文章编号：0258-0926(2021)03-0132-08; doi:10.13832/j.jnpe.2021.03.0132

大型加压重水反应堆隐蔽攻击方法研究

张 妍，樊登宁，黄 宇*，王东风，许培昊

华北电力大学自动化系，河北保定，071003

摘要：为了对大型加压重水反应堆（PHWR）安全防御系统研发提供帮助，研究了 PHWR 网络控制系统中潜在的攻击方式，并提出了一种基于樽海鞘群优化高斯过程回归算法的隐蔽攻击方法。该方法在对 PHWR 网络控制系统实施虚假数据注入时，通过樽海鞘群优化高斯过程回归算法进行系统辨识，获得 PHWR 受攻击区域高精度的估计模型，并利用该估计模型实现隐蔽攻击。仿真结果表明，该攻击方法对 PHWR 造成一定破坏性的同时具有高度的隐蔽性能。

关键词：加压重水反应堆（PHWR）；隐蔽攻击；高斯过程回归算法；系统辨识；虚假数据注入

中图分类号：TL364；TP273 文献标志码：A

Research on Covert Attack Method in Large Pressurized Heavy Water Reactors

Zhang Yan, Fan Dengning, Huang Yu*, Wang Dongfeng, Xu Peihao

Department of Automation, North China Electric Power University, Baoding, Hebei, 071003, China

Abstract: In order to facilitate the research and development of the security defense system for large pressurized heavy water reactors (PHWR), this paper studies the potential attack mode in PHWR networked control system and proposes a covert attack approach based on Gaussian process regression model optimized by salp swarm algorithm. In this method, when the false data is injected into the PHWR networked control system, the system identification is addressed by optimizing the Gaussian process regression algorithm and a high-precision estimation model of the attacked area in PHWR is obtained, and then the estimation model is used to realize the covert attack. The simulation results show that the attack method not only causes some damage to PHWR, but also has high concealment performance.

Key words: Pressurized heavy water reactor (PHWR), Covert attack, Gaussian process regression algorithm, System identification, False data injection

0 引 言

核科学相关研究论证了通过网络控制系统（NCS）控制大型加压重水反应堆（PHWR）的可行性与优越性[1-2]。然而，由于网络与物理系统的紧密结合，PHWR NCS 面临着众多潜在的网络攻击威胁[3]。NCS 中的网络攻击可能会严重破坏经济、环境，甚至危及人类生命，其中最具代表性的例子是震网病毒攻击，该攻击造成伊朗核电站铀浓缩离心机的损坏[4]。在众多的网络攻击问题中，以破坏物理过程并保持对异常检测器隐蔽为目的的隐蔽攻击是非常危险的一类攻击，受到了极大的关注[5-6]。

目前，研究者为构造隐蔽攻击进行了大量的研究。Pang 等[7]假设攻击者具有物理系统的详细参数，提出了一种针对 NCS 反馈和前向通道的隐蔽攻击，该攻击可以破坏闭环系统的稳定性，同

收稿日期：2020-04-08；修回日期：2020-04-17

基金项目：中央高校基本科研业务费专项资金（2019MS099）

作者简介：张 妍（1980—），女，博士研究生，现从事热力系统建模及工业控制系统信息安全研究，E-mail：zhangyan_07@126.com

*通讯作者：黄 宇，E-mail：huangyufish@ncepu.edu.cn

时避免被异常检测器检测到；Hu[8]等假设攻击者完全了解物理系统模型，在系统中注入了虚假数据，异常检测器无法感知到该攻击序列；Chen[9]和 Bai[10]等针对卡尔曼滤波器系统，假设攻击者具有系统的模型知识和噪声的分布参数，设计了特定的攻击方式；Smith 等[11]提出了一种反馈型虚假数据注入攻击结构，在该攻击结构下，物理对象的模型知识越完善，攻击的隐蔽性越好。在上述研究中，攻击者想要保持攻击的隐蔽性，必须拥有物理系统的先验模型知识，但在实际场景中很难具备完善的先验模型知识。

近年来，使用数据驱动的方法从控制器和传感器数据中获取物理系统估计模型知识来构造隐蔽攻击的研究受到越来越多关注。Kim 等[12]根据传感器测量数据得到了系统的估计子空间，将攻击向量隐藏在系统子空间中来影响系统状态。Sa 等[13]针对 PHWR NCS，提出了一种基于回溯搜索优化算法（BSA）的隐蔽攻击方式。但是，文献[12-13]所获物理系统的估计模型精度较低，控制器的输出会出现异常。

高斯过程回归算法（GPR）在数据驱动建模方面具有精度高、泛化性好等优点[14-15]，樽海鞘群算法（SSA）解决了众多参数优化问题[16]。因此，本研究在文献[11]的反馈型虚假数据注入结构基础上，提出了一种基于樽海鞘群优化高斯过程回归算法（SSA-GPR）的隐蔽攻击方法，设计了隐蔽攻击器来执行该攻击，并通过仿真实例验证了本研究提出的攻击方法的隐蔽性。

1　PHWR NCS 模型

核科学相关研究论证了通过用户数据报协议（UDP）以及网际互连协议（IP）的以太网，利用比例-积分-微分（PID）、状态反馈等控制器控制大型 PHWR 的可行性与优势。在 NCS 环路中，从传感器到控制器的传感器数据和从控制器到执行器的控制指令分别以单个数据包的形式在反馈和前向网络回路中传输。与有线控制环路相比，PHWR NCS 无需单独进行点对点布线，而是采用网络通道进行传输，这给网络攻击带来了极大的可能性。

本研究采用 540 MW 的大型 PHWR 模型，可以将其建模成 14 个独立的系统区域，每个系统都由 1 个离散 PID 控制器控制，其详细的动力学特性见文献[2]。

PHWR 中 14 个独立系统具有相似特征，都由被控对象（P）、反馈控制器（C）和异常检测器（D）组成，如图 1 所示。被控对象接收到的控制指令（u）以及反馈控制器接收到的传感器信号值（y_{cm}）通过通讯网络进行传输。

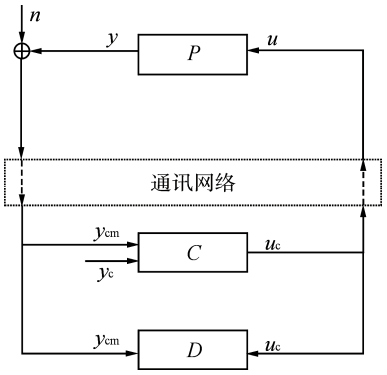


图 1　PHWR NCS 结构

Fig. 1　Structure of PHWR NCS

n—噪声信号；y—被控对象的输出；y_{c}—反馈控制器的设定值；u_{c}—反馈控制器发出的控制指令

被控对象的输入输出关系为：

$$y = Pu \tag{1}$$

反馈控制器接收到的传感器信号值和被控对象接收到的控制指令分别为：

$$y_{\mathrm{cm}} = y + n \tag{2}$$

$$u = u_{\mathrm{c}} \tag{3}$$

反馈控制器根据设定值和反馈值进行运算，发出的控制指令为：

$$u_{\mathrm{c}} = C_{\mathrm{c}} y_{\mathrm{c}} + C_{\mathrm{cm}} y_{\mathrm{cm}} \tag{4}$$

式中，C_{c} 和 C_{cm} 均为反馈控制器控制规律系数。得出反馈控制器接收到的传感器信号值为：

$$y_{\mathrm{cm}} = \left(1 - PC_{\mathrm{cm}}\right)^{-1}\left(PC_{\mathrm{c}} y_{\mathrm{c}} + n\right) \tag{5}$$

2　虚假数据注入

文献[11]针对 NCS 存在的信息安全问题提出一种反馈型虚假数据注入的隐蔽攻击结构。本研究假设攻击者可以截取和篡改通讯网络传输的信号。在实际中，可以通过非法控制物理执行器以及修改传感器传输值来实现。

图 2 给出了反馈型虚假数据注入结构，虚线框住的部分是攻击在通讯网络上的连接情况。

在攻击情况下，被控对象实际上由反馈控制器发出的控制指令和隐蔽攻击器前馈输出信号共

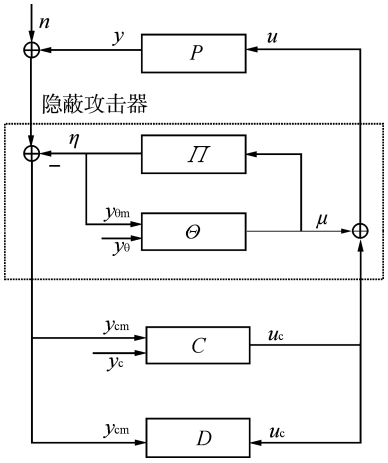


图 2 反馈型虚假数据注入结构

Fig. 2 Feedback False Data Injection Structure

η—隐蔽攻击器反馈输出信号；$y_{\theta m}$—攻击控制器接收到的反馈信号值；y_θ—攻击控制器的设定值；Π—对象模型；Θ—攻击控制器；μ—隐蔽攻击器前馈输出信号

同控制：

$$u = u_c + \mu \tag{6}$$

带有噪声信号的被控对象输出减去隐蔽攻击器反馈输出信号后送给反馈控制器和异常检测器：

$$y_{cm} = y + n - \eta \tag{7}$$

隐蔽攻击器的反馈输出信号和前馈输出信号为：

$$\eta = \left(1 - \Theta_{\theta m}\Pi\right)^{-1}\Pi\Theta_\theta y_\theta \tag{8}$$

$$\mu = \left(1 - \Theta_{\theta m}\Pi\right)^{-1}\Theta_\theta y_\theta \tag{9}$$

式中，Θ_θ 和 $\Theta_{\theta m}$ 均为攻击控制器控制规律系数。

被控对象的输出为：

$$y = \left(1 - PC_{cm}\right)^{-1}P\left(1 - \Pi C_{cm}\right)\left(1 - \Pi\Theta_{\theta m}\right)^{-1} \times \Theta_\theta y_\theta + \left(1 - PC_{cm}\right)^{-1}P\left(C_c y_c + C_{cm} n\right) \tag{10}$$

由式（10）可知，通过输入合适的攻击控制器设定值，攻击控制器可以将被控对象驱动到攻击者期望的状态。

反馈控制器和异常检测器接收到的传感器信号值为：

$$y_{cm} = \left(1 - PC_{cm}\right)^{-1}\left(PC_c y_c + n\right) + \left(1 - PC_{cm}\right)^{-1}\left(1 - \Theta_{\theta m}\Pi\right)^{-1}\Theta_\theta y_\theta\left(P - \Pi\right) \tag{11}$$

为检验异常检测器是否能检测到隐蔽攻击的存在，将无攻击情况下反馈控制器接收到的传感器信号[式（5）]和攻击情况下反馈控制器接收到的传感器信号[式（11）]进行比较，2 者之间的差异值（ω_{attack}）定义为：

$$\omega_{attack} = \left(1 - PC_{cm}\right)^{-1}\left(1 - \Theta_{\theta m}\Pi\right)^{-1}\Theta_\theta y_\theta\left(P - \Pi\right) \tag{12}$$

假设攻击者通过通讯网络截获了物理系统反馈控制器的相关参数，进而设计了一个攻击控制器来复现反馈控制器的控制作用。其中，被控对象的反馈或驱动限制都会被考虑在内。

由式（12）可知，如果攻击者对被控对象完全了解，即 $\Pi=P$，则 $\omega_{attack}=0$，异常检测器将不能检测到该虚假数据注入的存在。但是，考虑到现实情况，攻击者一般不可能完全掌握被控对象的模型知识，即 $\Pi\neq P$，若要尽量使异常检测器不能检测到虚假数据注入的存在，必须使对象模型的动态特性尽可能与被控对象保持一致，即要求具备高精度的对象模型。

3 系统辨识

为获得高精度的对象模型，本研究应用 SSA-GPR 对 PHWR 被控对象进行系统辨识。

3.1 GPR

高斯过程是服从联合高斯分布的任意有限随机变量[$f(x_1), f(x_2), \cdots, f(x_n)$]的集合。假定训练数据集（$\boldsymbol{E}$）为：

$$\boldsymbol{E} = \left[\left(\boldsymbol{X}, \boldsymbol{y}^*\right)\right] = \left[\left(x_i, y_i^*\right)_{i=1}^n\right] \tag{13}$$

$$x_i \in \mathbf{R}_e,\quad y_i^* \in \mathbf{R}$$

式中，$\boldsymbol{X}$ 为训练数据；x_i 和 y_i^* 分别为训练数据集中第 i 个输入变量和其相应的输出值；下标 e 表示输入变量维数；$\mathbf{R}$ 为实数集。均值函数[$m(x)$]和协方差函数[$k(x, x')$]确定了高斯过程的性质，定义如下：

$$f\left(x_i\right) \sim G\left[m\left(x\right), k\left(x, x'\right)\right] \tag{14}$$

式中，G 为高斯过程；随机变量 x、$x' \in \mathbf{R}_e$，为任意值。一般情况下，会预处理数据，故令 $m(x)=0$。GPR 是推断输入变量和相应输出值之间的关系[$f(x_i)$]。在实际回归问题中，噪声将会影响被控对象的输出值，则带有噪声的 GPR 模型为：

$$y^* = f\left(x_i\right) + \varepsilon \tag{15}$$

式中，ε 为独立于 $f(x_i)$的噪声，服从高斯分布，即 $\varepsilon \sim N(0, \sigma_n^2)$（$\sigma_n^2$ 为方差；N 为正态分布）。

可得到训练数据集输出值的高斯分布为：

$$\boldsymbol{y}^{*} \sim G\left[0, k\left(x, x^{'}\right)+\sigma_{n}^{2}\right] \tag{16}$$

假定测试数据集为 $E_{*}=[(\mathbf{X}_{*}, \boldsymbol{y}_{*})]$，根据贝叶斯原理，训练数据集的输出值和测试数据集的预测值（$\boldsymbol{y}_{*}$）之间的联合高斯分布为：

$$\begin{bmatrix}\boldsymbol{y}^{*}\\ \boldsymbol{y}_{*}\end{bmatrix} \sim \left\{0, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})+\sigma_{n}^{2}\boldsymbol{I}_{n} & \boldsymbol{K}(\boldsymbol{X}, x_{*})\\ \boldsymbol{K}(x_{*}, \boldsymbol{X}) & \boldsymbol{k}(x_{*}, x_{*})\end{bmatrix}\right\} \tag{17}$$

式中，x_{*} 为测试数据；$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ 为 $n\times n$ 阶对称正定的协方差矩阵，$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})=k(x_i,x_j)$；$\boldsymbol{I}_{n}$ 为 n 阶单位矩阵；$\boldsymbol{K}(\boldsymbol{X}, x_{*})$ 为测试数据与训练数据之间的 $n\times 1$ 阶协方差矩阵，$\boldsymbol{K}(\boldsymbol{X}, x_{*})=\boldsymbol{K}(x_{*}, \boldsymbol{X})^{\mathrm{T}}$；$k(x_{*}, x_{*})$ 为测试数据自身的协方差。

因此，在训练数据和测试数据给定后，$\boldsymbol{y}_{*}$ 的高斯分布为：

$$\boldsymbol{y}_{*} \mid \boldsymbol{X}, \boldsymbol{y}^{*}, x_{*} \sim N\left[\bar{\boldsymbol{y}}_{*}, \operatorname{cov}\left(\boldsymbol{y}_{*}\right)\right] \tag{18}$$

$$\bar{\boldsymbol{y}}_{*}=\boldsymbol{K}\left(x_{*}, \boldsymbol{X}\right)\left[\boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{X}\right)+\sigma_{n}^{2}\boldsymbol{I}_{n}\right]^{-1}\boldsymbol{y}^{*} \tag{19}$$

$$\operatorname{cov}\left(\boldsymbol{y}_{*}\right)=k\left(x_{*}, x_{*}\right)-\boldsymbol{K}\left(x_{*}, \boldsymbol{X}\right)\times \left[\boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{X}\right)+\sigma_{n}^{2}\boldsymbol{I}_{n}\right]^{-1}\boldsymbol{K}\left(\boldsymbol{X}, x_{*}\right) \tag{20}$$

式中，$\bar{\boldsymbol{y}}_{*}$ 和 cov($\boldsymbol{y}_{*}$)分别为 $\boldsymbol{y}_{*}$ 的均值和方差。

GPR 中“核函数”即协方差函数有多种选择，常用的核函数是平方指数函数：

$$k\left(x_{i}, x_{j}\right)=\sigma_{\mathrm{f}}^{2}\exp\left[\frac{-\left(x_{i}-x_{j}\right)^{2}}{2l^{2}}\right] \tag{21}$$

式中，σ_{f}^{2} 为核函数信号方差；l 为方差尺度。σ_{f}^{2} 和 l 分别用于控制输入变量的局部相关性和 GPR 模型的光滑程度。在确定了核函数信号方差和方差尺度后，可利用式（19）和式（20）得到测试数据对应的预测值均值和方差。

优秀的超参数可使得 GPR 获得良好的拟合精度与泛化能力，因此，超参数的优化问题至关重要。将 $\theta=(\sigma_{\mathrm{f}}^{2}, l)$ 设置为 GPR 的超参数，本研究选用 SSA 优化该超参数。

3.2　SSA

通过模拟樽海鞘群的觅食行为，Mirjalili提出 SSA[16]，在解决优化问题时，该算法首先在限定范围内随机初始化樽海鞘的位置，计算每只樽海鞘的适应度值，再将最优樽海鞘的位置分配给食物源，种群中领导者和追随者分别按照一定规律更新位置，食物源引导整个樽海鞘链向其逼近，追随者的顺次移动有效防止了算法轻易陷入局部最优的困境。

4　隐蔽攻击器设计

隐蔽攻击器的设计分为 3 个阶段：数据记录阶段、模型训练阶段和输出预测阶段。

4.1　数据记录阶段

攻击者通过通讯网络截获 PHWR 受攻击区域反馈控制器发出的控制指令和接收到的传感器信号值，从而生成训练数据集，训练数据集描述如下：

（1）$y_{\mathrm{cm}}=\{y_{\mathrm{cm},1}, y_{\mathrm{cm},2}, \cdots, y_{\mathrm{cm},k}, \cdots, y_{\mathrm{cm},m}\}$表示某一有限时间窗口上截获的一组传感器信号数据集，其中 k 表示该时间窗口上的采样时刻，k=1,2, …, m；$u_{\mathrm{c}}=\left\{u_{\mathrm{c},1}, u_{\mathrm{c},2}, \cdots, u_{\mathrm{c},k}, \cdots u_{\mathrm{c},m}\right\}$表示该时间窗口上截获的一组控制指令数据集。

（2）$y_{\mathrm{cm},k}=\left\{y_{\mathrm{cm},k}^{1}, y_{\mathrm{cm},k}^{2}, \cdots, y_{\mathrm{cm},k}^{j}, \cdots, y_{\mathrm{cm},k}^{p}\right\}$表示第 k 采样时刻上的一组传感器信号数据集，j 表示被控对象输出变量的编号，j=1, 2, … , p；$u_{\mathrm{c},k}=\left\{u_{\mathrm{c},k}^{1}, u_{\mathrm{c},k}^{2}, \cdots, u_{\mathrm{c},k}^{i}, \cdots, u_{\mathrm{c},k}^{q}\right\}$ 表示第 k 采样时刻上的一组控制指令数据集，i 表示被控对象输入变量的编号，i=1, 2, …, q。

4.2　模型训练阶段

隐蔽攻击器通过基于 SSA-GPR 的系统辨识方法获得 PHWR 受攻击区域被控对象的估计模型，即图 2 中的对象模型。训练对象模型时，利用 SSA 优化 GPR 的超参数，提高模型的拟合精度与泛化能力。由于不了解 $y_{\mathrm{cm},k+1}^{j}$ 与 $u_{\mathrm{c},k}$ 以及 $y_{\mathrm{cm},k+1}^{j}$ 与 $y_{\mathrm{cm},k}$ 之间的相关性，攻击者可以选择 $u_{\mathrm{c},k}$ 和 $y_{\mathrm{cm},k}^{j}$ 作为 GPR 的输入，$y_{\mathrm{cm},k+1}^{j}$ 作为输出，因此对象模型表示为：

$$\Pi=\left\{\pi^{1}, \pi^{2}, \cdots, \pi^{j}, \cdots, \pi^{p}\right\} \tag{22}$$

式中，π^{j} 为第 j 个被控对象输出变量对应的 SSA-GPR 模型训练过程。

4.3　输出预测阶段

保存训练结束的对象模型，将其用于被攻击系统中的虚假数据注入，如图 2 所示，对象模型在 t 时刻的输出（$y_{\theta\mathrm{m},t}$）为：

$$y_{\theta\mathrm{m},t}=\left\{y_{\theta\mathrm{m},t}^{1}, y_{\theta\mathrm{m},t}^{2}, \cdots, y_{\theta\mathrm{m},t}^{j}, \cdots, y_{\theta\mathrm{m},t}^{p}\right\} \tag{23}$$

$$y_{\theta\mathrm{m,t}}^{j}=\begin{cases}\pi^{j}\left(y_{\mathrm{m,Initial}}^{j},\mu_{t_s}\right) & t_{\mathrm{s}}=t-1\\ \pi^{j}\left(y_{\theta\mathrm{m},t-1}^{j},\mu_{t-1}\right) & t_{\mathrm{s}}<t-1\end{cases} \tag{24}$$

式中，$y_{\theta\mathrm{m},t}^{j}$ 为对象模型的第 j 个变量在 t 时刻的输出值；$y_{\mathrm{m,Initial}}^{j}$ 为被控对象的第 j 个变量的初始输出值；t_{s} 为隐蔽攻击开始的时间；$\mu_{t_{\mathrm{s}}}$ 为攻击开始时刻隐蔽攻击器前馈输出信号；μ_{t-1} 为 t-1 时刻隐蔽攻击器前馈输出信号。

5 实验分析

5.1 实验准备

考虑实验的简便性，本实验选用 540 MW PHWR 14 个独立系统中的一个来评估本研究隐蔽攻击方法的影响。与文献[13]描述的系统参数一致，被控对象在满功率附近的离散化传递函数[$P(z)$]为：

$$P(z)=\frac{0.0001889z}{z^{2}-1.289z+0.2891} \tag{25}$$

式中，z 为复变量。

反馈控制器的传递函数[$C(z)$]为：

$$C(z)=k_{\mathrm{p}}+T_{\mathrm{s}}k_{\mathrm{i}}\left(\frac{z}{z-1}\right)+\frac{k_{\mathrm{d}}}{T_{\mathrm{s}}}\left(\frac{z-1}{z}\right) \tag{26}$$

式中，k_{p} 为比例系数，k_{p}=348.52；k_{i} 为积分系数，k_{i}=17.25；k_{d} 为微分系数，k_{d}=10.79；T_{s} 为采样周期，T_{s}=0.5 s。施加到反馈控制器设定值的指令是一个斜坡信号，上升速率为 0.66 MW/s，上升时间为 10 s，之后保持恒定。

本实验对 PHWR 中的隐蔽攻击进行了无噪声和有噪声 2 种情况的仿真分析。为了模拟噪声，将 $n\sim N(0,\sigma_n^2)$ 的高斯白噪声插入到 PHWR NCS 中，如图 1 所示。调整标准差的方式是令 $I=2\sigma_n$，其中 I=0.02 rad/s。

进行系统辨识时，GPR 选用平方指数函数为核函数；SSA 的种群规模设置为 10，其搜索控制的下界和上界分别设置为[1, 1]和[800, 100]，优化迭代次数为 100。

实验设置了 2 个观测器来获得实验数据，如图 3 所示。观测器 1 用于捕获反馈控制器接收到的传感器信号值，观测器 2 用于捕获反馈控制器发出的控制指令。

为了更好地评估基于 SSA-GPR 系统辨识的准确性以及虚假数据注入的隐蔽性，在无噪声模拟实验中，用文献[13]中提出的 BSA 设计的系统辨识和虚假数据注入实验作为对比。由于文献[13]未对有噪声系统进行实验分析，所以不设置有噪声对比实验。

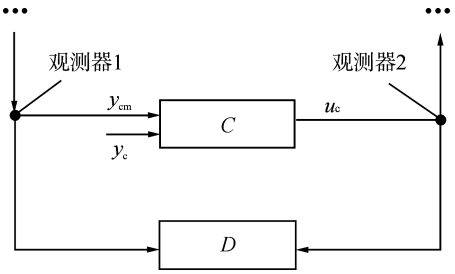


图 3 实验数据观测器设置

Fig. 3 Setting of Experimental Data Observers

5.2 实验结果分析

5.2.1 无噪声 PHWR NCS 在无攻击情况下，记录 2 个观测器在时间窗口[0 s, 200 s]内捕获的数据，并做必要的数据预处理，生成用于系统辨识的训练数据集。

根据第 4 节所述，将用于系统辨识的数据集导入 SSA-GPR 模型进行训练，利用绝对值误差来评估系统辨识的拟合情况。

图 4 为每个采样时刻拟合值与实际值之间的绝对值误差。由图 4 可以看出，误差逐渐减小，没有出现较大的异常值，最大绝对值误差为 5.69×10^{-4} MW，平均绝对值误差为 1.46×10^{-5} MW，系统辨识的拟合精度较高。

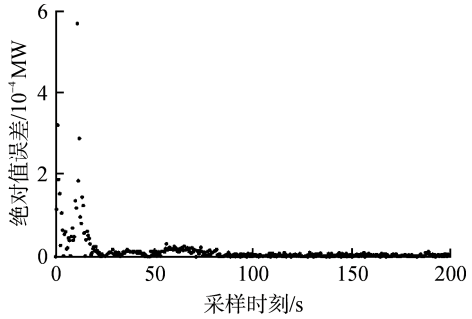


图 4 每个采样时刻拟合值与实际值之间的绝对值误差

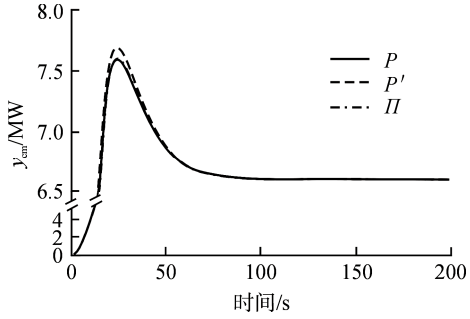
Fig. 4 Absolute Error between Fitting Value and Actual Value at Each Sampling Time

为了验证辨识模型在闭环控制系统中的有效性，将图 1 中的被控对象替换为对象模型，再通过 2 个观测器捕获在闭环控制下系统的响应数据。文献[13]利用 BSA 辨识被控对象的传递函数，该方法通过 600 次算法迭代获得的最佳辨识模型

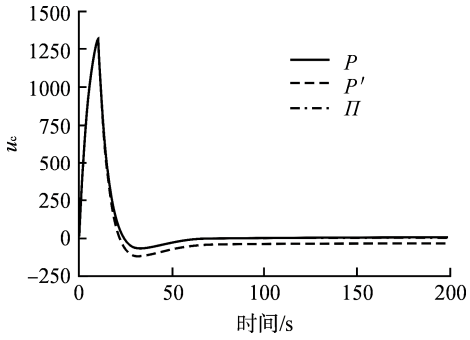
（P'）为：

$$P'(z)=\frac{0.000189z}{z^2-1.289z+0.288} \tag{27}$$

作为对比实验，同样地将图 1 中的被控对象替换为 BSA 中的最佳辨识模型获得响应数据。图 5 为不同系统辨识模型的闭环响应。



a　反馈控制器接收到的传感器信号值对比



b　反馈控制器发出的控制指令对比

图 5　不同系统辨识模型的闭环响应

Fig. 5　Closed-Loop Response of Different System Identification Models

由 2 种不同的系统辨识方法获得的辨识模型与实际被控对象在闭环控制下反馈控制器接收到的传感器信号值如图 5a 所示，基于 SSA-GPR 的系统辨识方法相比于基于 BSA 获得的辨识模型，其输出更加接近原系统被控对象输出，2 者与原系统被控对象输出的平均绝对值误差分别为 4.4×10^{-5} MW 和 0.012 MW，在动态响应曲线波峰位置可以看出明显差异。反馈控制器发出的控制指令如图 5b 所示，基于 BSA 模型的控制器控制指令与原系统控制器控制指令有较大的差别，平均绝对值误差为 37.49；而基于 SSA-GPR 的系统辨识模型的控制器控制指令与原系统控制指令差异较小，平均绝对值误差仅为 0.016。

为评估基于 SSA-GPR 的虚假数据注入攻击的隐蔽性，利用反馈型虚假数据注入结构（图 2）进行虚假数据注入。此攻击的目的是使被控对象功率增加 6.6 MW，同时要求反馈控制器接收到的传感器信号值以及发出的控制指令变化很小（与无攻击情况相比）。攻击控制器的结构和参数与反馈控制器一致，其设定值是开始于 30 s 的斜坡信号，以 0.66 MW/s 的速率增加 10 s，之后保持恒定。图 6 为有攻击和无攻击情况下的 PHWR 响应。

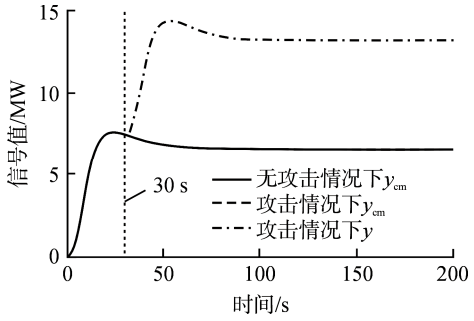


图 6　有攻击和无攻击情况下的 PHWR 响应

Fig. 6　PHWR Response with and without Attack

从图 6 中可以看出，隐蔽攻击器使被控对象的输出收敛到 13.2 MW 附近（正常功率为 6.6 MW），这可能会造成该 PHWR 区域处于一个危险的状态；但是反馈控制器接收到的传感器信号值（异常检测器的输入信号）基本正常。

图 7 为不同攻击情况下无噪声 PHWR 观测器捕获的数据。

通过比较图 7a 中的反馈控制器发出的控制指令可以看出，基于 SSA-GPR 的隐蔽攻击中反馈控制器控制指令比基于 BSA 的隐蔽攻击相似性更高（与无攻击情况相比），基于 BSA 的隐蔽攻击较容易被异常检测器感知。另外，图 7b 中基于 SSA-GPR 和 BSA 的隐蔽攻击以及无攻击情况下反馈控制器接收到的传感器信号值都非常相似，为了更加清晰地展示 2 种攻击方法隐蔽性的差异，通过计算差异值来比较，2 种攻击方法差异值分别为[$\omega_{\text{attack1}}(k)$、$\omega_{\text{attack2}}(k)$]，计算公式为：

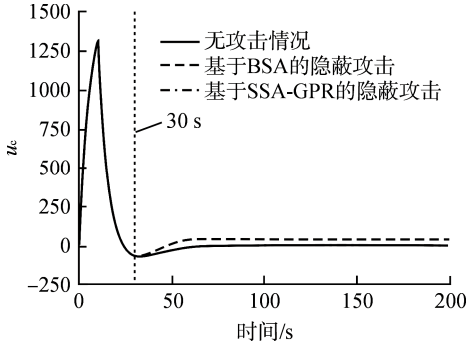
$$\omega_{\text{attack1}}(k)=y_{\text{cm}}^{\text{attack1}}(k)-y_{\text{cm}}^{\text{normal}}(k) \tag{28}$$

$$\omega_{\text{attack2}}(k)=y_{\text{cm}}^{\text{attack2}}(k)-y_{\text{cm}}^{\text{normal}}(k) \tag{29}$$

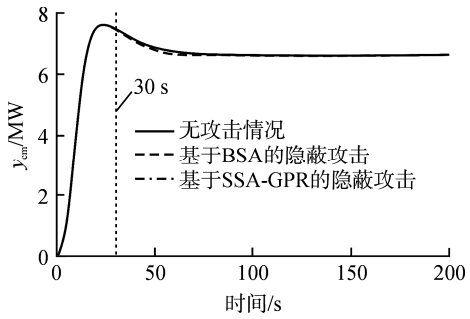
式中，$y_{\text{cm}}^{\text{attack1}}(k)$ 和 $y_{\text{cm}}^{\text{attack2}}(k)$ 分别为在基于 SSA-GPR 和基于 BSA 的隐蔽攻击情况下 k 时刻的反馈控制器接收到的传感器信号值；$y_{\text{cm}}^{\text{normal}}(k)$ 为无攻击情况下 k 时刻的反馈控制器接收到的传

感器信号值。

图 8 为反馈控制器接收到传感器信号的差异值。可以看出，基于 BSA 的隐蔽攻击导致的差异值曲线的峰值要高于基于 SSA-GPR。从攻击者的角度看，本研究提出的隐蔽攻击方法的隐蔽性能更好。



a 不同攻击情况下反馈控制器发出的控制指令



b 不同攻击情况下反馈控制器接收到的传感器信号值

图 7 不同攻击情况下无噪声 PHWR 观测器捕获的数据

Fig. 7 Data Captured by Noiseless PHWR Observers in the Case of Different Attacks

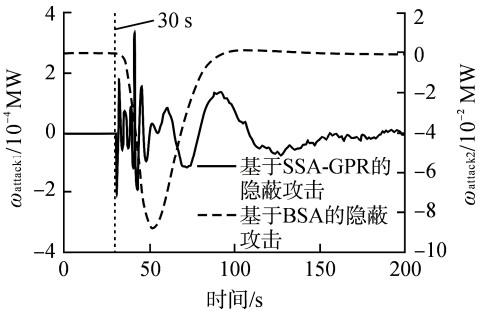
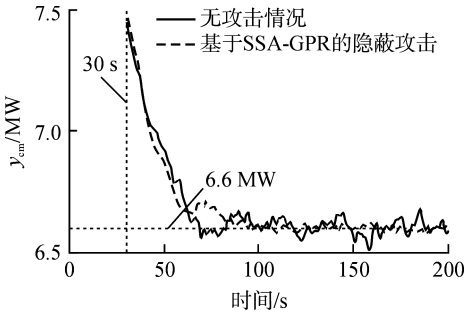


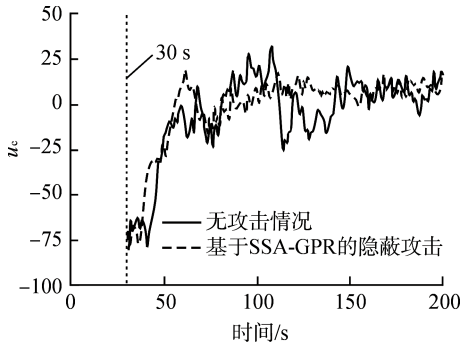
图 8 反馈控制器接收到传感器信号的差异值

Fig. 8 Difference of Sensor Signal Received by Feedback Controller

5.2.2 有噪声 PHWR NCS 与 5.2.1 节所述的实验步骤一致，图 9 为不同攻击情况下有噪声 PHWR 观测器捕获的数据。由图 9 可以看出，在 30 s 时加入攻击后，反馈控制器接收到的传感器信号和发出的控制指令都没有出现较大的异常



a 反馈控制器接收到的传感器信号值对比



b 反馈控制器发出的控制指令对比

图 9 不同攻击情况下有噪声 PHWR 观测器捕获的数据

Fig. 9 Data Captured by Noisy PHWR Observers in the Case of Different Attacks

值，与无攻击情况的信号值相比，差异较小，微小的差异将被噪声信号掩盖。为降低误诊率，基于阈值限制等算法的异常检测器将阈值设置在带有噪声的正常信号范围外，故很难感知到受攻击后反馈控制器接收到的传感器信号和发出的控制指令异常。另外，与 5.2.1 节相同的是受攻击后 PHWR 被控对象的输出收敛到 13.2 MW 附近（正常功率为 6.6 MW）。

6 结 论

（1）在无噪声 PHWR NCS 中，通过 SSA-GPR 执行的系统辨识可以高度准确地模拟 PHWR 被控对象行为，为实施隐蔽性虚假数据注入提供有力帮助。在闭环控制下，与原系统相比，基于 BSA 的模型输出以及反馈控制器发出的控制指令平均绝对值误差分别为 0.012 MW 和 37.49，而基于 SSA-GPR 的模型输出以及反馈控制器发出的控制指令平均绝对值误差仅为 4.4×10^{-5} MW 和 0.016。

（2）在无噪声 PHWR NCS 中，攻击者执行基于 SSA-GPR 的隐蔽攻击与基于 BSA 的隐蔽攻击相比，基于 SSA-GPR 的隐蔽攻击方法具有更高的

隐蔽性。

（3）在有噪声 PHWR NCS 中，攻击者通过执行基于 SSA-GPR 的隐蔽攻击，在造成 PHWR 受攻击区域功率增加 6.6 MW 的情况下，反馈控制器接收到的传感器信号以及发出的控制指令与无攻击情况的信号值差异较小，微小的差异将被噪声信号掩盖。因此，基于 SSA-GPR 的隐蔽攻击具有较高的隐蔽性。

本研究不是要对 PHWR NCS 进行攻击，而是揭示潜在的攻击方式，以此促进防御措施的研究。在今后的工作中，将会研究设计基于 SSA-GPR 隐蔽攻击的检测方法，进一步提高 PHWR 安全防御系统的安全性能。

参考文献：

[1] DASGUPTA S, ROUTH A, BANERJEE S, et al. Networked control of a large pressurized heavy water reactor(PHWR) with discrete Proportional-Integral-Derivative(PID) controllers[J]. IEEE Transactions on Nuclear Science, 2013, 60(5): 3879-3888.

[2] DAS M, GHOSH R, GOSWAMI B, et al. Network Control System applied to a large pressurized heavy water reactor[J]. IEEE Transactions on Nuclear Science, 2006, 53(5): 2948-2956.

[3] 孙子文，张炎棋. 工业信息物理系统的攻击建模研究[J]. 控制与决策，2019, 34(11): 2323-2329.

[4] KUSHNER D. The real story of stuxnet[J]. Spectrum, IEEE, 2013, 50(3):48-53.

[5] LI Z, YANG G H . A data-driven covert attack strategy in the closed-loop cyber-physical systems[J]. Journal of the Franklin Institute, 2018, 355(14): 6454-6468.

[6] WANG D, WANG Z, SHEN B, et al. Recent advances on filtering and control for cyber-physical systems under security and resource constraints[J]. Journal of the Franklin Institute, 2016, 353(11): 2451-2466.

[7] PANG Z H, LIU G P, ZHOU D, et al. Two-Channel false data injection attacks against output tracking control of networked systems[J]. IEEE Transactions on Industrial Electronics, 2016, 63(5):3242-3251..

[8] HU L, WANG Z, HAN Q L, et al. State estimation under false data injection attacks: Security analysis and system protection[J]. Automatica, 2018(87): 176-183.

[9] CHEN Y, KAR S, MOURA J M F. Cyber physical attacks constrained by control objectives[C]//2016 American Control Conference (ACC), IEEE, 2016: 1185-1190.

[10] BAI C Z, PASQUALETTI F, GUPTA V. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs[J]. Automatica, 2017(82): 251-260.

[11] SMITH R S. Covert misappropriation of networked control systems: Presenting a feedback structure[J]. IEEE Control Systems Magazine, 2015, 35(1): 82-92.

[12] KIM J , TONG L , THOMAS R J . Subspace methods for data attack on state estimation: a data driven approach[J]. IEEE Transactions on Signal Processing, 2015, 63(5): 1102-1114.

[13] DE SÁ A O, CARMO L F R C, MACHADO R C S. Evaluation on passive system identification and covert misappropriation attacks in large pressurized heavy water reactors[C]//2018 Workshop on Metrology for Industry 4.0 and IoT, IEEE, 2018: 203-208.

[14] WILLIAMS C K I, RASMUSSEN C E. Gaussian processes for machine learning[M]. Cambridge, MA: MIT press, 2006: 11-28.

[15] 蒋波涛，黄新波. 基于高斯过程回归的临界热流密度预测[J]. 核动力工程，2019, 40(05): 46-50.

[16] MIRJALILI S, GANDOMI A H , MIRJALILI S Z , et al. Salp swarm algorithm: A bio-inspired optimizer for engineering design problems[J]. Advances in Engineering Software, 2017(114): 163-191.

（责任编辑：杨灵芳）