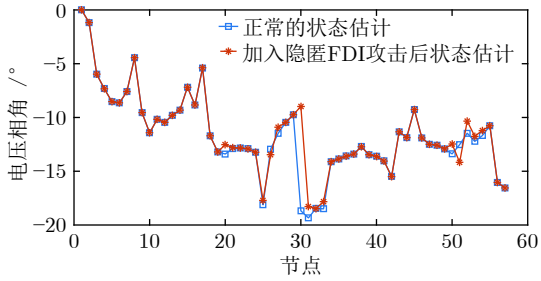

图 4 隐匿 FDI 攻击对系统状态估计的影响

Fig. 4 The effect of stealthy FDI attack on system state estimation

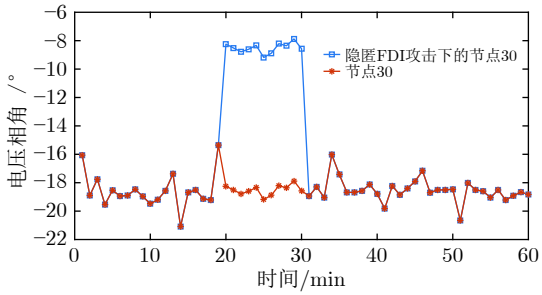

图 5 节点 30 的状态变化曲线

Fig. 5 The state curve of node 30

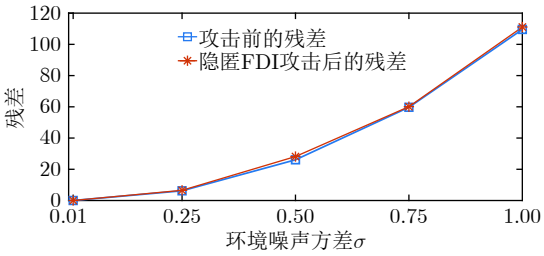

图 6 不同环境噪声下的残差变化

Fig. 6 Residual change under different environmental noise

3.2 降维预处理

将样本集进行拉普拉斯特征映射降维处理, 取最小的两个非零特征值对应的广义特征向量, 数据降维后的二维空间分布如图 7 所示.

此外, 我们也比较了 PCA 降维的效果, 采用 PCA 方法选择协方差矩阵最大的两个特征值对应的特征向量, 将高维数据压缩到二维, 降维后的样本点分布如图 8 所示.

由图 7 和图 8 可以看到直接用 PCA 方法将数据降到二维丢失了许多主成分信息, 降维后样本点是杂糅在一起的, 而应用拉普拉斯特征映射降维后, 正常的测量数据都聚集在一起, 且与被攻击样本有明显的分离, 便于机器学习方法找到决策平面. 拉普拉斯特征映射降维方法之所以能够很好地区分数据点, 是因为两类数据间的距离存在如下的关系[10]:

$$\|\bar{z}_i-\bar{z}_j\|_2=\begin{cases}\|z_i-z_j\|_2+\|a_i-a_j\|_2 & \text{if } i,j\in\bar{S}\\ \|z_i-z_j\|_2+\|a_i\|_2 & \text{if } i\in\bar{S},j\in S\\ \|z_i-z_j\|_2 & \text{if } i,j\in S\end{cases} \tag{16}$$

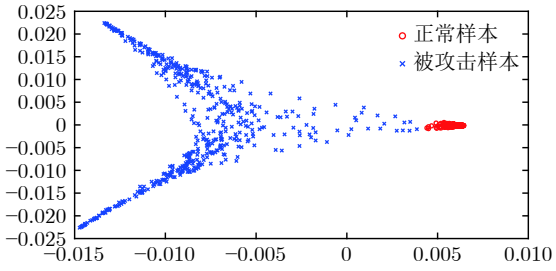

图 7 LE 降维后的样本点分布

Fig. 7 Sample distribution after LE dimension reduction

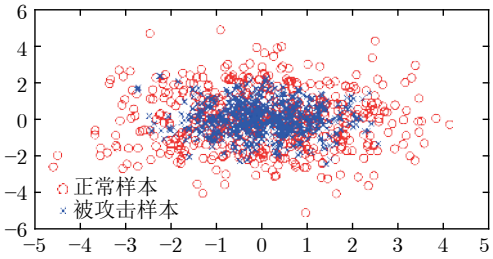

图 8 PCA 降维后的样本点分布

Fig. 8 Sample distribution after PCA dimension reduction

其中, S 表示正常样本的集合, $\bar{S}$ 表示被攻击样本的集合, z_i , $\bar{z}_i$ 分别为正常测量和被攻击的测量, a_i 为攻击向量, 可以看出被攻击的样本和未被攻击的样本存在一定的距离 $\|a\|_2$. 拉普拉斯特征映射在构建邻接图的时候也抓取了这些信息, 只要选取合适的邻居个数 k , 就可以使得所有未被攻击的样本点之间有一个非零的权重, 且与被攻击的样本无连接. 最后通过求解优化问题, 使得 S 中的样本降维后尽可能接近, 而且尽可能不包含 $\bar{S}$ 中的样本, 因此具有区分异常点的特性.

3.3 仿真结果分析

本文采用接受者操作特征 (Receiver operating characteristic, ROC) 曲线中的假阳性率 (False positive Rate, FPR) 和准确率 (Accuracy, ACC) 作为评价各个算法优劣的指标, FPR 和 ACC 计算方式如下:

$$\begin{cases}FPR=FP/\left(FP+TN\right)\\ ACC=\left(TP+TN\right)/\left(TP+FN+FP+TN\right)\end{cases} \tag{17}$$

其中, TP、FP、TN 和 FN 分别表示正确分类的被攻击样本、错误分类的正常样本、正确分类的正常样本和错误分类的被攻击样本. FPR 表示正常样本被误分为被攻击的概率, 定义为误报率, ACC 则为所有样本被正确分类的概率, 定义为检测精度. 我们希望检测精度高的同时, 发生误报的概率也尽可能的低, 因为即使是 1 % 的误报率, 在不断生成的测量数据面前, 也会产生频繁的误报, 对电网控制带来很大影响, 所以我们的目标是 ACC 指标尽可能高, 而 FPR 指标尽可能低, 或者为零.

这里取噪声方差为 $\sigma=0.01$, 状态变化阈值 $\tau=10$, 生成原始样本集, 用于训练深度神经网络和一个三层的神经网络, 并将 LE 降维处理后和 PCA 降维处理后的样本集

分别训练神经网络, 其中深度神经网络我们采用了长短时记忆网络 (Long short-term memory, LSTM)[20], 由输入特征数为 217 的输入层, 具有 100 个隐藏单元的双向 LSTM 层, 大小为 9 的全连接层, softmax 层和分类层 5 层结构组成. 则它们的迭代收敛效果如图 9 所示. 由此图可知发现, 基于拉普拉斯特征映射降维的神经网络均方误差最小, 与深度神经网络的均方误差一致, 且收敛速度要比深度神经网络快很多, 与而基于主成分分析的神经网络收敛效果不明显, 均方误差较大.

然后将训练好的检测模型在另外的测试样本中检验检测精度与误报率, 通过多次的训练并测试, 得到各个算法的 ROC 曲线如图 10–11 所示.

从图 10–11 可以看出, 基于神经网络的检测方法有较高的检测精度, 精度可以达到 90 % 左右, 但是误报率达到了 8 % 左右, 这意味着平均每 100 次检测, 会错误报警 8 次, 因此在实际当中不能将神经网络方法直接用于隐匿 FDI 攻击检测. 而采用 PCA 降维预处理的训练结果, 由于丢掉了大部分主成分信息, 检测效果反而下降明显, 其误报率也达到了 30 % 以上. 此外, 深度神经网络具有很高的检测精度, 均达到了 98 % 以上, 且误报率都在 0.3 % 以下. 最后本文的检测机制的检测精度均达到了 95 % 以上, 且误报率均在 0.5 % 以下, 最少能达到 0.1 %, 相比神经网络的方法, 基于 LE 降维学习方法的检测精度提升明显, 且十分接近深度神经网络的检测效果.

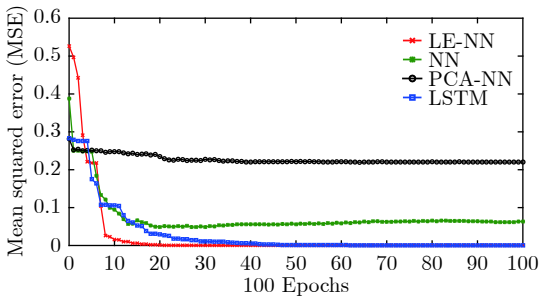

图 9　收敛效果

Fig. 9　Convergence performance

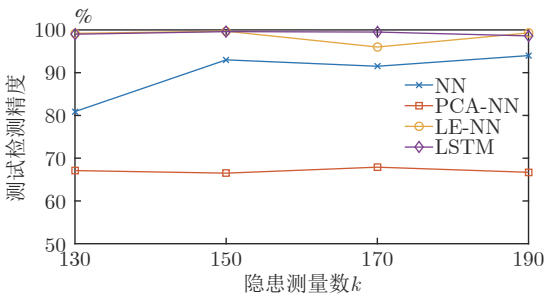

图 10　四种检测机制在不同隐患测量数 k 下的检测精度 ACC

Fig. 10　Detection accuracy of four detection mechanisms

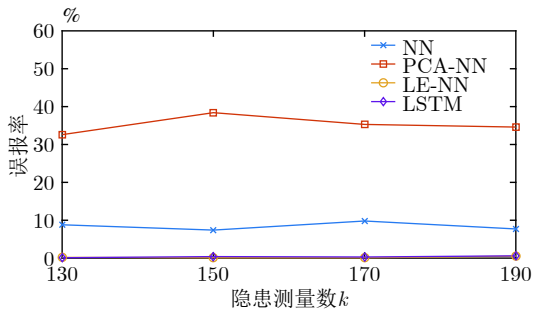

图 11　四种检测机制在不同隐患测量数 k 下的误报率 FPR

Fig. 11　The false positive rate of four detection mechanisms

此外, 为了验证检测模型的泛化能力, 我们在隐患测量数 $k = 150$, 状态变化阈值 $\tau = 10$ 的情景下, 用上述训练得到的检测模型分别对不同噪声环境下的测试样本做了检测, 其中 PCA 降维预处理的检测模型由于测试精度不高, 便不再讨论其泛化性能, 检测结果如图 12–13 所示.

从图 12–13 中的仿真结果可以看出, 单纯神经网络检测方法的性能易受到环境噪声变化的影响, 噪声变大时, 检测精度下降明显, 误报率也在 7 % 以上. 深度神经网络的检测精度也在噪声变大时, 出现了一定幅度的下降, 但也保持了 90 % 以上的检测精度和 5 % 以下的误报率. 而本文提出的检测机制几乎不受噪声变化的影响, 随着噪声增大, 检测精度并没有显著下降, 仍均有 95 % 以上的检测精度, 误报率也不超过 0.8 %. 因此, 与神经网络方法相比, 所提出的 LE 降维学习方法具有更好的泛化性能和鲁棒性.

最后, 考虑到状态变化阈值的选取对检测结果会有明显的影响, 我们在隐患测量数 $k = 150$, 噪声方差为 $\sigma = 0.01$ 的情景下, 用上述训练得到的检测模型对不同的 τ 值的测试样本做了检测, 检测结果如图 14 所示.

从图 14 可以看出, 系统状态量篡改的幅值越大, 检测

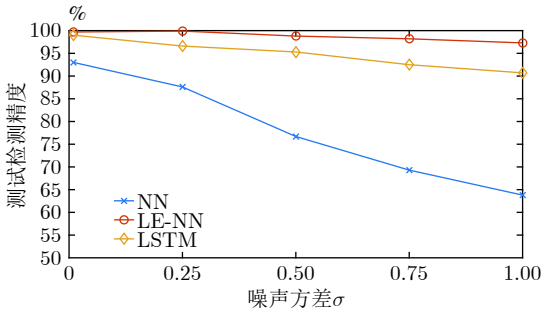

图 12　四种检测方法在不同环境噪声中的检测精度 ACC 变化

Fig. 12　Detection accuracy of three detection mechanisms in different environmental noises

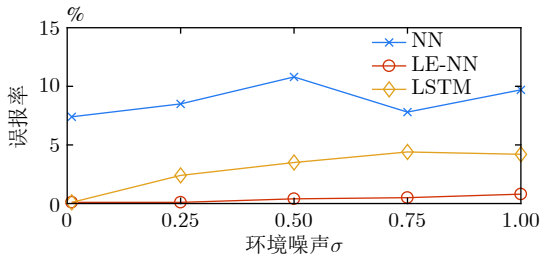

图 13　四种检测方法在不同环境噪声中的误报率 FPR 变化

Fig. 13　False positive rate of three detection mechanisms in different environmental noises

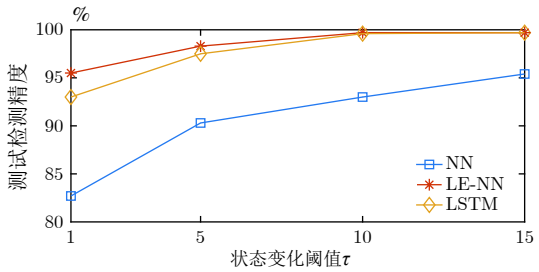

图 14 阈值 τ 对检测精度的影响

Fig. 14 The effect of threshold τ on detection accuracy

效率也越高, 而当攻击改变的状态量较小时, 三种方法的检测精度都有显著的降低, 其中本文提出的检测机制受阈值影响最小, 可见本文所提出方法具有很好的鲁棒性.

4 结语

本文针对电力系统中隐匿 FDI 攻击信号的检测问题, 利用拉普拉斯特征映射将历史数据映射到低维空间, 然后通过构建合适的神经网络结构以建立相应的检测模型, 从而形成基于拉普拉斯特征映射降维学习的隐匿 FDI 攻击信号检测机制. 最后通过 IEEE 57-Bus 模型验证了这种检测机制的有效性. 仿真结果表明采用拉普拉斯特征映射方法能够使正常的测量数据与受攻击的数据很好地分离; 相比于神经网络方法, 这种检测机制能明显提升检测精度, 达到与深度神经网络接近的检测效果. 进一步的, 相比于深度神经网络, 本文的方法不仅能有相似的检测精度, 并且在训练时间上花费更少, 且具有更好的泛化能力.

References

1 Eklas H, Imtiaj K, Fuad U N, Sarder S S, Samiul H S. Application of big data and machine learning in smart grid, and associated security concerns: A Review. *IEEE Access*, 2019, **7**: 13960−13988

2 Liang G Q, Weller S R, Zhao J H, Luo F J, Dong Z Y. The 2015 Ukraine blackout: Implications for false data injection attacks. *IEEE Transactions on Power Systems*, 2017, **32**(4): 3317−3318

3 Wang Qi, Tai Wei, Tang Yi, Ni Ming. A review on false data injection attack toward cyber-physical power system. *Acta Automatica Sinica*, 2019, **45**(1): 72−83
(王琦, 邰伟, 汤奕, 倪明. 面向电力信息物理系统的虚假数据注入攻击研究综述. 自动化学报, 2019, **45**(1): 72−83)

4 Yao L, Peng N, Michael K R. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 2011, **14**(1): No. 13, 33 pages

5 Kim T T, Poor H V. Strategic protection against data injection attacks on power grids. *IEEE Transactions on Smart Grid*, 2011, **2**(2): 326−333

6 Ansari M H, Vakili V T, Bahrak B, Tavassoli P. Graph theoretical defense mechanisms against false data injection attacks in smart grids. *Journal of Modern Power Systems and Clean Energy*, 2018, **6**(5): 860−871

7 Liu L C, Esmalifalak M, Han Z. Detection of false data injection in power grid exploiting low rank and sparsity. International Conference on Communications. Budapest, Hungary: IEEE, 2013.

8 Liang G Q, Zhao J H, Luo F J, Weller S R, Dong Z Y. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 2017, **8**(4): 1630−1638

9 Shan Ke-Meng, Qi Dong-Lian. Distributed dection of false data injection in smart grid and location of error estimation. In: Proceedings of the 36th Chinese Control Conference. Dalian, China: 2017.

10 Ozay M, Esnaola I, Vural F T Y, Kulkarni S R, Poor H V. Machine learning methods for attack detection in the smart grid. *IEEE Transactions on Neural Networks and Learing Systems*, 2015, **27**(8): 1773−1786

11 Esmalifalak M, Liu L C, Nguyen N, Zheng R, Han Z. Detecting stealthy false data injection using machine learning in smart grid. *IEEE Systems Journal*, 2014, **11**(3): 1644−1652

12 He Y B, Mendis G J, Wei J. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. *IEEE Transactions on Smart Grid*, 2017, **8**(5): 2505−2516

13 Yu J Q, Huo Y H, Li V O K. Online false data injection attack detection with wavelet transform and deep neural networks. *IEEE Transactions on Industrial Informatics*, 2018, **14**(7): 3271−3280

14 Sun Y B, Fu M Y, Wang B C, Zhang H S, Marelli D. Dynamic state estimation for power networks using distributed MAP technique. *Automatica*, 2016, **73**: 27−37

15 Ali A, Antonio G E. *Power System State Estimation: Theory and Implementation*. CRC Press, 2004.115−142

16 Yu Z H, Chin W L. Blind false data injection attack using PCA approximation method in smart grid. *IEEE Transactions on Smart Grid*, 2015, **6**(3): 1219−1226

17 Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003, **15**(6): 1373−1396

18 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

19 Ray D Z, Carlos E M S, Robert J T. MATPOWER: Steadystate operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 2011, **26**(1): 12−19

20 Hochreiter S, Schmidhuber J. Long Short-term Memory. *Neural Computation*, 1997, **9**(8): 1735−1780

石家宇 浙江工业大学硕士研究生. 主要研究方向为信息物理系统安全. E-mail: jiayu_shi0621@163.com
(**SHI Jia-Yu** Master student at Zhejiang University of Technology. His main research interest is cyber-physical systems security.)

陈 博 浙江工业大学信息工程学院教授. 主要研究方向为信息融合, 攻击信号检测, 安全估计与控制, 信息物理系统. 本文通信作者. E-mail: bchen@aliyun.com
(**CHEN Bo** Professor at the College of Information Engineering, Zhejiang University of Technology. His research interest covers information fusion, attack signal detection, security estimation and control, and cyber physical system. Corresponding author of this paper.)

俞 立 浙江工业大学信息工程学院教授. 主要研究方向为网络化控制, 信息融合, 信息物理系统. E-mail: lyu@zjut.edu.cn
(**YU Li** Professor at the College of Information Engineering, Zhejiang University of Technology. His research interest covers networked control, information fusion, and cyber physical system.)