様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ３０ 日現在

機関番号：１２６０８
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２１２７２

研究課題名（和文）連続状態空間モデルに基づくCPSセキュリティリスク評価基盤の構築

研究課題名（英文）Building Foundation for CPS Security Risk Evaluation based on Continuous State-Space Model

研究代表者

笹原　帆平（Sasahara, Hampei）

東京工業大学・工学院・助教

研究者番号：５０９５４６０８

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究ではCPS（Cyber-Physical-System)のセキュリティリスク評価基盤の構築を実施した。CPSの特徴としてシステムの挙動が物理法則に従うという点に焦点を当て、状態空間モデルに基づく効率的なリスク評価アルゴリズムを開発した。また、実応用に向けてマイクログリッドにおける評価プラットフォームの開発を実施した。さらに、リザバーコンピューティングを用いた攻撃検知手法を開発した。分散型電源の不確かな発電量のもとでの実時間高速再学習を提案し、従来手法と比較してほぼ同等の性能を達成しつつ、学習速度を100倍程度高速化できることが確かめられた。

研究成果の学術的意義や社会的意義
本研究ではCPS（Cyber-Physical-System)のセキュリティリスク評価基盤の構築に成功した。CPSでは既存の機密性，完全性，可用性に加えて実空間における物理的安全性による評価が求められるため、物理現象を陽に考慮する新たな評価手法が必要であるが、体系的かつ汎用的な方法論はこれまで提案されていなかった。本研究は各種防御技術の有効性を検証するための基礎として貢献でき、さらに，リスク評価とセキュリティ向上を繰り返すCPSセキュリティマネジメント（PDCAサイクル）の促進に繋がることが期待される。

研究成果の概要（英文）：In this project, we conducted the construction of a security risk assessment framework for CPS (Cyber-Physical Systems). Focusing on the characteristic of CPS where system behavior adheres to physical laws, we developed an efficient risk assessment algorithm based on state-space models. Additionally, we implemented the development of an evaluation platform for microgrids for practical applications. Furthermore, we developed an attack detection method using reservoir computing. We proposed real-time fast retraining under uncertain generation of distributed power sources, confirming that it achieves nearly equivalent performance to conventional methods while accelerating learning speed by approximately 100 times.

研究分野：制御理論

キーワード：CPSセキュリティ　リスク評価　制御理論

１．研究開始当初の背景

情報通信及びデータ処理技術により物理システムの高度な制御を実現する CPS の概念に基づき，エネルギー等の基幹産業のスマート化が加速している．しかし，物理層におけるリアルタイム性の要求等のために，暗号通信等の高機能防御技術の実装は必ずしも容易でない．このことから CPS がバリューチェーンにおける“ weakest link ”となり得る点が認識され，そのセキュリティリスク評価の重要性が高まっている．しかし，CPS では従来の情報セキュリティ三要素（機密性，完全性，可用性）に加えて，実空間における物理的安全性による評価が必要である．よって情報システムを想定した既存の枠組みは適用できず，物理現象を陽に考慮する CPS セキュリティリスク評価の原理と方法を明らかにすることが求められている．

２．研究の目的

（１）第一の目的として，CPS セキュリティにおいて標準的な手法であるモデルベース防御を対象として，理論的な性能評価を与える． 特に，Bayes 推論に基づく防御手法を考え，巧妙な攻撃に対する堅牢性を解析する．さらに実応用に向けて，CPS セキュリティリスク評価のための汎用的なリスク評価フレームワークおよびスケーラブルな計算アルゴリズムを構築する．

（２）第二の目的として，リスク評価に加えて，ニューラルネットワーク及び CPS の動的な観測データに基づく攻撃検知手法を構築することで，リスク低減の枠組みを提案する．

（３）第三の目的として，新しい脅威シナリオとして観測データの改ざんを考え，その影響について解析することで，これまで取り扱っていなかったリスクの評価を可能とする手法を構築する．

３．研究の方法

（１）第一の目的に対して，CPS の挙動と想定脅威シナリオを組み込み，Markov 決定過程を用いて動的確率モデルを構築する．適切な数理モデルのもとで，想定脅威シナリオの影響は検出確率が一定値以下の攻撃が及ぼし得る影響の最悪値として定量化される．まず理論解析のために不完備情報ゲームに基づく均衡解析を行う．得られた均衡から導き出される CPS の挙動を調べることで，動的な堅牢性を解析する．次に，リスク評価の定式化を行う．一般に，各時刻における最悪攻撃は現在の状態だけでなく過去の全履歴に依存する．そのため探索空間の次元が区間長に応じて指数的に増大し，この点が本枠組みにおける最大の技術的課題となっている．この問題に対し，警報履歴導入による状態空間拡大を導入することで，時間スケールに対する探索空間の増大を抑制することを可能とする．このもとでスケーラブルな計算アルゴリズムを構築する．

（２）第二の目的に対して，リザバーコンピューティングを用いる攻撃検出を行う．リザバーコンピューティングは再帰的ニューラルネットワークの一種であるが，中間層をランダムに生成した上で出力層だけをデータから学習する特徴を持ち，多くのタスクにおいて高速かつ高精度な学習を達成することができることが知られている．特に電力システムに注目し、電力システムセキュリティにおける標準的なデータセットを利用し、提案法の性能を評価する．

（３）第三の目的に対して，機械学習分野の特に画像識別タスクにおいて整理されている敵対的学習の概念を拡張することで，微小ながらも制御性能に大きな影響を与える敵対的摂動を生成しデータを汚染するアルゴリズムの構築を行う．特に，ニューラルネットワークに対する最も標準的な敵対的摂動生成アルゴリズムである Fast Gradient Sign Method（FGSM）を拡張することを考える．FGSM の基本的なアイデアは，摂動の絶対値を一定以下に抑えた上で損失関数の勾配が正となる摂動を作成することである．CPS において，制御システムの安定度を損失関数とみなすことで，システムを不安定化させる摂動を生成するアルゴリズムを構築する．

４．研究成果

（１：発表論文 [1]）図１に示すモデルベース防御機構を考える．攻撃側（sender）は何らかの脆弱性を利用した行動を起こし，システムの振る舞いを変化させることができるとする．この攻撃に対処するため，防御側（receiver）は状態を観測し，攻撃を受けているか判断した上で適切な対応を行う．例えば，もしもシステムの振る舞いが正常時の典型的な挙動と異なっている場合，攻撃者の存在に関する信念は上昇し，ログ解析等の対応を行う．逆に，システムの振る舞いが典型的なものであった場合，攻撃者は存在しないと判断される．本研究の興味の一つは，

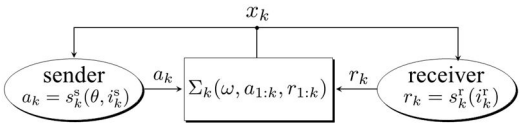

図 １：モデルベース防御機構

どんな攻撃に対してもモデルベース防御機構は攻撃を検出できるのかを調べることである．この問題に対して，不完備情報動的ゲーム理論の枠組みを用いて攻撃側及び防御側の妥当な戦略を Bayesian-Nash 均衡としてモデル化し，均衡が導くシステムの挙動を理論的に解析した．結果として，任意の均衡の下で信念が劣マルチンゲール性を持つことを証明した．このことは信念および Bayes 推論の収束性を意味する．すなわち，モデルベース防御機構はどんな攻撃に対しても漸近的に検出できることを理論的に証明することに成功した．

加えて，非漸近解析のために，モデルベース防御機構として具体的な検出器を想定した場合の有限時間リスク評価手法を構築した．この場合，リスクは結合機会制約付き最適制御問題の解として定量化される．この問題のテクニカルな困難は，結合機会制約のために最悪攻撃が過去の全履歴に依存するため，探索空間の次元が区間長に応じて指数的に増大する点である．この問題に対し，最悪攻撃特定のためには全履歴の詳細は不要であり，警報が過去に鳴動したか否かという情報が十分統計量であることを証明した．この二値情報を新たな状態として参照することを警報履歴導入による状態空間拡大と呼び，時間スケールに対する探索空間の増大を抑制することが可能であることを示した．

（２：発表論文 [2.3]）提案攻撃検知手法の検証データとして，IEEE68 バス電力モデルを用いたシミュレータを構築した．IEEE68 バス電力モデルニューイングランドテストシステム（NETS）とニューヨークパワーシステム（NYPS）および隣接する３つの地域が接続されたモデルであり，個別に負荷周波数制御（LFC）が行われている．それぞれの LFC に対して，周波数偏差データに対する偽データ注入攻撃が入ったデータセットを作成した．このデータセットに対して，リザバーコンピューティングの代表的なモデルであるエコーステートネットワーク（ESN）を用いた検出器を設計した．代表的な再帰的ニューラルネットワークモデルである LSTM との比較を表１に示す．まず性能に関して，提案法は LSTM とほぼ同等の分類性能を達成できることがわかった．さらに，ESN モデルは計算時間を大幅に短縮しており，ESN モデルの学習にかかる時間は LSTM モデルの学習の１エポックにかかる時間よりも短く，学習時間の大幅な短縮化が実現できていることがわかった．また，誤分類された場合のシナリオの分類結果を表２に示す．４つのパターンのうちで最も深刻なのは，攻撃シナリオを正常シナリオとして分類するものであるが，その割合は最も低いことがわかった．

表１： ESN と LSTM の比較

	正解率 [%]	学習時間 [秒]
ESN（提案法）	99.11	1.568
LSTM	99.28	276.7

表 2：誤分類結果

分類結果 / 真値	正常	攻撃
正常	1.84%	2.61%
攻撃	0.38%	2.87%

（３：発表論文 [4,5]）直接データ駆動型制御手法を前提として，制御システムの閉ループ系の最大固有値を最大化するようなデータ摂動の設計方法（DGSM: Directed Gradient Sign Method）を構築した．具体的には，摂動サイズである行列の Max ノルムを制約条件として，評価関数の線形近似関数を最大化する手法である．また，射影勾配法に基づき，反復的に DGSM を適用する IDGSM（Iterative DGSM）という手法に拡張した．制御理論分野における代表的なベンチマークモデルである倒立振子モデルを用いて，攻撃手法の効果を検証した．表３に攻撃による不安定化率を示す．全ての状況において IDGSM で生成された攻撃が最も不安定化率が高く，最も深刻な攻撃であることがわかった．また，これらの結果を基礎として，ロバスト安定化のための正則化手法を同時に提案し，有効性を検証した．

表 3：各攻撃による不安定化率

摂動サイズ	ランダム	DGSM	IDGSM
0.0001	3.1%	7.2%	22.1%
0.001	11.9%	26.1%	69.6%
0.01	14.5%	36.9%	86.2%
0.1	11.4%	34.6%	89.4%

発表論文

[1] H. Sasahara and H. Sandberg, Asymptotic Security using Bayesian Defense Mechanism with Application to Cyber Deception, IEEE Transactions on Automatic Control, 2024. (掲載決定)

[2] 金，笹原，井村，リザバーコンピューティングによるスマートグリッドへの攻撃検出，第 10 回計測自動制御学会制御部門マルチシンポジウム，2023.

[3] K. Kim, H. Sasahara, and J. Imura, Cyberattack Detection in Smart Grids based on Reservoir Computing, IFAC World Congress, 2023.

[4] 神永，笹原，井村，Direct Data-driven Control に対する射影勾配法を用いた敵対的攻撃，第 11 回計測自動制御学会制御部門マルチシンポジウム，2024.

[5] H. Sasahara, Adversarial Attacks to Direct Data-driven Control for Destabilization, The 62nd IEEE Conference on Decision and Control, 2023.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Sasahara Hampei、Sandberg Henrik	４．巻 -
２．論文標題 Asymptotic Security using Bayesian Defense Mechanism with Application to Cyber Deception	５．発行年 2023年
３．雑誌名 IEEE Transactions on Automatic Control	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TAC.2023.3340978	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 金起誠　笹原帆平　井村順一
２．発表標題 リザバーコンピューティングによるスマートグリッドへの攻撃検出
３．学会等名 第10回計測自動制御学会制御部門マルチシンポジウム
４．発表年 2023年

１．発表者名 Sasahara Hampei
２．発表標題 Adversarial Attacks to Direct Data-driven Control for Destabilization
３．学会等名 The 62nd IEEE Conference on Decision and Control（国際学会）
４．発表年 2023年

１．発表者名 Kim Kisong, Sasahara Hampei, Imura Jun-ichi
２．発表標題 Cyberattack Detection in Smart Grids based on Reservoir Computing
３．学会等名 IFAC World Congress 2023（国際学会）
４．発表年 2023年

１．発表者名 神永泰良、笹原帆平、井村順一
２．発表標題 Direct data-driven controlに対する射影勾配法を用いた敵対的攻撃
３．学会等名 第11回計測自動制御学会制御部門マルチシンポジウム
４．発表年 2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関