动态水印攻击检测方法的鲁棒性研究

杜大军 [1,2] 张竞帆 [1,2] 张长达 [1,2] 费敏锐 [1,2] YANG Tai-Cheng [3]

摘 要 针对传统动态水印 (Dynamic-watermarking, DWM) 检测方法无法适用模型不确定系统的攻击检测问题, 首先分析模型不确定项导致的传统动态水印检测失效原因, 然后考虑模型不确定项和过程噪声的统计规律, 将其影响转化为对方差变化特性进行分析, 提出两个具有鲁棒性的攻击检测式以及检测式中关键时变方差阈值的确定方法; 其次采用系统失真信号功率定量刻画攻击信号造成系统性能损失程度, 理论证明了系统失真信号功率上界; 在此基础上考虑最坏情况下攻击能够躲过检测, 基于水印信号与其他混合信号相互独立性新增第三检测式, 同时理论证明了系统失真信号功率上界进一步受限范围, 进而提升不确定系统的安全性; 最后仿真算例验证了所提方法的有效性和可行性.

关键词 动态水印, 模型不确定, 失真信号功率, 攻击检测

引用格式 杜大军, 张竞帆, 张长达, 费敏锐, YANG Tai-Cheng. 动态水印攻击检测方法的鲁棒性研究. 自动化学报, 2023, **49**(12): 2557−2568

DOI 10.16383/j.aas.c200614

Robustness of Dynamic-Watermarking Attack-detection Method

DU Da-Jun[1,2] ZHANG Jing-Fan[1,2] ZHANG Chang-Da[1,2] FEI Min-Rui[1,2] YANG Tai-Cheng[3]

Abstract The traditional dynamic-watermarking (DWM) cyber-attack detection cannot directly apply to uncertain systems. In this paper, we first analyze this failure due to model uncertainty. Secondly, based on the statistics of model uncertainty and process noise, we derive two robust attack-detection formulas, and the method of determining the critical time-varying-variance threshold in the formulas. Then, the system performance loss caused by the attack signal, and the upper limit of the distortion signal power are quantified. Furthermore, in the worst case, to avoid that an attack can pass-through the above two-formula detection, the third detection formula is derived. In addition, the upper limit of the distorted signal power of the system is also derived. This is based on the mutual independence of the watermark signal and other mixed signals. Thus, the three-formula detection strengthens the security of the system with model uncertainty. Finally, we use a simulation example to demonstrate our theoretical results.

Key words Dynamic-watermarking (DWM), model uncertainty, distorted signal power, attack detection

Citation Du Da-Jun, Zhang Jing-Fan, Zhang Chang-Da, Fei Min-Rui, YANG Tai-Cheng. Robustness of dynamic-watermarking attack-detection method. *Acta Automatica Sinica*, 2023, **49**(12): 2557−2568

分布式通信网络在工业控制系统中大量的部署和应用, 给信息传输带来便利的同时, 也打破了传统控制系统的孤岛壁垒, 使得其从“封闭”走向“开放”, 导致网络化控制系统极易面临恶意攻击[1−4]. 中国经济周刊报道: 工业控制系统中受网络安全事件影响的企业占比达到了 28.6%, 造成工控网络停机的企业高达 19.1%, 带来极大的危害. 如 2010 年伊朗核电站遭受到震网病毒攻击[5]; 2014 年德国一家钢铁制造厂的熔炉控制系统遭受网络攻击, 导致整个生产线被迫停止运转; 同年流行于欧洲的新型木马病毒 Havex 对工控系统造成了重大损害[6]; 2015 年乌克兰国内多个区域的电网因遭受网络攻击而导致大规模停电[7]; 2019 年委内瑞拉国内的电网遭受网络攻击造成了 6 天的大规模停电事故[8]. 这些网络攻击不但带来大规模物理破坏, 而且经济损失惨重.

为了应对网络攻击, 目前国内外学者主要从攻击检测、攻击后的隔离与恢复、安全控制等角度展开研究[9−10]. 其中, 有助于迅速发现网络攻击的检测方法尤其受到国内外学者重视[11−12]. 根据防御者在

收稿日期 2020-08-03 录用日期 2020-11-04
Manuscript received August 3, 2020; accepted November 4, 2020
国家自然科学基金 (61773253, 61803252, 61633016, 61833011), 111 引智基地项目 (D18003), 上海市科委项目 (20JC1414000, 19500712300, 19510750300) 资助
Supported by National Natural Science Foundation of China (61773253, 61803252, 61633016, 61833011), 111 Project (D18003), and Project of Science and Technology Commission of Shanghai Municipality (20JC1414000, 19500712300, 19510750300)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 上海大学机电工程与自动化学院 上海 200072 中国 2. 上海市电站自动化技术重点实验室 上海 200072 中国 3. 萨赛克斯大学工程系 布莱顿 BN1 9QT 英国
1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China 2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China 3. Department of Engineering, University of Sussex, Brighton BN1 9QT, UK

设计检测机制时是否主动增加额外激励信号以助检测, 可将检测方法分为被动检测和主动检测方法.

被动检测方法根据已知的系统信息进行攻击检测, 并不干涉和影响原系统正常运行, 如针对服从高斯分布的网络攻击信号[13−15], 采用卡尔曼滤波−卡方法[16] 进行检测; 针对拒绝服务攻击, 采用基于分组接收速率分析法[17] 进行检测; 针对分布式系统遭受虚假数据注入攻击, 采用一致性分析法[18], 根据各子系统物理耦合、状态变量和控制决策之间相关性进行检测; 针对监视控制与数据采集系统 (Supervisory control and data acquisition, SCADA) 的网络攻击, 采用布谷鸟优化算法和神经网络[19] 进行检测. 然而, 最近一些攻击者精心构造网络攻击信号, 以躲过被动攻击检测机制, 进而破坏系统运行的稳定性和经济性[20−22]. 因此, 为了提高网络攻击检测能力, 主动检测技术应运而生.

主动检测方法不仅需要已知的系统信息, 而且主动增加额外的激励信号以助检测, 这将在一定程度上影响系统性能. 例如, 动态水印 (Dynamic watermarking, DWM) 技术属于典型的主动检测方法[23], 首先主动向系统增加一个私有激励信号 (水印信号), 然后根据水印信号的统计特征设计攻击检测机制, 进而提高攻击检测能力. 针对重放攻击, 在系统输入中增加水印信号, 然后根据水印信号的系统响应是否和预计的系统响应相一致以进行主动检测[23]; 针对线性时不变系统遭受网络攻击, 分别采用多水印[24]、周期水印[25]、乘性水印[26] 和递归水印[27] 等技术进行主动检测以提高检测精度; 在此基础上, 文献 [28] 针对确定性系统, 提出基于动态水印的新型攻击检测方法; 文献 [29] 提出任意噪声分布情况下的动态水印信号设计方法; 文献 [30] 将基于动态水印的网络攻击检测算法应用于自动发电系统中, 以提高系统安全性.

以上基于动态水印的主动检测方法主要针对确定性系统展开, 然而由于对实际系统特性缺乏足够了解或环境变化导致某些物理参数产生漂移等, 使得建立的模型往往具有不确定性. 这时基于误差方差与过程噪声方差一致性原理的渐进检测式显然不适用, 这是因为正常情况下模型不确定项的变化会导致检测式计算出的值偏离设定值, 从而导致检测式在正常情况下发生误判并持续报警, 使得系统无法正常运行.

为了解决以上问题, 本文的主要贡献如下: 1) 分析系统模型中不确定性因素导致传统水印检测式失效的原因, 然后考虑模型不确定项和过程噪声的统计规律, 将其影响转化为对方差变化特性进行分析, 提出两个具有鲁棒性的攻击检测式以及检测式中关键时变方差阈值的确定方法; 2) 采用系统失真信号功率定量刻画攻击信号造成系统性能损失程度, 理论证明了系统失真信号功率上界; 3) 基于水印信号与过程噪声信号和模型不确定项组成的混合信号相互独立的性质新增第三检测式, 理论证明了系统失真信号功率上界进一步受限范围, 进而提升系统的安全性.

1 问题描述

基于动态水印的主动检测框架如图 1 所示. 被控对象为模型不确定系统且受到高斯噪声 $w(k)$ 干扰; 系统输出 $y(k)$ 经过传感器测量后通过网络传输到缓存器, 由于 $y(k)$ 可能遭受网络攻击, 故到达缓存器时记为 $z(k)$. 缓存器一方面将 $z(k)$ 送到控制器计算得到控制信号 $u_c(k)$, 进一步为了提高攻击检测能力, 通过水印产生器主动向 $u_c(k)$ 中注入动态水印信号 $e(k)$, 进而形成带有水印的控制信号 $u(k)$, 最后传送到执行器实现对系统控制; 另一方面将 $z(k)$ 传送到攻击检测器, 同时运用 $u_c(k)$ 和 $e(k)$ 通过攻击检测机制 (即 3 个攻击检测式) 进行检测, 当其中任意一个检测结果超出设定阈值即判定系统遭受攻击.

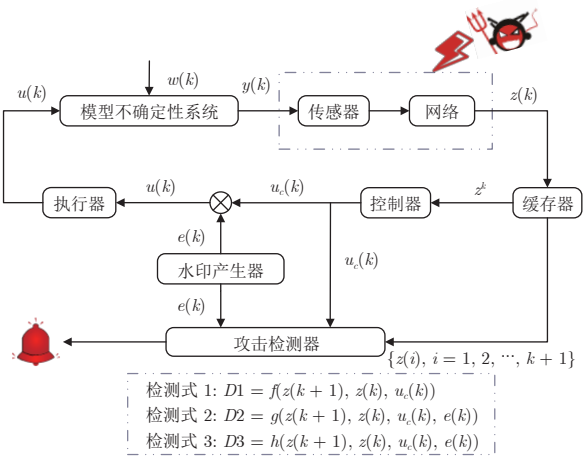


图 1 基于 DWM 的主动检测框架

Fig. 1 Active detection framework based on DWM

考虑图 1 中模型不确定系统的状态方程为

$$x(k+1) = (a + \Delta a(k))x(k) + bu(k) + w(k+1) \quad (1)$$

$$y(k) = x(k) \quad (2)$$

其中, $x(k)$ 是系统状态变量, $u(k)$ 是控制输入, $y(k)$ 是系统输出, $w(k+1)$ 是均值为零、方差为 Σ_w^2 的独立同分布高斯噪声; $\Delta a(k) = E_a \Lambda_a(k) F_a$ 为系统的

不确定性项[31], E_a 和 F_a 是反映模型不确定参数结构信息的已知常数, $\Lambda_a^2(k) \le 1$.

系统输出 $y(k)$ 通过传感器测量并经过网络以及缓存器传输到控制器为 $z(k)$, 则

$$u_c(k) = g_k(z^k) \tag{3}$$

其中, $z^k = (z(0), z(1), \cdots, z(k))$, $g_k(\cdot)$ 为系统正常运行的控制率, 如状态反馈控制、PID 控制等.

当系统输出 $y(k)$ 未遭受网络攻击时, 则 $z(k) = y(k)$, 控制器会计算出正确的控制率; 当 $y(k)$ 遭受网络攻击 (如虚假数据注入攻击等) 时, 则 $z(k) \ne y(k)$, 控制器计算出错误的控制率, 这势必影响控制系统性能. 因此, 需要设计攻击检测器以判断 $z(k)$ 是否遭受到网络攻击.

当控制器计算出控制信号 $u_c(k)$ 后, 水印生成器生成动态水印信号 $e(k)$, 主动注入到控制信号作为认证信号, 即

$$u(k) = u_c(k) + e(k) = g_k(z^k) + e(k) \tag{4}$$

其中, $e(k)$ 是均值为零、方差为 Σ_e^2 的独立同分布高斯随机信号.

将式 (4) 代入式 (1), 则系统闭环模型为

$$\begin{aligned} x(k+1) = {} & (a + \Delta a(k))x(k) + bu_c(k) + \\ & be(k) + w(k+1) \end{aligned} \tag{5}$$

当系统控制信号被注入动态水印信号后, 攻击者通常无法将过程噪声信号 $w(k)$ 和动态水印信号 $e(k)$ 分离. 在此基础上, 设计的攻击检测器能有效提高攻击检测率. 针对确定性系统, 传统动态水印检测根据系统闭环方程、水印信号和过程噪声均为高斯分布特性[28] 进行设计

$$\begin{aligned} & \{x(k+1) - ax(k) - bu_c(k)\}_k \sim \\ & \quad \text{i.i.d. N}(0, \Sigma_w^2 + b^2\Sigma_e^2) \end{aligned} \tag{6}$$

$$\begin{aligned} & \{x(k+1) - ax(k) - bu_c(k) - be(k)\}_k \sim \\ & \quad \text{i.i.d. N}(0, \Sigma_w^2) \end{aligned} \tag{7}$$

根据式 (6) 和式 (7) 中检测序列 $\{x(k+1) - ax(k) - bu_c(k)\}_k$ 和 $\{x(k+1) - ax(k) - bu_c(k) - be(k)\}_k$ 的方差是否与设定的方差值相一致, 可判定系统有无遭受网络攻击.

然而, 针对模型不确定系统, 根据闭环方程 (5), 可得

$$\begin{aligned} & x(k+1) - ax(k) - bu_c(k) = \\ & \quad \Delta a(k)x(k) + be(k) + w(k+1) \end{aligned} \tag{8}$$

$$\begin{aligned} & x(k+1) - ax(k) - bu_c(k) - be(k) = \\ & \quad \Delta a(k)x(k) + w(k+1) \end{aligned} \tag{9}$$

注 1. 与式 (6) 和式 (7) 不同, 式 (8) 和式 (9) 中含有的模型不确定项 $\Delta a(k)x(k)$ 导致其方差值不固定, 若仍然采用传统动态水印检测方法, 将使得检测结果错误, 进而导致检测失效, 这表明传统基于方差一致性的水印检测方法不适用于模型不确定系统. 因此, 需分析模型不确定项 $\Delta a(k)x(k)$ 对式 (8) 和式 (9) 中两个序列的方差影响, 并提出针对不确定系统的动态水印检测方法.

2 模型不确定系统的动态水印检测方法设计与分析

2.1 模型不确定系统网络攻击检测方法设计

根据上述分析可知, 在设计针对不确定性系统的动态水印攻击检测式时, 关键是要处理不确定项 $\Delta a(k)x(k)$ 的影响. 式 (2) 中 $x(k)$ 符合均值为零的高斯分布, 但不同于确定性系统中 $x(k)$ 的方差为固定值 Σ_x^2, 此时 $x(k)$ 的方差 $\Sigma_{x(k)}^2$ 是时变的. 进而可得模型不确定项 $\Delta a(k)x(k)$ 也符合均值为零的高斯分布, 时变方差符合 $(\Delta a(k))^2\Sigma_{x(k)}^2 = (E_a\Lambda_a(k)F_a)^2\Sigma_{x(k)}^2$. 在 k 时刻, $w(k+1)$ 和 $x(k)$ 是相互独立的, 则 $w(k+1)$ 和 $\Delta a(k)x(k)$ 也是相互独立的. 由于过程噪声信号和系统的模型不确定项均符合高斯分布, 故这两个独立的高斯分布信号叠加组成的混合信号依然符合零均值高斯分布, 该混合信号的方差为 $(E_a\Lambda_a(k)F_a)^2\Sigma_{x(k)}^2 + \Sigma_w^2$.

因此, 式 (8) 和式 (9) 在每一时刻的统计规律为

$$\begin{aligned} & \{x(k+1) - ax(k) - bu_c(k)\}_k \sim \\ & \quad \text{i.i.d. N}(0, (E_a\Lambda_a(k)F_a)^2\Sigma_{x(k)}^2 + \Sigma_w^2 + b^2\Sigma_e^2) \end{aligned} \tag{10}$$

$$\begin{aligned} & \{x(k+1) - ax(k) - bu_c(k) - be(k)\}_k \sim \\ & \quad \text{i.i.d. N}(0, (E_a\Lambda_a(k)F_a)^2\Sigma_{x(k)}^2 + \Sigma_w^2) \end{aligned} \tag{11}$$

注 2. 为了解决传统检测式的不足, 将模型不确定项 $\Delta a(k)x(k)$ 的影响转化为对方差进行分析, 进而根据方差范围大小重新设计攻击检测式. 根据 $\Lambda_a(k)$ 的范围 $\Lambda_a^2(k) \le 1$, 确定 $\Delta a(k)x(k)$ 的方差范围 $0 \le (\Delta a(k))^2\Sigma_{x(k)}^2 \le (E_aF_a)^2\Sigma_{x(k)}^2$. 同理可知, 混合信号 $w(k+1) + \Delta a(k)x(k)$ 的方差满足不等式 $\Sigma_w^2 \le (E_a\Lambda_a(k)F_a)^2\Sigma_{x(k)}^2 + \Sigma_w^2 \le (E_aF_a)^2\Sigma_{x(k)}^2 + \Sigma_w^2$.

基于式 (10) 和式 (11), 将模型不确定项和过程噪声统计规律的影响转化为对方差变化特性进行分析, 设计了如下两个具有鲁棒性的攻击检测式:

检测式 1 (Test 1).

$$\begin{aligned} & D1 = f(z(k+1), z(k), u_c(k)) = \\ & \quad \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z(k+1) - az(k) - bu_c(k))^2 \end{aligned}$$

正常情况下, $D1$ 需满足

$$\Sigma_w^2 + b^2\Sigma_e^2 \le D1 \le (E_aF_a)^2\Sigma_\infty^2 + \Sigma_w^2 + b^2\Sigma_e^2 \quad (12)$$

检测式 2 (Test 2).

$$D2 = g(z(k+1),\, z(k),\, u_c(k),\, e(k)) = \lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(z(k+1)-az(k)-bu_c(k)-be(k))^2$$

正常情况下, $D2$ 需满足

$$\Sigma_w^2 \le D2 \le (E_aF_a)^2\Sigma_\infty^2 + \Sigma_w^2 \quad (13)$$

其中, Σ_∞^2 代表 $x(k)$ 的时变方差阈值.

当 $z(k+1)$ 同时满足 Test 1 和 Test 2 时, 则认为系统正常运行; 否则认为系统遭受到攻击, 攻击检测器将会触发报警. 从 Test 1 和 Test 2 可以看出, 相比较传统水印检测式无法适用于不确定系统的问题, 本文将其转换成具有一定鲁棒性的不等式, 避免了传统检测式将系统正常运行误判为攻击的行为, 从而提升模型不确定系统的攻击检测鲁棒性.

注 3. Test 1 和 Test 2 中的时变方差阈值 Σ_∞^2 的选择直接影响检测性能. 当阈值 Σ_∞^2 过小时, 不等式范围相应变小, 将出现模型不确定项 $\Delta a(k)x(k)$ 的变化被误认为是攻击的情况; 当阈值 Σ_∞^2 较大时, 不等式范围相应变大, 将出现攻击无法被有效检测出的情况, 因此合理选择阈值 Σ_∞^2 至关重要.

阈值 Σ_∞^2 的选取与 $x(k)$ 在无穷时刻的方差有关, 随着时间趋于无穷, $x(k)$ 收敛, 其方差将趋于稳定. 因此, 当控制率确定时, 可通过计算状态变量在无穷时刻的方差 $\Sigma_{x(\infty)}^2$, 选择合适的阈值 Σ_∞^2. 例如: 当 $u_c(k) = Kz(k)$ 时, 其中 K 是反馈增益, 系统闭环方程为

$$x(k+1) = (a+\Delta a(k)+bK)x(k) + be(k) + w(k+1) \quad (14)$$

对式 (14) 两边进行求方差计算, 可得

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(x(k+1))^2 = \\ &\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}((a+\Delta a(k)+bK)x(k))^2 + \\ &2\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(a+\Delta a(k)+bK)x(k)be(k) + \\ &2\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(a+\Delta a(k)+bK)x(k)w(k+1) + \\ &2\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}be(k)w(k+1) + b^2\Sigma_e^2 + \Sigma_w^2\end{aligned}$$

由于状态变量、动态水印信号以及过程噪声信号之间相互独立, 上式进一步化简, 可得

$$\Sigma_{x(\infty)}^2 = \left|\frac{b^2\Sigma_e^2 + \Sigma_w^2}{1 - \lim\limits_{T\to\infty}\frac{1}{T}\sum\limits_{k=0}^{T-1}(a+\Delta a(k)+bK)^2}\right| \quad (15)$$

注 4. 当 $\Delta a(k)$ 的分布已知时, 可通过式 (15) 计算得到 $\Sigma_{x(\infty)}^2$, 并将其记作阈值 Σ_∞^2; 当 $\Delta a(k)$ 的分布未知时, 通过选择 $\Delta a(k)$ 使得方差 $\Sigma_{x(\infty)}^2$ 的值最大并记作阈值 Σ_∞^2, 但这会导致检测式的保守性较强. 因此, 对不确定性了解得越多, 构造的检测式保守性越小, 检测效果也越好.

2.2 网络攻击造成的失真信号功率分析

针对模型不确定系统, 采用 Test 1 和 Test 2 能够提高检测率. 然而, 在最坏情况下, 攻击者若能够躲过检测, 则攻击造成的失真信号功率的大小以及对系统造成破坏的程度将是一个亟待解决的问题.

为了解决上述问题, 首先定义每一时刻网络攻击造成的系统失真信号为

$$v(k+1) = z(k+1) - az(k) - \Delta a(k)z(k) - bu_c(k) - be(k) - w(k+1) \quad (16)$$

当系统未遭受网络攻击时, $z(k) = y(k) = x(k)$, 此时失真信号 $v(k) = 0$; 当输出数据遭受网络攻击被篡改时, 则 $v(k) \ne 0$.

为了定量评价攻击信号对系统性能损失程度, 采用信号功率对失真信号进行刻画, 网络攻击造成的系统失真信号功率定义为

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\|v(k)\|^2 \quad (17)$$

式 (17) 反映了攻击信号造成系统性能损失的大小, 功率越大代表网络攻击造成系统性能损失越大. 基于 Test 1 和 Test 2, 并结合系统失真信号功率定义, 可得定理 1, 其定量给出了系统失真信号功率的上界.

定理 1. 当模型不确定系统遭受网络攻击并躲过 Test 1 和 Test 2 时, 攻击造成的系统失真信号功率为

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\|v(k)\|^2 \le \frac{(E_aF_a)^2\Sigma_\infty^2 + \Sigma_w^2 + b^2\Sigma_e^2}{b^2\Sigma_e^2}(E_aF_a)^2\Sigma_\infty^2 \quad (18)$$

证明. 若 $z(k)$ 躲过攻击检测, 则需满足 Test 1:

$$\Sigma_w^2+b^2\Sigma_e^2\le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v(k)+l(k)+be(k-1)\right)^2= \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v(k)+l(k)\right)^2+ b^2e^2(k-1)+2be(k-1)(v(k)+l(k))\le (E_aF_a)^2\Sigma_\infty^2+\Sigma_w^2+b^2\Sigma_e^2 \tag{19}$$

其中, $l(k)=\Delta a(k)z(k-1)+w(k)$ 表示模型不确定项和噪声构成的混合信号, 其方差需满足

$$\Sigma_w^2\le\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}l^2(k)\le(E_aF_a)^2\Sigma_\infty^2+\Sigma_w^2$$

同理, $z(k)$ 也需满足 Test 2:

$$\Sigma_w^2\le\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v(k)+l(k)\right)^2\le (E_aF_a)^2\Sigma_\infty^2+\Sigma_w^2 \tag{20}$$

为了进一步简化证明过程, 定义 3 个新的变量: $\varepsilon_1(k)$, $\varepsilon_2(k)$, $\varepsilon_3(k)$, 分别为

$$\varepsilon_1(k)=(E_aF_a)^2\Sigma_\infty^2+\Sigma_w^2-\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}l^2(k)\ge 0 \tag{21}$$

$$\varepsilon_2(k)=(E_aF_a)^2\Sigma_\infty^2+\Sigma_w^2- \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v(k)+l(k)\right)^2\ge 0 \tag{22}$$

$$-(E_aF_a)^2\Sigma_\infty^2\le\varepsilon_3(k)= \Sigma_w^2-\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v(k)+l(k)\right)^2\le 0 \tag{23}$$

进一步, 将 $\varepsilon_1(k)$ 与 $\varepsilon_2(k)$ 相减, 可得

$$\varepsilon_1(k)-\varepsilon_2(k)=\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(v^2(k)+2v(k)l(k)\right) \tag{24}$$

将 $\varepsilon_2(k)$ 与 $\varepsilon_3(k)$ 代入式 (20), 可得

$$\frac{1}{2}\varepsilon_3(k)\le\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e(k-1)[v(k)+l(k)]\le\frac{1}{2}\varepsilon_2(k) \tag{25}$$

由于状态变量、动态水印信号以及噪声信号之间相互独立, 故 $e(k-1)$ 与 $l(k)$ 相互独立, 则可将式 (25) 化简为

$$\frac{1}{2}\varepsilon_3(k)\le\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e(k-1)v(k)\le\frac{1}{2}\varepsilon_2(k) \tag{26}$$

式 (26) 表明了水印信号与失真信号之间的相关性, 并且其上界值和下界值满足: $\varepsilon_2(k)/2-\varepsilon_3(k)/2=(E_aF_a)^2\Sigma_\infty^2/2$, 即两个信号的相关性大小和模型不确定项大小有关, 当模型不确定项取值较小时, 两个信号之间的相关性相对较弱. 如果系统无模型不确定性时 ($E_a=F_a=0$), 则水印信号和失真信号不相关.

接下来, 针对模型不确定项和噪声构成的混合信号 $l(k)=\Delta a(k)z(k-1)+w(k)=x(k)-ax(k-1)-bg_{k-1}(z^{k-1})-be(k-1)$, 设一个可测集 $S_k=\sigma(x^k, z^k, e^{k-2})$ 进行描述, 即

$$(x^{k-2}, e^{k-2})\to(x(k-1), x(k), z^k)\to l(k)$$

由于攻击者很难分离混合信号和水印信号, 故通过 $\hat{l}(k)=\mathrm{E}[l(k)|S_k]$ 对 $l(k)$ 进行估计. 然而, 由于攻击者无法获得水印信号, 故该估计由 $\hat{l}(k)=\mathrm{E}[l(k)|\sigma(x^{k-2}, e^{k-2}, z^k, x(k-1), x(k))]$ 变为 $\hat{l}(k)=\mathrm{E}[l(k)|\sigma(x(k-1), x(k), z^k)]$.

攻击者能获得混合信号和水印信号的组合 $l(k)+be(k-1)=x(k)-ax(k-1)-bg_{k-1}(z^{k-1})$, 并且 $l(k)+be(k-1)$ 在每一时刻符合独立同高斯分布, 故通过条件估计获得 $\hat{l}(k)$, 条件估计分布为

$$f(l(k)|be(k-1)+l(k))=\frac{f(l(k), be(k-1)+l(k))}{f(be(k-1)+l(k))}$$

定义 Σ_Δ^2 为混合信号 $l(k)$ 的方差, 通过计算估计信号 $\hat{l}(k)$ 的概率分布密度, 可得

$$\hat{l}(k)=\frac{\Sigma_\Delta^2}{\Sigma_\Delta^2+b^2\Sigma_e^2}(be(k-1)+l(k))= \beta(be(k-1)+l(k)) \tag{27}$$

其中,

$$\beta=\frac{\Sigma_\Delta^2}{\Sigma_\Delta^2+b^2\Sigma_e^2}= \frac{\Sigma_w^2+(E_aF_a)^2\Sigma_{x(k)}^2}{\Sigma_w^2+(E_aF_a)^2\Sigma_{x(k)}^2+b^2\Sigma_e^2}<1$$

混合信号 $l(k)$ 的方差 Σ_Δ^2 并非确定不变, 由式 (15) 可知, 当已知 $\Delta a(k)$ 分布时才可获得无穷时刻混合信号 $l(k)$ 的方差. 但 Σ_Δ^2 可以从混合信号的方差范围 $\Sigma_w^2\sim(E_aF_a)^2\Sigma_{x(k)}^2+\Sigma_w^2$ 内取值, 对于攻击者估计混合信号 $l(k)$ 来讲是可行的.

令 $\tilde{l}(k)=l(k)-\hat{l}(k)$ 表示实际混合信号和估计信号之间的差值信号, 则差值信号 $\tilde{l}(k)$ 的期望为 0. 同时, $l(k)$ 与 $\hat{l}(k)$ 均由观测集 S_k 获得, 所以有 $\tilde{l}(k-1)\in S_k$, 差值信号在可测集合下的期望值为零: $\mathrm{E}[\tilde{l}(k)|S_k]=0$. 因此, 根据鞅差分序列定义, $\tilde{l}(k-1)$ 是关

于 S_k 的鞅差分序列. 另一方面, 失真信号 $v(k) \in S_k$ ($v(k)$ 可由 $\sigma(x^k, z^k)$ 得到), 根据文献 [32] 提出的鞅稳定性, 可得

$$\sum_{k=1}^{T} v(k)\tilde{l}(k) = o\left(\sum_{k=1}^{T} v^2(k)\right) + O(1) \tag{28}$$

则

$$\sum_{k=1}^{T} v(k)l(k) = \sum_{k=1}^{T} v(k)\hat{l}(k) + o\left(\sum_{k=1}^{T} v^2(k)\right) + O(1)$$

将式 (27) 代入上式, 进一步有

$$\sum_{k=1}^{T} v(k)l(k) = \frac{\beta}{1-\beta}\sum_{k=1}^{T} v(k)be(k-1) + \frac{\beta}{1-\beta}o\left(\sum_{k=1}^{T} v^2(k) + \frac{\beta}{1-\beta}O(1)\right) \tag{29}$$

对上式两边同时除以 T, 并取极限有

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v(k)l(k) = \frac{\beta}{1-\beta}\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v(k)be(k-1) \tag{30}$$

将式 (30) 代入式 (26), 可得

$$\frac{\beta}{1-\beta}\varepsilon_3(k) \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} 2v(k)l(k) \le \frac{\beta}{1-\beta}\varepsilon_2(k) \tag{31}$$

进一步, 将式 (24) 代入式 (31), 可得

$$\varepsilon_1(k) - \varepsilon_2(k) - \frac{\beta}{1-\beta}\varepsilon_2(k) \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2(k) \le \varepsilon_1(k) - \varepsilon_2(k) - \frac{\beta}{1-\beta}\varepsilon_3(k) \tag{32}$$

式 (32) 中不等式右边为

$$\varepsilon_1(k) - \varepsilon_2(k) - \frac{\beta}{1-\beta}\varepsilon_3(k) = \varepsilon_1(k) - \varepsilon_2(k) + \varepsilon_3(k) - \frac{1}{1-\beta}\varepsilon_3(k)$$

为了进一步简化证明过程, 定义一个新的变量 Π, 则

$$\Pi = \frac{\Sigma_w^2 + (E_aF_a)^2\Sigma_\infty^2 + b^2\Sigma_e^2}{-b^2\Sigma_e^2} \le -\frac{1}{1-\beta} < -1$$

由于 $\varepsilon_3(k) \le 0$, 所以 $-\frac{1}{1-\beta}\varepsilon_3(k) \le \Pi\varepsilon_3(k)$, 即

$$\begin{aligned}
&\varepsilon_1(k) - \varepsilon_2(k) + \varepsilon_3(k) - \frac{1}{1-\beta}\varepsilon_3(k) \le \\
&\varepsilon_1(k) - \varepsilon_2(k) + \varepsilon_3(k) + \Pi\varepsilon_3(k) = \\
&\Sigma_w^2 + \Pi\varepsilon_3(k) - \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} l^2(k) \le \\
&\Sigma_w^2 - \Pi(E_aF_a)^2\Sigma_\infty^2 - \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} l^2(k) \le \\
&\Sigma_w^2 - \Pi(E_aF_a)^2\Sigma_\infty^2 - \Sigma_w^2 = \\
&\frac{\Sigma_w^2 + (E_aF_a)^2\Sigma_\infty^2 + b^2\Sigma_e^2}{b^2\Sigma_e^2}(E_aF_a)^2\Sigma_\infty^2
\end{aligned}$$

另一方面, 当系统无模型不确定性时[8], 则

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2(k) = 0$$

当系统存在模型不确定性时, 则

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2(k) \ge 0$$

因此, 可得

$$0 \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} \|v(k)\|^2 \le \frac{(E_aF_a)^2\Sigma_\infty^2 + \Sigma_w^2 + b^2\Sigma_e^2}{b^2\Sigma_e^2}(E_aF_a)^2\Sigma_\infty^2 \tag{33}$$

□

注 5. 采用动态水印主动检测方法后, 利用攻击者无法分离出水印信号这一特性, 攻击检测器根据采集数据形成等式, 即

$$z(k+1) - az(k) - bg_k(z^k) - be(k) = \Delta a(k)x(k) + w(k+1) + v(k)$$

该式表明, 序列 $\{\Delta a(k)x(k) + w(k+1) + v(k)\}$ 可以计算获得, 将该序列在正常情况下的统计规则作为标准, 以判断系统是否遭受到攻击. 当系统正常时, 序列 $\{\Delta a(k)x(k) + w(k+1) + v(k)\}$ 中的 $v(k)$ 为零; 当攻击存在时, $v(k)$ 则不恒等于 0. 然而在最坏情况下, 攻击者能够躲过检测, 此时 $v(k)$ 不恒等于 0, 会对系统造成一定的损害. 定理 1 定量给出攻击造成的失真信号功率的上界, 展示了两个检测式对攻击造成系统失真信号功率的限制能力, 这也表明了两个检测式具有抑制攻击对系统造成破坏的能力, 同时还表明当模型确定时, 式 (33) 中失真信号的功率将恒等于 0.

2.3 改进的攻击检测策略

当只使用 Test 1 和 Test 2 进行攻击检测时,

不可避免地存在以下一些不足:

1) 当模型不确定存在时, 若阈值选择不当, 躲过攻击检测的失真信号功率上界有时会过大, 这表明在最坏情况下躲过检测的攻击将会对系统造成比较大的损害;

2) 当无法保证水印信号 $e(k-1)$ 和失真信号 $v(k)$ 无关时, 攻击者能够使用包含水印信号信息的数据构造攻击, 以通过 Test 1 和 Test 2;

3) 当无法保证混合信号 $l(k)$ 和失真信号 $v(k)$ 无关时, 攻击者能够使用包含混合信号信息的数据构造攻击, 以通过 Test 1 和 Test 2;

4) 当对模型不确定性了解较少时, Test 1 和 Test 2 的阈值选择的保守性增加, 会出现漏报攻击情况, 导致检测失效.

由于以上缺陷的存在, 攻击者能够构造新的攻击信号躲避攻击检测, 且对系统可能造成的破坏相对较大, 需添加新的检测式进行解决. 因为水印信号并不依赖系统状态变量和过程噪声而产生, 所以水印信号和混合信号必然是独立的. 当攻击者篡改输出信号 $z(k)$ 时, 有可能改变水印信号和混合信号之间相关性, 故需增加第三检测式进行检测.

检测式 3 (Test 3).

$$D3 = h(z(k+1),\, z(k),\, u_c(k),\, e(k)) = \lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} e(k)(z(k+1) - az(k) - bu_c(k) - be(k))$$

正常情况下, $D3$ 需满足: $D3 = 0$.

当 $z(k)$ 同时满足 Test 1 ~ Test 3 时, 则认为系统正常运行; 否则认为系统遭受到攻击, 攻击检测器将会触发报警. Test 3 作为 Test 1 和 Test 2 的补充, 能提升攻击检测能力, 接下来将通过定理 2 分析 Test 3 如何解决 Test 1 和 Test 2 的不足, 结合系统失真信号功率定义, 进一步定量限制系统失真信号功率的上界.

定理 2. 当模型不确定系统遭受网络攻击并躲过 Test 1 ~ Test 3 时, 攻击造成的系统失真信号功率为

$$0 \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\|v(k)\|^2 \le (E_aF_a)^2\left(\Sigma_\infty^2 - \Delta_a^2(\infty)\Sigma_{z(\infty)}^2\right) \tag{34}$$

证明. 在定理 1 的基础上进一步推导, 若 $z(k)$ 躲过攻击检测, 则需满足 Test 3, 即

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} e(k)\left(v(k)+l(k)\right) = 0 \tag{35}$$

水印信号 $e(k-1)$ 与混合信号 $l(k)$ 对于 $\forall k$ 相互独立, 则可将上式化简为

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} e(k-1)v(k) = 0 \tag{36}$$

式 (36) 表明水印信号与失真信号是无关的, 即攻击所造成的失真信号不能带有与水印相关的信息, 否则将无法通过 Test 3. 同时, 失真信号与混合信号之间的相关性也随之发生改变, 即定理 1 证明中的式 (29) 变为

$$\sum_{k=1}^{T} v(k)l(k) = \frac{\beta}{1-\beta}o(T) + \frac{\beta}{1-\beta}o\left(\sum_{k=1}^{T} v^2(k)\right) + \frac{\beta}{1-\beta}O(1) \tag{37}$$

对上式两边同时除以 T, 并取极限

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v(k)l(k) = 0 \tag{38}$$

即混合信号与失真信号无关, 结合式 (19) 和式 (20), 则

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2(k) \ge 0$$

进一步, 可得

$$0 \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2(k) \le \varepsilon_1(k) \le (E_aF_a)^2\Sigma_\infty^2 + \Sigma_w^2 - \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\left(E_a\Lambda_a(\infty)F_az(k) + w(k+1)\right)^2 = (E_aF_a)^2(\Sigma_\infty^2 - \Lambda_a^2(\infty)\Sigma_{z(\infty)}^2) \tag{39}$$

□

注 6. 定理 2 表明新增的 Test 3 能解决 Test 1 和 Test 2 的不足, 即能够对包含水印信号或混合信号信息的攻击进行有效检测, 同时证明了系统失真信号功率上界进一步受限范围, 即大大限制了躲过检测式的失真信号功率上界, 该上界值和阈值 Σ_∞^2 的选择紧密相关, 并且 Σ_∞^2 与 $\Lambda_a^2(\infty)\Sigma_{z(\infty)}^2$ 之差越小, 说明阈值 Σ_∞^2 越接近真实值, 失真信号功率的上界值越小, 越能更好地限制攻击. 即使在最坏情况下, 攻击者也能够躲过检测, 但是定理 2 给出了攻击造成的失真信号功率的上界, 这表明 Test 1 ~ Test 3 对攻击造成的系统失真信号功率具有限制能力, 抑制了攻击对系统造成的破坏. 特别地, 当模型确定时, 代数差为零, 失真功率也等于零, 这表明模型不确定性信息越明确, 防御者对阈值 Σ_∞^2 的选择

将会越合适, 最终限制攻击的效果将更好, 反之则可能会选择偏离程度较大的阈值 Σ_∞^2, 限制攻击的效果或许不尽人意.

定理 2 表明模型不确定系统中动态水印技术对攻击信号造成系统性能损失程度的限制, 进一步以最为常见的虚假数据注入攻击信号为例, 表现攻击信号本身也会受到动态水印技术的限制. 若虚假数据注入攻击中的虚假数据信号为 $m(k)=z(k)-x(k)$, 当该攻击能够同时躲过 Test 1 ~ Test 3 时, 虚假数据信号 $m(k)$ 的功率需满足

$$0 \le \lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\|m(k)\|^2 \le (E_aF_a)^2\left(\Sigma_\infty^2-\Lambda_a^2(\infty)\Sigma_{z(\infty)}^2\right) \tag{40}$$

式 (40) 的上界值与式 (34) 一样, 都取决于对阈值 Σ_∞^2 的选择. 当模型不确定性信息越明确, 防御者对阈值 Σ_∞^2 的选择将会越合适, 最终限制攻击的效果将更好. 特别地, 当模型确定时, 虚假数据攻击信号功率将被限制为 0.

3 仿真实验

为了验证本文所提出方法的可行性和有效性, 通过数值和实例仿真进行验证.

3.1 数值仿真

由于 Test 1 ~ Test 3 是在无穷时间统计下的渐进表达式, 故采用工程实际中常用的窗口检测统计方法, 转换为有限时间内的统计形式进行实验. 在本实验中, 窗口大小设为 1 000 个采样时刻, 持续时间为 20 000 个采样时刻.

3.1.1 Test 1 和 Test 2 的可行性和有效性

首先验证 Test 1 和 Test 2 的可行性和有效性, 考虑如下模型不确定系统

$$x(k+1)=0.4x(k)+\Delta a(k)x(k)+u_c(k)+e(k)+w(k+1) \tag{41}$$

$$y(k)=x(k) \tag{42}$$

其中, $\Delta a(k)x(k)=E_a\Lambda_a(k)F_ax(k)$ 为模型不确定项, 设 $E_a=0.3$, $F_a=1.5$, $\Lambda_a(k)$ 为 $-1\sim1$ 的均匀分布. 水印信号 $e(k)$ 的方差为 $\Sigma_e^2=0.01$, 过程噪声 $w(k+1)$ 方差为 $\Sigma_w^2=0.01$, 控制率采用 $u_c(k)=0.1z(k)$.

若根据基于式 (6) 和式 (7) 的传统水印检测方法对系统进行攻击检测, 正常情况下两个传统水印检测式的方差值分别设为 0.02 和 0.01, 系统运行结果如图 2 和图 3 所示. 由图 2 和图 3 中可以看出, 当系统正常运行时, 模型不确定项的变化会引起检测式计算出的值偏离阈值, 导致传统水印检测式在正常情况下发生误判并持续报警, 从而使得系统无法正常运行.

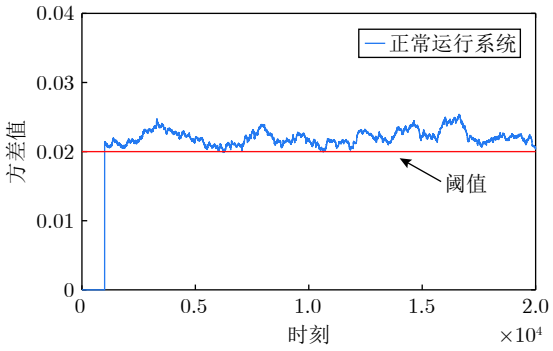


图 2 基于式 (6) 的传统水印检测结果

Fig. 2 Traditional watermark detection results based on (6)

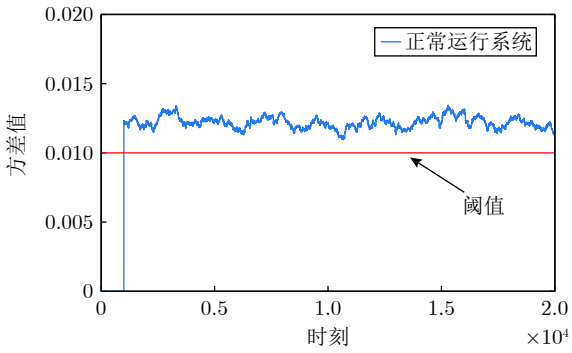


图 3 基于式 (7) 的传统水印检测结果

Fig. 3 Traditional watermark detection results based on (7)

为了解决以上传统水印检测方法无法适用问题, 采用本文所提的两个具有鲁棒性的检测式, 其考虑模型不确定项 $\Delta a(k)x(k)$ 带来的影响, 设定合适的阈值, 能够避免以上误报发生. 当控制率确定时, 根据式 (15) 选择的阈值 Σ_∞^2 为

$$\Sigma_{x(\infty)}^2=\left|\frac{0.01+0.01}{1-\lim\limits_{T\to\infty}\dfrac{1}{T}\sum\limits_{k=0}^{T-1}\left(0.5+0.45\Lambda_a(k)\right)^2}\right|$$

因为已知 $\Lambda_a(k)$ 符合 $-1\sim1$ 的均匀分布, 故可选取阈值: $\Sigma_\infty^2=\Sigma_{x(\infty)}^2=0.029$, 则 $(E_aF_a)^2\Sigma_\infty^2=0.0059$, 所以 Test 1 的上限值设为 0.0259, 下限值设为 0.02; Test 2 的上限值设为 0.0159, 下限值设为 0.01. 采用 Test 1 和 Test 2 进行攻击检测, 系统

正常运行时的结果如图 4 和图 5 所示. 由图 4 和图 5 中可以发现, 模型不确定项变化引起的检测值 $D1$ 和 $D2$ 变化不会出现持续报警情况, 从而使得系统能够正常运行.

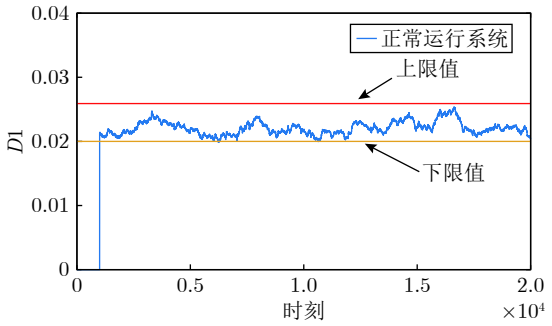


图 4　基于 Test 1 的检测结果

Fig. 4　Detection results based on Test 1

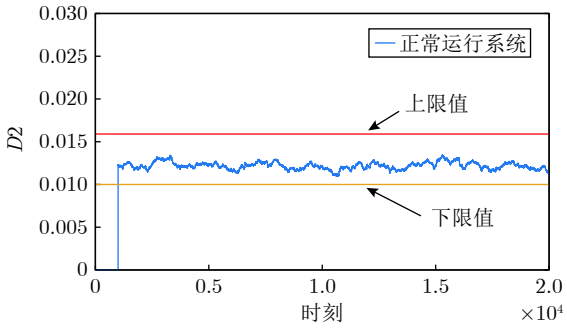


图 5　基于 Test 2 的检测结果

Fig. 5　Detection results based on Test 2

3.1.2　Test 3 的可行性和有效性

当对模型不确定性了解较少时, Test 1 和 Test 2 的方差阈值选择不一定合适, 则攻击将能够躲过检测, 接下来验证 Test 3 的有效性. 考虑如下模型不确定性系统:

$$x(k+1) = 0.1x(k) + \Delta a(k)x(k) + u_c(k) + e(k) + w(k+1) \tag{43}$$

$$y(k) = x(k) \tag{44}$$

其中, $\Delta a(k)x(k) = E_a\Lambda_a(k)F_a x(k)$ 为模型不确定项, 设 $E_a = 2$, $F_a = 0.4$, 不确定项 $\Lambda_a(k)$ 分布未知. 水印信号 $e(k)$ 的方差为 $\Sigma_e^2 = 0.01$, 过程噪声 $w(k+1)$ 的方差为 $\Sigma_w^2 = 0.01$, 控制率采用 $u_c(k) = 0.1z(k)$.

在未知不确定项 $\Lambda_a(k)$ 分布下, 选取阈值的方法有两种: 1) 根据式 (15), 选择 $\Delta a(k)$ 使得方差 $\Sigma_{x(\infty)}^2$ 最大值作为阈值 Σ_∞^2, 但这会使得检测式的保守性较强; 2) 采用系统正常运行一段时间的状态变量方差值作为阈值 Σ_∞^2. 本实验统计持续运行时间 10 000 个时刻的状态变量方差值均值, 最终设 Test 1 的上限值为 0.039, 下限值为 0.02; Test 2 的上限值为 0.029, 下限值为 0.01.

本实验中采用虚假数据注入攻击, 将虚假数据注入到真实数据 $z(k)$ 中, 构造的虚假数据信号为 $m(k) = n(k) - 0.1\hat{l}(k) - 0.13z(k)$, 其中 $n(k)$ 是均值为零、$\Sigma_n^2 = 0.01$ 的高斯信号, $\hat{l}(k)$ 是根据式 (27) 得到的混合信号 $l(k)$ 的估计 $\hat{l}(k+1) = 0.7\,(z(k+1) - 0.6z(k))$.

当攻击者估计出每一时刻的混合信号后, 便可构造虚假数据注入攻击信号 $m(k)$. 在 $k = 8\,500$ 时, 将虚假数据信号开始注入到系统中, 传感器输出数据被篡改为 $z(k) = y(k) + m(k)$. 接着, 采用 Test 1 ~ Test 3 对 $z(k)$ 进行检测, 检测结果如图 6 ~ 8 所示.

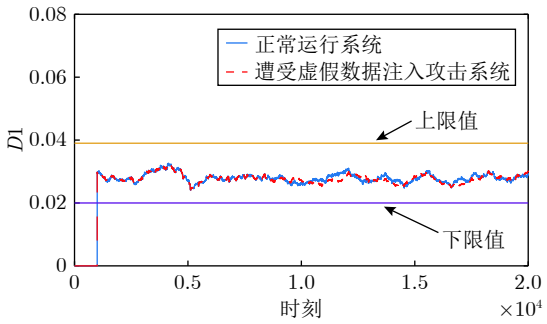


图 6　基于 Test 1 的虚假数据注入攻击检测结果 (数值仿真)

Fig. 6　Detection results based on Test 1 under false data injection attack (numerical simulation)

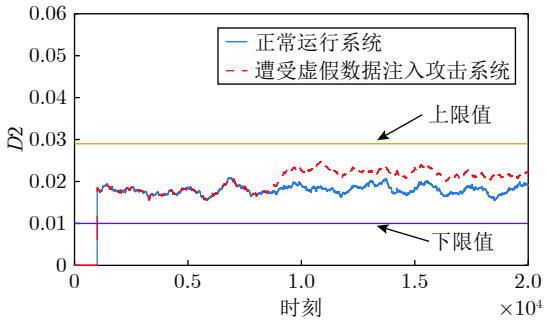


图 7　基于 Test 2 的虚假数据注入攻击检测结果 (数值仿真)

Fig. 7　Detection results based on Test 2 under false data injection attack (numerical simulation)

由图 6 和图 7 可以看出, 当虚假数据注入攻击发生时, 检测值 $D1$ 和 $D2$ 将无法超过上限值, 即 Test 1 和 Test 2 不能够有效检测出攻击. 图 8 表明 Test 3 的结果与正常运行系统结果在攻击发生时有明显偏离, 即 Test 3 能有效检测攻击并触发警报. 因此, Test 3 作为 Test 1 和 Test 2 的补充, 能够解决这两个检测式带来的不足.

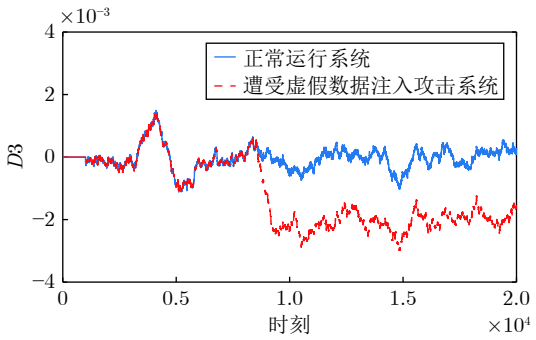


图 8 基于 Test 3 的虚假数据注入攻击检测结果 (数值仿真)

Fig. 8 Detection results based on Test 3 under false data injection attack (numerical simulation)

根据定理 1, 当攻击能够躲过 Test 1 和 Test 2 时, 设定系统失真信号功率阈值为 0.0741, 系统失真信号功率变化如图 9 所示. 图 9 展示了系统失真信号功率的变化, 表明当攻击能够躲过 Test 1 和 Test 2 时, 攻击信号造成的失真信号功率小于定理 1 设定的阈值, 即 Test 1 和 Test 2 在一定程度上抑制了攻击所能造成的破坏.

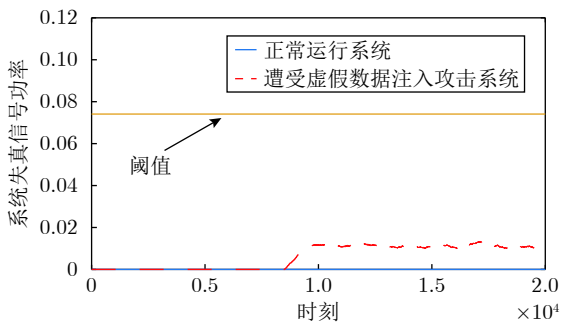


图 9 系统失真信号功率变化

Fig. 9 Variation of system distortion signal power

3.2 实例仿真

工业电加热炉系统[33] 的温度特性建模为

$$x(k+1)=(0.85+\Delta a(k))x(k)+0.154u(k-1)+w(k+1) \tag{45}$$

$$y(k)=x(k) \tag{46}$$

其中, $x(k)$ 是系统状态变量 (即 k 时刻的温度), $y(k)$ 是系统输出, $w(k+1)$ 表示均值为零、方差为 $\Sigma_w^2=1$ 的独立同分布高斯过程噪声; $\Delta a(k)=E_a\Lambda_a(k)F_a$ 为系统的模型不确定性, 设 $E_a=0.2$, $F_a=0.02$, $\Lambda_a(k)$ 为 $-1\sim1$ 的均匀分布; $u(k)$ 采用 PI 控制.

采用动态水印策略, 则闭环系统模型为

$$x(k+1)=(0.85+\Delta a(k))x(k)+0.154(u_c(k-1)+e(k))+w(k+1) \tag{47}$$

其中, $e(k)$ 是动态水印信号, 其均值为零、方差为 $\Sigma_e^2=16$.

为了验证本文提出的方法能够提高网络攻击检测能力, 本实验采用虚假数据注入攻击, 即攻击信号为 $m(k)=n(k)-\hat{l}(k)$, 其中 $n(k)$ 是均值为零、$\Sigma_n^2=4$ 的高斯信号, 混合信号为 $\hat{l}(k+1)=0.7(z(k+1)-0.85z(k)-0.154u(k-1))$, 并且与文献 [34] 中的阈值检测方法进行对比分析. 文献 [34] 中阈值检测方法的检测式为 $\left\|\begin{bmatrix}z(k)\\u(k)\end{bmatrix}\right\|>\theta$, 即遭受到攻击, 否则未遭受攻击, 在此设 $\theta=310$. 本文所提方法 Test 1 上限值设为 2.4, 下限值设为 1.4; Test 2 上限值设为 2.1, 下限值设为 1. 实验中窗口大小设定为 500 个采样时刻, 持续时间为 10 000 个采样时刻.

在图 10 ~ 12 中, 当 $k=5\,000$ 时, 攻击者将虚假数据信号注入到系统中, 即传感器输出数据被篡

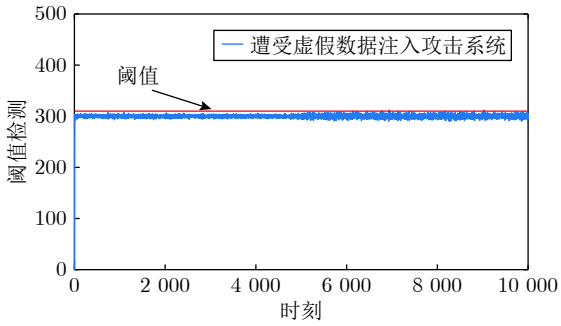


图 10 基于阈值检测的虚假数据注入攻击检测结果 (实例仿真)

Fig. 10 Detection results based on threshold detection under false data injection attack (real example simulation)

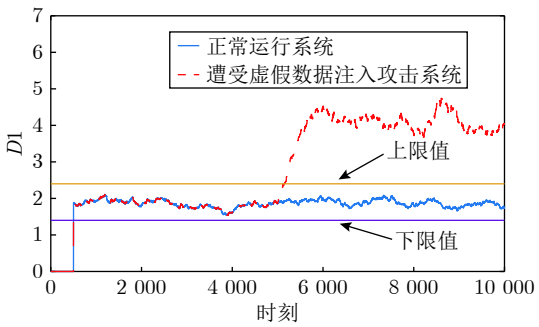


图 11 基于 Test 1 的虚假数据注入攻击检测结果 (实例仿真)

Fig. 11 Detection results based on Test 1 under false data injection attack (real example simulation)

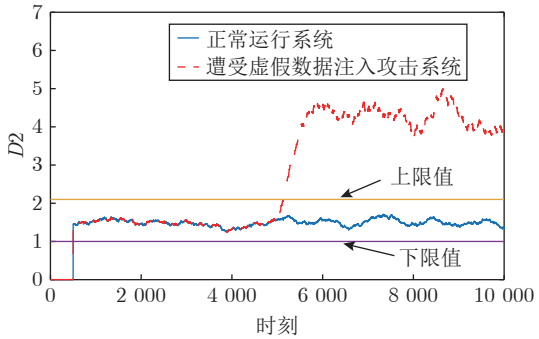


图 12　基于 Test 2 的虚假数据注入攻击检测结果 (实例仿真)

Fig. 12　Detection results based on Test 2 under false data injection attack (real example simulation)

改为 $z(k) = y(k) + m(k)$. 从图 10 中可以看出, 当虚假数据注入攻击发生时, 文献 [34] 中阈值检测方法的结果几乎没有变化, 导致该方法无法有效检测出攻击. 然而, 从图 11 和图 12 可以看出, 当虚假数据注入攻击发生时, 本文所提方法的检测值 $D1$ 和 $D2$ 均超过上限值, 能够有效检测出攻击, 再次验证了本文所提方法的可行性和有效性.

4　结束语

本文提出了基于动态水印的模型不确定系统攻击检测鲁棒性方法. 首先分析了模型不确定项导致传统动态水印检测方法失效的原因, 然后提出了两个具有鲁棒性的攻击检测式以及检测式中关键时变方差阈值的确定方法, 理论证明了系统失真信号功率上界以定量刻画攻击信号造成系统性能损失程度, 进一步考虑攻击能够躲过检测的最坏情况新增加第三检测式, 以进一步限制系统失真信号功率上界, 从而提升不确定系统的安全性. 与文献 [28] 提出的针对确定系统水印检测方法相比, 本文解决了模型不确定系统中的攻击检测问题. 然而水印信号的添加会对系统的性能产生一定的影响, 如何衡量水印信号对控制性能的影响和检测效果是未来研究的一个重要研究方向.

References

1　Zhang X M, Han Q L, Ge X H, Ding D R, Ding L, Yue D, et al. Networked control systems: A survey of trends and techniques. *IEEE/CAA Journal of Automatica Sinica*, 2020, **7**(1): 1–17

2　Sun Q, Lim C C, Shi P, Liu F. Design and stability of moving horizon estimator for Markov jump linear systems. *IEEE Transactions on Automatic Control*, 2019, **64**(3): 1109–1124

3　Mousavinejad E, Yang F, Han Q L, Vlacic L. A novel cyber attack detection method in networked control systems. *IEEE Transactions on Cybernetics*, 2018, **48**(11): 3254–3264

4　Dolk V S, Tesi P, De Persis C, Heemels W P M H. Event-triggered control systems under denial-of-service attacks. *IEEE Transactions on Control of Network Systems*, 2017, **4**(1): 93–105

5　Langner R. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security and Privacy*, 2011, **9**(3): 49–51

6　Cherdantseva Y, Burnap P, Blyth A, Eden P, Jones K, Soulsby H, et al. A review of cyber security risk assessment methods for SCADA systems. *Computers and Security*, 2016, **56**: 1–27

7　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)

8　Kurmanaev A, Herrera I. No end in sight to Venezuelas Blackout, experts warn [Online], available: https://www.nytimes.com/2019/03/11/world/americas/venzuela-blackout-maduro.html, July 15, 2019

9　Liu Ting, Tian Jue, Wang Jia-Zhou, Wu Hong-Yu, Sun Li-Min, Zhou Ya-Dong, et al. Integrated security threats and defense of cyber-physical systems. *Acta Automatica Sinica*, 2019, **45**(1): 5–24
(刘烃, 田决, 王稼舟, 吴宏宇, 孙利民, 周亚东, 等. 信息物理融合系统综合安全威胁与防御研究. 自动化学报, 2019, **45**(1): 5–24)

10　Dibaji S M, Pirani M, Flamholz D, Annaswamy A M, Johansson K H, Chakrabortty A. A systems and control perspective of CPS security. *Annual Reviews in Control*, 2019, **47**: 394–411

11　Murguia C, Van De Wouw N, Ruths J. Reachable sets of hidden CPS sensor attacks: Analysis and synthesis tools. *IFAC-PapersOnLine*, 2017, **50**(1): 2088–2094

12　Macwan R, Drew C, Panumpabi P, Valdes A, Vaidya N, Sauer P, et al. Collaborative defense against data injection attack in IEC61850 based smart substations. In: Proceedings of the IEEE Power and Energy Society General Meeting. Boston, MA, USA: IEEE, 2016. 1–5

13　An L W, Yang G H. Secure state estimation against sparse sensor attacks with adaptive switching mechanism. *IEEE Transactions on Automatic Control*, 2018, **63**(8): 2596–2603

14　Zhou Y, Vamvoudakis K G, Haddad W M, Jiang Z P. A secure control learning framework for cyber-physical systems under sensor and actuator attack. *IEEE Transactions on Cybernetics*, 2021, **51**(9): 4648–4660

15　Li Xue, Li Wen-Ting, Du Da-Jun, Sun Qing, Fei Min-Rui. Dynamic state estimation of smart grid based on UKF under denial of service attacks. *Acta Automatica Sinica*, 2019, **45**(1): 120–131
(李雪, 李雯婷, 杜大军, 孙庆, 费敏锐. 拒绝服务攻击下基于 UKF 的智能电网动态状态估计研究. 自动化学报, 2019, **45**(1): 120–131)

16　Ma L F, Wang Z D, Han Q L, Lam H K. Variance-constrained distributed filtering for time-varying systems with multiplicative noises and deception attacks over sensor networks. *IEEE Sensors Journal*, 2017, **17**(7): 2279–2288

17　Su L, Ye D. A cooperative detection and compensation mechanism against Denial-of-Service attack for cyber-physical systems. *Information Sciences*, 2018, **444**: 122–134

18　Zhao C H, Mallada E, Dirfler F. Distributed frequency control for stability and economic dispatch in power networks. In: Proceedings of the American Control Conference. Chicago, IL, USA: IEEE, 2015. 2359–2364

19　Shitharth S, Prince W D. An enhanced optimization based algorithm for intrusion detection in SCADA network. *Computers and Security*, 2017, **70**: 16–26

20　Guo Z Y, Shi D W, Johansson K H, Shi L. Optimal linear cyberattack on remote state estimation. *IEEE Transactions on Control of Network Systems*, 2017, **4**(1): 4–13

21　Mo Y L, Sinopoli B. On the performance degradation of cyber-physical systems under stealthy integrity attacks. *IEEE Transactions on Automatic Control*, 2016, **61**(9): 2618–2624

22 Chen Y, Kar S, Moura J M F. Cyber-physical attacks with control objectives. *IEEE Transactions on Automatic Control*, 2018, **63**(5): 1418–1425

23 Mo Y L, Sinopoli B. Secure control against replay attacks. In: Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2009. Monticello, USA: IEEE, 2009. 911–918

24 Rubio Hernan J, De Cicco L, Garcia Alfaro J. On the use of watermark-based schemes to detect cyber-physical attacks. *Eurasip Journal on Information Security*, 2017: Article No. 8

25 Fang C G, Qi Y F, Cheng P, Zheng W X. Optimal periodic watermarking schedule for replay attack detection in cyber-physical systems. *Automatica*, 2020, **112**: Article No. 108698

26 Teixeira A M, Ferrari R M. Detection of sensor data injection attacks with multiplicative watermarking. In: Proceedings of 2018 European Control Conference. Limassol, Cyprus: IEEE, 2018. 338–343

27 Song Z, Skuric A, Ji K. A recursive watermark method for hard real-time industrial control system cyber-resilience enhancement. *IEEE Transactions on Automation Science and Engineering*, 2020, **17**(2): 1030–1043

28 Satchidanandan B, Kumar P R. Dynamic watermarking: Active defense of networked cyber-physical systems. *Proceedings of the IEEE*, 2017, **105**(2): 219–240

29 Satchidanandan B, Kumar P R. On the design of securityguaranteeing dynamic watermarks. *IEEE Control Systems Letters*, 2020, **4**(2): 307–312

30 Huang T, Satchidanandan B, Kumar P R, Xie L. An online detection framework for cyber-attacks on automatic generation control. *IEEE Transactions on Power Systems*, 2018, **33**(6): 6816–6827

31 Zhao D, Ding S X, Karimi H R, Li Y Y. Robust H_∞ filtering for two-dimensional uncertain linear discrete time-varying systems: A Krein space-based method. *IEEE Transactions on Automatic Control*, 2019, **64**(12): 5124–5131

32 Lai T L, Wei C Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 1982, **10**(1): 154–166

33 Hu X M, Zou Q, Zou H B. Design and application of fractional order predictive functional control for industrial heating furnace. *IEEE Access*, 2018, **6**: 66565–66575

34 Naghnaeian M, Hirzallah N H, Voulgaris P G. Security via multirate control in cyber-physical systems. *Systems and Control Letters*, 2019, **124**: 12–18

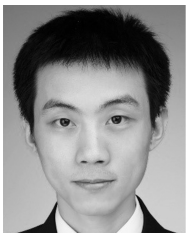
杜大军 上海大学机电工程与自动化学院教授. 主要研究方向为机器视觉和网络化系统安全控制. 本文通信作者. E-mail: ddj@shu.edu.cn

(**DU Da-Jun** Professor at the School of Mechatronics Engineering and Automation, Shanghai University. His research interest covers machine vision and security control for networked systems. Corresponding author of this paper.)

张竞帆 上海大学机电工程与自动化学院硕士研究生. 主要研究方向为网络化系统安全控制.

E-mail: shuzoooe@shu.edu.cn

(**ZHANG Jing-Fan** Master student at the School of Mechatronics Engineering and Automation, Shanghai University. His main research interest is security control for networked systems.)



张长达 上海大学机电工程与自动化学院博士研究生. 主要研究方向为网络化系统安全控制.

E-mail: zhangweiran@shu.edu.cn

(**ZHANG Chang-Da** Ph.D. candidate at the School of Mechatronics Engineering and Automation, Shanghai University. His main research interest is security control for networked systems.)

费敏锐 上海大学机电工程与自动化学院教授. 主要研究方向为网络化控制系统及实现.

E-mail: mrfei@staff.shu.edu.cn

(**FEI Min-Rui** Professor at the School of Mechatronics Engineering and Automation, Shanghai University. His research interest covers networked control system and its implementation.)

YANG Tai-Cheng 英国萨赛克斯大学工程系 Reader. 主要研究方向为物联网和网络化控制系统恶意攻击的检测及防范.

E-mail: t.c.yang@sussex.ac.uk

(**YANG Tai-Cheng** Reader in the Department of Engineering, University of Sussex, UK. His research interest covers detection and prevention of malicious cyber-attacks for networked control systems and internet of things.)