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　５ 年　 ６ 月　 ８ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１８００

研究課題名（和文）機械学習を用いた悪性ドメイン名検知システムのホワイトボックス化に関する研究

研究課題名（英文）A Study on White-Boxing of Malicious Domain Name Detection System Using Machine Learning

研究代表者
内田　真人（Uchida, Masato）

早稲田大学・理工学術院・教授

研究者番号：２０４１９６１７

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、機械学習を用いた悪性ドメイン名検知システムのホワイトボックス化を目指し、それに必要となる要素について検討した。まず、ドメイン名を含むインターネット資源に関わる悪性活動の実態について調査した。次に、検知された悪性活動の解釈性を向上させるための可視化手法を提案した。その上で、本研究の主要な目標である識別モデル、説明モデル、及び人間の専門知識を統合させる手法について検討した。また、研究目的を達成する上で欠かすことのできない識別モデルや説明モデルの信頼性についても検討した。

研究成果の学術的意義や社会的意義
本研究では、機械学習を用いた悪性ドメイン名検知システムのホワイトボックス化を実現する上で必要となる要素について検討した。これにより、検知結果の解釈性や信頼性が向上し、セキュリティアナリストがより正確かつ効果的なリスク分析やインシデント対応を行うための支援が可能になる。また、透明性の高いセキュリティ対策の実現や信頼性の高いシステムの構築にも貢献することが期待される。これらの成果は、悪性ドメイン名検知システムの透明性の向上や信頼できるセキュリティ対策の実現に寄与し、学術的・社会的な意義を持つものといえる。

研究成果の概要（英文）：In this study, we aimed to achieve the white-boxing of a malicious domain name detection system using machine learning and examined the necessary elements. First, we investigated the nature of malicious activities related to internet resources involving domain names. Next, we proposed visualization techniques to enhance the interpretability of detected malicious activities. Furthermore, we discussed the integration of identification models, explanation models, and human expertise, which are the main objectives of this research. We also investigated the reliability of identification models and explanation models, which are indispensable for achieving the research goals.

研究分野：機械学習の理論と応用

キーワード：悪性ドメイン名検知　判断根拠説明

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

悪意のあるドメイン名（悪性ドメイン名）を攻撃のインフラとして利用するサイバー攻撃が頻発している。Cisco によると、2015 年に観測されたサイバー攻撃のうち、悪性ドメイン名を利用したものは 91.3%を占めていた[A]。このような悪性ドメイン名を早期に検知することは、マルウェア感染やフィッシングサイトへのアクセス、組織外への情報漏洩などを防ぐ上で重要な役割を果たす。このため、これまでに機械学習を用いて悪性ドメイン名を検知するシステムが検討されてきた。例えば、文献[B]では、ドメイン名の利用状況に関する経時変化を特徴量とすることで、99%という非常に高い検知精度を実現する手法が提案されている。また、文献[C]では、文献[B]を上回る検知精度を、わずかな訓練データを使用して実現する手法を提案している。このように、これまでの機械学習を用いた悪性ドメイン名検知に関する先行研究では、検知精度の向上、すなわち「検知結果の正確性」の向上が中心課題とされてきた。

一方で、検知されたセキュリティ脅威に対する的確なリスク分析やインシデント対応が求められるセキュリティアナリストを支援するためには、「検知結果の正確性」だけでなく、「検知結果の解釈性」の向上も必要である。例えば、悪性であると疑われるドメイン名が検知された場合、そのドメイン名を差し押さえたりサイトを閉鎖したりするなどの措置を講じることがあるが、このような過失責任が問われかねない措置に踏み切るためには、検知結果に至る判断根拠を正確に解釈する必要がある。しかしながら、悪性ドメイン名を検知するために使用される機械学習モデルの内部構造や内部動作は検知精度の向上とともに複雑化しており、検知結果に至る判断根拠はブラックボックス化され、セキュリティアナリストが直接解釈することは困難である。以上より、機械学習を用いた悪性ドメイン名検知に関する研究フェーズは、「検知結果の正確性」の向上に加え、「検知結果の解釈性」の向上も追求する段階に移行すべき状況にあるといえる。

[A] http://mkto.cisco.com/rs/564-whv-323/images/cisco-asr-2016.pdf
[B] D. Chiba, T. Yagi, M. Akiyama, T. Shibahara, T. Yada, T. Mori, and S. Goto: DomainProfiler: toward accurate and early discovery of domain names abused in future, Int. J. Inf. Secur., Nov. 2018.
[C] N. Fukushi, D. Chiba, M. Akiyama, and M. Uchida, “Exploration into Gray Area: Toward Efficient Labeling for Detecting Malicious Domain Names”, IEICE Trans. Commn., Vol.E103-B, No.4, pp.375-388, 2020.

２．研究の目的

本研究ではこうした学術的背景を踏まえ、以下の学術的「問い」を掲げた。

問い：ブラックボックス化された悪性ドメイン名検知のホワイトボックス化は可能か？
- 検知結果に影響を及ぼす特徴量間の相互作用を考慮した判断根拠説明はできないか？
- 多様な判断根拠説明を組み合わせることで説明に対する信頼性を向上できないか？
- 判断根拠説明の内容に対する人間による評価結果に基づき、悪性ドメイン名検知システムを再構築できないか？

本研究の目的は、機械学習を用いた悪性ドメイン名検知システムのホワイトボックス化、すなわち「検知結果の解釈性」の向上である。この目的を達成するために、本研究では、悪性ドメイン名検知システムの構成要素である、識別モデル(入力されたドメイン名を識別し、検知結果を出力するモデル)、説明モデル(検知結果に至る判断根拠を説明するためのモデル)、人間(識別モデルや説明モデルを利用するセキュリティアナリスト)の三者を連携させるというコンセプトに基づき検討を行った。

３．研究の方法

上記の研究目的を達成するために、以下に示すように、課題を細分化して研究を進めた。

まず、ドメイン名を含むインターネット資源に関わる悪性活動の実態について調査した。具体的には、ドメインパーキングを利用するドメイン名に注目し、ドメインパーキングを利用する期間と悪性活動を行う期間の時系列関係を調査した。また、クラウドサービスを悪用するサイバー攻撃の実態や、いかなるエンドユーザにも割り当てられていないにも関わらず経路広告された IP アドレスの実態について調査した。

次に、検知された悪性活動の解釈性を高めるための可視化手法を提案した。具体的には、新たに検知された Android マルウェアが、既存の Android マルウェアやその種別（ファミリ）とどのような関係にあるか、またそれらが時間経過とともにどう変化してきたかを表現することが可能な「Android マルウェアの家系図」を機械学習を用いて自動作成する手法について検討した。

その上で、本研究の主目的である識別モデル、説明モデル、及び人間を連携させる手法について検討した。具体的には、識別モデルによる検知結果と、それに対する説明モデルによる説明結果を照らし合わせることで識別結果の信頼性を評価し、信頼性に疑義がある場合には、その識別結果に対する異議を唱える機械学習モデル（異議判定モデル）について検討した。異議判定モデルの判定結果を踏まえることで、優先的な検証が必要となる識別結果（異議あり）と、そうでないもの（異議なし）をより分けられるようになる。これにより、誤検知や見逃しを効率的に特定

できるようになり、大量のセキュリティインシデントに対する判断を求められる現場のアナリストの労力が削減され、人間を含めたシステム全体としての識別精度を向上できる。

また、研究目的を達成する上での前提となる説明モデルの信頼性について検討した。機械学習を用いた予測モデルの解釈性を確保するための代表的なアルゴリズムでは、入力データに対する摂動に基づいて予測モデルの出力を説明する。そこで本研究では、この摂動を無効化する手法について検討し、既存の代表的な解釈性アルゴリズムにおける脆弱性を明らかにした。

さらに、研究目的を達成する上での前提となる識別モデルの信頼性についても検討した。具体的には、入力に対する微弱な摂動により機械学習モデルの出力を誤らせるという敵対的攻撃について、その動特性の解析、性能評価の検証、防御手法の提案などに関する基礎検討を開始した。

４．研究成果

機械学習を用いた悪性ドメイン名検知システムのホワイトボックス化（検知結果の解釈性の向上）に向けて以下の研究を行った。

４．１．悪性活動の実態調査

４．１．１．ドメインパーキングサービスを利用した悪性ドメイン名の実態調査

インターネット上には、実際に利用されていない未使用のドメイン名が多数存在する。このような未使用ドメインにオンライン広告を表示させてマネタイズするのが、ドメインパーキングである。サイバー攻撃で使用されたドメイン名の中には、攻撃後にドメインパーキングサービスを利用するものがあることが知られている。しかしながら、ドメインパーキングサービスと悪意のあるドメイン名の時間的関係については、これまでほとんど研究されてこなかった。本研究では、ドメインパーキングサービスを利用して悪質なドメイン名が時間的にどのように変化するかを調査した。過去 19 ヶ月間に、6,680 万以上のドメイン名を調査対象とし、大規模な計測調査を行った。その結果、ドメインパーキングを利用した後に悪質化したドメイン名が 3,964 件存在することが明らかになった。さらに、そのような悪意のあるドメイン名がどのような種類の悪意ある活動（フィッシングやマルウェアなど）に利用される傾向があるのかを明らかにした。また、複数のパーキングサービスを同時に、または切り替えながら利用した 302 万件のドメイン名の存在も明らかになった。本研究は、ドメインパーキングサービスを利用した悪質なドメイン名の効率的な分析に貢献することができる。

４．１．２．未割り当て IP アドレスの経路広告の実態調査

インターネット上の経路制御において、どのエンドユーザにも割り当てられていない IP アドレスが経路広告されるという問題が知られている。この問題は、現在の BGP によるインターネットの経路制御の仕組みでは、未割り当て IP アドレスですら経路広告が可能であることに起因している。しかし、技術的な原因は明らかである一方、その実態は明らかでなく、また調査手法も確立されていない。そこで本研究では、日本国内において未割り当て IP アドレスの経路広告実態調査を行い、その簡便かつ有効な手法を提案した。提案手法では、日本が割り振りを受けたアドレスプールから未割り当て IP アドレスを抽出し、インターネット上で公開されている経路情報と比較する。検出された経路広告の詳細を追加調査した結果、国内外の AS から数年にわたり未割り当て IP アドレスが経路広告されていたことが明らかになった。また、その全てがネットワーク管理者の設定ミスによるものであったこともわかった。以上の問題は、レジストリや ISP、通信事業者だけでなく、一般のエンドユーザにも影響を及ぼすものであり、本研究が経路広告設定の正当性を保つための注意喚起となることが期待される。

４．２．悪性活動の可視化：Android マルウェアの家系図作成

Android を標的としたマルウェアの脅威は増加し続けている。Android マルウェアは時間経過とともに種別（ファミリ）が増加するため、その対策のためにはある時点においてマルウェアを検知するだけでなく、マルウェアの時系列変化を考慮した解析を行う必要がある。そこで、本研究では、新たに検知された Android マルウェアが、既存の Android マルウェアやそのファミリとどのように関連しているか、そしてそれらが時間とともにどのように変化しているかを表現できる「Android マルウェアの家系図」を自動作成する手法を提案した。実際の Android マルウェア 18,958 個を利用した評価の結果、本研究で提案するマルウェアの時系列変化に対応した家系図を作成することで、マルウェアのファミリ間の時間的変化を正確に表現できることがわかった。また、マルウェアのトレンドの把握やファミリ推定など、多くの示唆が得られることがわかった。提案手法を用いることで、日々収集されるマルウェアをより効率的に解析、分類し、様々な情報と組み合わせて脅威情報として利用できることが期待される。

４．３．検知結果の信頼性評価：XAI による誤識別された悪性活動の特定

サイバー空間における様々な悪性活動を機械学習で構築した識別モデルで検知する手法が多数検討されている。しかし、どのような識別モデルであっても誤検知や見逃しはつきものであり、人間による検証が欠かせない。これを補助する手法に、識別結果の判断根拠を提示する説明可能 AI（eXplainable AI: XAI）がある。しかし、検証の対象となる識別結果の件数が膨大である場合、全件について XAI の出力を確認するのは現実的ではない。また、XAI の出力を解釈すること

自体が難しい場合もある。そこで本研究では、XAI の出力を特徴量として用いることで識別結果を検証し、信頼性に疑義がある場合には異議を唱える機械学習モデル（異議判定モデル）を提案した。悪性サイト検知とマルウェア検知に関する実験の結果、異議判定モデルを用いることで、誤識別（誤検知、見逃し）された悪性活動を効率的に特定できることがわかった。これにより、異議判定モデルを利用することで、誤検知や見逃しを特定するために必要となるセキュリティアナリストの労力を軽減できることを確認した。

４．４．説明モデルの信頼性に関する検討：XAI における脆弱性の発見

悪性活動を検知するために使用される識別モデルにおいて、その出力に対する判断根拠の解釈可能性を確保することは、ブラックボックスな識別モデルに対するユーザーの信頼を得る上で重要である。解釈可能性を確保するための代表的なアルゴリズムである LIME と SHAP は、識別モデルへの入力に対して摂動を与えたときの出力の変化から説明内容を生成するというメカニズムに基づいている。本研究では、このような摂動ベースの解釈可能性アルゴリズムによる説明を操作するための原理を定式化し、予測モデルが重要視する特徴を隠す手法について検討した。具体的には、予測の理由を隠すためのマスカレード層を提案し、それを利用することで、与えられた摂動を無効化できることを示した。この層はどの識別モデルに対しても、識別モデル自体を変更することなく接続することができる。これにより、予測モデルの動作を変更せずに、解釈可能性アルゴリズムが提供する説明を操作することができる。実際のデータを用いた実験では、提案された方法により、効果的かつ柔軟に解釈可能性アルゴリズムの説明を操作できることを示した。本研究の結果から、既存の摂動ベースの解釈可能性アルゴリズムには、信頼性に関する重大な弱点があることが示唆される。

４．５．識別モデルの信頼性に関する検討：機械学習への攻撃とその防御に関する基礎検討

機械学習モデルへの脆弱性攻撃の一つとして、入力データに微小なノイズを加えることで、人間に気が付かれることなく、機械学習モデルの誤分類を作為的に誘発させる Adversarial Example（AE）がある。機械学習モデルを識別モデルとして使用する場合、この脆弱性は脅威となる。そこで本研究では、機械学習への攻撃とその防御に関する基礎検討として、AE による攻撃性能の「人間と機械学習モデルの両者をだます」という本来の攻撃成功の定義に基づいた評価、サンプル単位での攻撃の成否というミクロな挙動の分析に基づいた攻撃手法の評価、ノイズのバランスを崩すという新たな着想に基づいた防御手法の提案に関する基礎検討を開始した。

参考文献

[1] Takayuki Tomatsuri, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida: Time-series Measurement of Parked Domain Names and Their Malicious Uses, IEICE Transactions on Communications, Vol.E104-B, No.7, pp.770-780, 2021.

[2] 五島 健太朗，澁谷 晃，岡田 雅之，内田 真人：未割り当て IP アドレスの経路広告の実態調査，コンピュータセキュリティシンポジウム 2020，pp.159-166，2020 年 10 月.

[3] Kentaro Goto, Akira Shibuya, Masayuki Okada and Masato Uchdia: Analysis of Route Announcements of Unassigned IP Addresses, The 9th IEEE International Workshop on Architecture, Design, Deployment & Management of Networks & Applications (ADMNET 2021), July 2021.

[4] 野村 和也，千葉 大紀，秋山 満昭，内田 真人：Android マルウェアの家系図作成，コンピュータセキュリティシンポジウム 2020，pp.527-534，2020 年 10 月 26 日.

[5] Kazuya Nomura, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida: Auto-creation of Android Malware Family Tree, IEEE International Conference on Communications (ICC 2021), Online, June 2021.

[6] 藤田 晃治，芝原 俊樹，千葉 大紀，秋山 満昭，内田 真人，異議あり！：XAI による誤識別された悪性活動の特定，コンピュータセキュリティシンポジウム 2021，pp.898-905，2021 年 10 月.

[7] Koji Fujita, Toshiki Shibahara, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida: Objection!: Identifying Misclassified Malicious Activities with XAI, IEEE International Conference on Communications (ICC 2022), May 2022.

[8] 藤森 洸，芝原 俊樹，千葉 大紀，秋山 満昭，内田 真人：本来の定義に則った Adversarial Example の評価,セキュリティ心理学とトラスト研究会，Vol.2023-SPT-50，No.34，pp.1-8，2023 年 3 月.

[9] 五島 健太朗，内田 真人：敵対的攻撃手法の動的解析,情報通信システムセキュリティ研究会，信学技報 Vol.122，No.422，pp.25-30(ICSS2022-52)，2023 年 3 月

[10] 添田 遼，内田 真人：CMYK 防御モデルを用いた色変換による Adversarial Example の無効化，電子情報通信学会 情報通信システムセキュリティ研究会，信学技報 Vol.122，No.422，pp.37-42(ICSS2022-54)，2023 年 3 月.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Takayuki Tomatsuri, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida	４．巻 E104-B
２．論文標題 Time-series Measurement of Parked Domain Names and Their Malicious Uses	５．発行年 2021年
３．雑誌名 IEICE Transactions on Communications	６．最初と最後の頁 770-780
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transcom.2020CQP0007	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Katsura Yasuhiro and Uchida Masato	４．巻 2
２．論文標題 Candidate-Label Learning: A Generalization of Ordinary-Label Learning and Complementary-Label Learning	５．発行年 2021年
３．雑誌名 SN Computer Science	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） 10.1007/s42979-021-00681-x	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nomura Kazuya, Chiba Daiki, Akiyama Mitsuaki and Uchida Masato	４．巻 29
２．論文標題 Auto-creation of Robust Android Malware Family Trees	５．発行年 2021年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 801～811
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.29.801	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Naoki Fukushi, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida	４．巻 29
２．論文標題 A Comprehensive Measurement of Cloud Service Abuse	５．発行年 2021年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 93～102
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.29.93	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計14件（うち招待講演　1件／うち国際学会　7件）

1．発表者名
藤森 洸，芝原 俊樹，千葉 大紀，秋山 満昭，内田 真人

2．発表標題
本来の定義に則ったAdversarial Exampleの評価

3．学会等名
セキュリティ心理学とトラスト研究会

4．発表年
2023年

1．発表者名
添田 遼，内田 真人

2．発表標題
CMYK防御モデルを用いた色変換によるAdversarial Exampleの無効化

3．学会等名
情報通信システムセキュリティ研究会

4．発表年
2023年

1．発表者名
五島 健太朗，内田 真人

2．発表標題
敵対的攻撃手法の動的解析

3．学会等名
情報通信システムセキュリティ研究会

4．発表年
2023年

1．発表者名
Koji Fujita, Toshiki Shibahara, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida

2．発表標題
Objection!: Identifying Misclassified Malicious Activities with XAI

3．学会等名
IEEE International Conference on Communications (ICC 2022)（国際学会）

4．発表年
2022年

１．発表者名
Kazuya Nomura, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida

２．発表標題
Auto-creation of Android Malware Family Tree

３．学会等名
IEEE International Conference on Communications (ICC 2021)（国際学会）

４．発表年
2021年

１．発表者名
Kentaro Goto, Akira Shibuya, Masayuki Okada and Masato Uchdia

２．発表標題
Analysis of Route Announcements of Unassigned IP Addresses

３．学会等名
The 9th IEEE International Workshop on Architecture, Design, Deployment & Management of Networks & Applications (ADMNET 2021)（国際学会）

４．発表年
2021年

１．発表者名
Tomohiro Koide and Masato Uchida

２．発表標題
Behind The Mask: Masquerading The Reason for Prediction

３．学会等名
The 45th IEEE Annual Computer Software and Applications Conference (COMPSAC 2021)（国際学会）

４．発表年
2021年

１．発表者名
藤田 晃治，芝原 俊樹，千葉 大紀，秋山 満昭，内田 真人

２．発表標題
異議あり！：XAIによる誤識別された悪性活動の特定

３．学会等名
コンピュータセキュリティシンポジウム2021

４．発表年
2021年

１．発表者名
内田 真人

２．発表標題
人間参加型機械学習によるサイバーセキュリティ

３．学会等名
情報通信マネジメント研究会（招待講演）

４．発表年
2021年

１．発表者名
Yasuhiro Katsura and Masato Uchida

２．発表標題
Bridging Ordinary-Label Learning and Complementary-Label Learning

３．学会等名
The 12th Asian Conference on Machine Learning (ACML 2020)（国際学会）

４．発表年
2020年

１．発表者名
Takayuki Tomatsuri, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida

２．発表標題
Time-series Measurement of Parked Domain Names

３．学会等名
IEEE Global Telecommunications Conference (GLOBECOM 2020)（国際学会）

４．発表年
2020年

１．発表者名
Kentaro Goto and Masato Uchida

２．発表標題
Tsallis Entropy Based Labelling

３．学会等名
IEEE International Conference on Machine Learning and Applications (IEEE ICMLA 2020)（国際学会）

４．発表年
2020年

１．発表者名
野村 和也，千葉 大紀，秋山 満昭，内田 真人

２．発表標題
Androidマルウェアの家系図作成

３．学会等名
コンピュータセキュリティシンポジウム2020

４．発表年
2020年

１．発表者名
五島 健太朗，澁谷 晃，岡田 雅之，内田 真人

２．発表標題
未割り当てIPアドレスの経路広告の実態調査

３．学会等名
コンピュータセキュリティシンポジウム2020

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

内田研究室
https://uchida-lab.jp/
内田研究室
https://uchida-lab.jp/

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