

異議あり！：XAIによる誤識別された悪性活動の特定

藤田 晃治^{1,a)} 芝原 俊樹² 千葉 大紀² 秋山 満昭² 内田 真人¹

概要：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。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) を提案する。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、誤識別された悪性活動を効率的に特定できることがわかった。

キーワード：XAI, 機械学習, 悪性サイト検知, マルウェア検知

Objection!: Identifying Misidentified Malicious Activities with XAI

KOJI FUJITA^{1,a)} TOSHIKI SHIBAHARA² DAIKI CHIBA² MITSUAKI AKIYAMA² MASATO UCHIDA¹

Abstract: There have been many studies to detect various malicious activities in cyberspace using discriminative models built by machine learning. However, any discriminative model is inevitable to make mistakes, so human verification is necessary. One method to address this issue is XAI (eXplainable AI), which provides a reason for identification. However, when the number of identification results to be verified is huge, it is not realistic to check the output of XAI for all cases. In addition, it is sometimes difficult to interpret the output of XAI. In this study, we propose a machine learning model called objection judgment model that verifies the identification results by using the output of XAI as a feature, and raises objections when there is doubt about the reliability. The results of experiments on malicious website detection and malware detection show that the objection judgment model can efficiently identify misidentified malicious activities.

Keywords: XAI, Machine learning, Malicious website detection, Malware detection

1. はじめに

企業や組織ではサイバーセキュリティを担保するために、セキュリティ管理や脅威検知のためのシステムを多数導入している。セキュリティオペレーションセンタ (SOC) は、このようなシステムの運用を行う組織であり、SOC のアナリストは、システムから出力される大量のログやアラートを監視、分析し、必要な対処を行う。文献 [1-3] によると、日々大量に発生するアラートを処理するアナリストが、ア

ラート疲労と呼ばれる状況をおこし、アナリストの燃え尽きにつながることが長年問題視されてきた。

一方、アラートの精度を高める手段の一つとして、サイバー空間における各種の悪性活動を機械学習を用いて検知する手法が多数検討されている [4, 5]。機械学習を用いて構築した識別モデルによる悪性／良性活動の識別精度は向上しており、その有効性は広く認められている。しかし、どのような機械学習の手法によって構築した識別モデルであっても必ず誤検知や見逃しが生じ、100%の識別精度を達成することは現実的に不可能である。そのため、識別結果を無条件に信頼するのではなく、誤検知や見逃しが本当に生じていないかを人間 (アナリスト) の目によって検証す

¹ 早稲田大学 / Waseda University

² 日本電信電話株式会社 / NTT

^{a)} fujitakoji@fuji.waseda.jp

る必要がある。ところが、識別モデルの精度が向上することによって生ずる以下のような逆説的な問題がある。

- (a) 高精度な識別モデルですら区別するのが難しい巧妙な悪性活動と真の良性活動は、人間の目で区別することも同様に難しい。
- (b) 高精度な識別モデルにおいては誤検知や見逃しがわずかにしか生じない。そのため、識別モデルから出力される大量な識別結果の中にわずかに含まれる誤った識別結果を特定する労力が膨大となる。例えば、無作為に選択した識別結果を人間の目で検証したとしても、正しく識別されていたということが改めて確認されることがほとんどで、誤った識別結果を特定するためには何度も「空振り」を繰り返すことになる。

機械学習を用いて構築した識別モデルによる誤検知や見逃しを特定する上で、その識別モデルがどのような判断根拠に基づいて悪性／良性活動を識別したのかを解釈することは重要である。しかし、識別モデルの内部構造や内部動作は識別精度の向上に伴い複雑化しており、そこから出力される識別結果に至る判断根拠はブラックボックス化され、アナリストが直接解釈することは困難である。この問題を解決するために、近年では、識別結果に加えて、その判断根拠や付随する情報を提示する説明可能 AI (eXplainable AI: XAI) と呼ばれる手法が盛んに研究されている。XAI の代表的な手法である LIME [6] では、特徴量に関する線形モデルを用いた説明モデルによって説明結果を生成する。また、LIME の拡張手法である SHAP [7] では、ゲーム理論におけるシャプレイ値を用いた説明モデルによって説明結果を生成する。このような XAI の手法により判断根拠説明を得ることができれば、誤検知や見逃しの特定に役立てることができる。しかし、XAI による判断根拠説明を解釈するのは人間であり、XAI の原理に精通する専門家でも難しい場合がある [8]。そのため、XAI を単純に用いるのみでは上記の問題 (a) を解決できない。また、大量の識別結果に対する判断根拠説明の全件を人間が手動で解析するのは現実的ではない。そのため、上記の問題 (b) を解決するためには、識別モデルによる識別結果の中から優先的に検証すべきものを選択する必要がある。

そこで本研究では、識別モデルによる識別結果と、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 (以下、異議判定モデル) を提案する。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) をより分けることができる。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

向上することができる。異議判定モデルの学習アルゴリズムの概要は以下のとおりである。

- (1) 識別モデルによる識別結果に対して適当な XAI の手法を用いて判断根拠説明を行い、識別結果に対する説明結果を得る。
- (2) この説明結果を特徴量 (以下、XAI 特徴量) として用いて識別結果の正誤を予測し、識別結果に対する異議の有無 (異議あり／異議なし) を判定するための異議判定モデルを学習する。

識別結果に至る判断根拠が XAI の手法によつて的確に説明されていると仮定すれば、区別が難しい悪性／良性活動の識別結果に対する説明結果の差異は微細であると考えられる。そのような微細な差異は識別結果の正誤の判定において重要な情報になり得る一方で、人間にとっては直感的ではなく解釈しにくい情報でもあることが想定される。本研究では、誤識別の際の説明結果に表れるこのようなわずかな特徴を機械学習でモデル化する。これにより、人間にとっては気が付きにくい情報であっても、悪性活動の誤検知や見逃しの効率的な特定に有効活用する。識別結果と説明結果を照らし合わせ、識別結果の判断根拠を解釈し、識別結果の正誤を判断する、という一連の作業を人間が行うのではなく異議判定モデルに代行させることで、アナリストが XAI の原理に精通する必要もなくなり、上述の問題 (a)、(b) を解決することができる。

提案手法の有効性を確認するために、悪性サイト検知、及びマルウェア検知を対象とした実験を行った。実験の結果、異議判定モデルを用いることで、識別モデルによって悪性と識別されたものの中から実は良性だったものを特定したり、良性と識別されたものから実は悪性だったものを特定したりすることが効率良く行えることがわかった。

2. 関連研究

2.1 機械学習を用いたサイバー空間上の悪性活動検知

機械学習を用いてサイバー空間上の様々な悪性活動を検知する技術の提案は 10 年以上前から多岐にわたって行われてきた。例えば、多くのメールサービスで実利用されているスパムメール検知 [9] に加えて、マルウェア対策用のソフトウェアやサービスで使われているマルウェア検知 [10,11] や悪性サイト (URL やドメイン名) 検知 [12-14] がある。これらの検知技術は、対象の悪性活動の種類が異なっていたとしても、(1) 悪性／良性のラベルがついたデータセットを入力し、(2) 悪性／良性を識別可能な特徴量を設計し、(3) 機械学習アルゴリズム (例: ナイブベイズ, SVM, ランダムフォレスト, ニューラルネットワーク) で識別モデルを作成し、(4) 新規のデータを識別モデルで悪性／良性に二値分類した結果を利用する、という 4 段階で構成される。

一方で、このような検知技術の提案においては、悪性活

動を高精度に検知することに焦点が当てられており、識別モデルを利用する主体である人間（アナリスト）の存在が必ずしも意識されていない。そのため、識別モデルによる検知結果を知ったアナリストによる実際の運用（誤検知や見逃しへの対応）についての議論が不足している。こうした問題は悪性活動の検知に限らず機械学習全般に共通するものであり、それを解決しようとする手法の一つが、機械学習の判断根拠を説明しようとする説明可能 AI（eXplainable AI: XAI）と呼ばれる手法である。

2.2 機械学習の判断根拠説明

識別モデルの判断根拠を説明する直接的なアプローチは、もともと解釈性が高いホワイトボックスな識別モデルを作成することである。例えば、線形モデル、決定木、ルールリスト、ルールセットなどが識別モデルとして使用される。しかし、このアプローチでは、ブラックボックスな識別モデルによる判断根拠を事後的に説明することはできない。ブラックボックスな識別モデルの判断根拠説明を与えるアプローチは以下の 2 つに大別される。

一つ目は、識別モデルそのものに対する説明を与えようとするアプローチである。このアプローチは大域的説明と呼ばれ、識別モデルの入出力関係を解釈性の高いモデルで近似する手法がある [15, 16]。二つ目は、与えられた入力に対する識別モデルの識別結果について、その判断根拠に対する説明を個別的に与えようとするアプローチである。このアプローチは局所的説明と呼ばれる。識別モデルそのものに対する説明ではなく、個々の識別結果に対する個別的な説明が求められる場合には局所的説明のアプローチが有効である。本研究では局所的説明による説明結果を異議判定モデルの特徴量（XAI 特徴量）として利用する。以下では、局所的説明の代表的な手法である LIME [6] と、その拡張手法である SHAP [7] について説明する。

LIME は、与えられた入力に対する予測を行う際に、識別モデルが重要視する特徴量を出力するアルゴリズムである。このアルゴリズムでは、まず、識別モデルへの入力に対してランダムな摂動を加えたときの出力を取得する。そして、この入力と出力の組からなるデータセットで局所的リッジ回帰を行い、回帰係数の大小関係に基づいて識別モデルで用いられている各特徴量の重要度を算出する。

SHAP は、LIME と同様、ランダムな摂動が加えられた入力と、それに対する識別モデルの出力を得ることで、識別モデルで用いられている各特徴量の重要度を算出するアルゴリズムである。重要度の算出には、ゲーム理論におけるシャプレイ値が用いられている。これにより、特徴量間の相関を加味し、LIME よりも詳細な判断根拠説明を出力することが可能となる。しかし、特徴量の数が増えると計算量が膨大になるという欠点があり、大量の識別結果に対する説明を行う必要がある場合には不向きである。

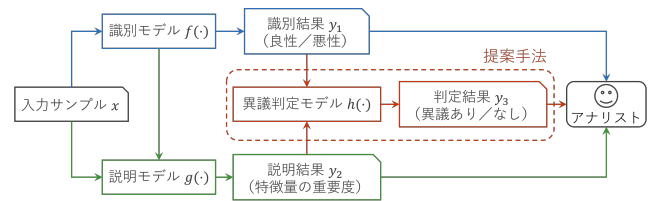


図 1 提案手法の概観

3. 提案手法

本研究では、識別モデルが出力した悪性／良性活動の識別結果のうち、人間による検証が必要となるものを特定するための異議判定モデルを提案する。異議判定モデルの学習においては、識別モデルの出力である識別結果と説明モデルの出力である説明結果を特徴量として用いる。

3.1 異議判定モデルの構造

本研究で提案する異議判定モデルと、既存の識別モデルや説明モデルとの関係を図 1 に示す。

識別モデル $f(\cdot)$ では、入力されたサンプル x が悪性か良性かを識別する。サンプル x を識別モデル $f(\cdot)$ に入力すると識別結果 $y_1 = f(x)$ が出力される。識別結果 y_1 は、サンプル x が悪性であるか、良性であるかの度合いを表す連続値（悪性の確信度を p 、良性の確信度を $1 - p$ とした連続値）として与えられる。アナリストはこの識別結果を観測することで、サンプル x が悪性であるか良性であるかについての情報を得ることができる。

識別モデル $f(\cdot)$ による出力結果の判断根拠を知るためには説明モデル $g(\cdot)$ を利用する。説明モデルには、サンプル x と識別モデル $f(\cdot)$ が入力され、説明結果 $y_2 = g(x, f(\cdot))$ が出力される。説明モデルの構築には LIME や SHAP といった説明手法を用いる。LIME や SHAP の場合、説明結果 y_2 は識別モデルに使用された各特徴量の重要度が数値として与えられる。アナリストは、識別モデル $f(\cdot)$ から得られた識別結果 y_1 に加え、説明モデル $g(\cdot)$ から得られた説明結果 y_2 を観測することでサンプル x に対する識別結果と共に、その識別結果が得られた判断根拠についての情報を得ることができる。

本研究で提案する異議判定モデル $h(\cdot)$ では、識別モデルによる識別結果 y_1 と説明モデルによる説明結果 y_2 を特徴量として入力し、判定結果 $y_3 = h(y_1, y_2)$ を出力する。判定結果 y_3 は、識別結果 y_1 が誤っている（異議あり）か、誤っていない（異議なし）かを表す連続値（異議ありの確信度を q 、異議なしの確信度を $1 - q$ とした連続値）として与えられる。アナリストは異議判定結果 y_3 を観測することで、識別結果 y_1 に対する異議の有無についての情報を得ることができる。これにより、正誤を検証する識別結果の優先順位を決定し、アナリストの労力を削減することが

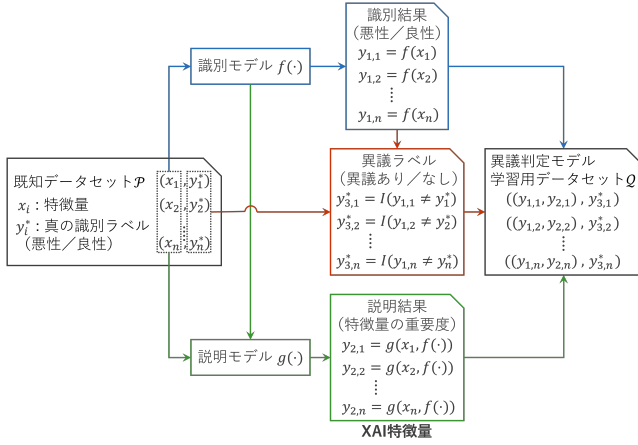


図 2 異議判定モデル学習用データセットの作成方法

できる．なお，説明モデルによる説明結果は y_2 は，識別モデルによる識別結果 y_1 から得られた判断根拠を人間が解釈するために用いるのが本来意図された利用用途である．これに対し本研究では，本来の利用用途とは異なり，説明モデルによる説明結果を人間が解釈するだけではなく，異議判定モデルの特徴量としても用いる点に特色がある．

3.2 異議判定モデルの学習

異議判定モデルを学習するために用いるデータセット Q を作成する手順を図 2 に示す．

データセット Q を作成するために必要なのは，真の識別ラベルが既知であるデータセット P ，識別モデル $f(\cdot)$ ，説明モデル $g(\cdot)$ である．データセット $P = \{(x_i, y_i^*)\}_{i=1,2,\dots,n}$ は，調査の対象となるサンプルを特徴付ける m 次元の特徴量ベクトル $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$ とその真の識別ラベル $y_i^* \in \{\text{悪性}, \text{良性}\}$ の組からなる n 個のサンプルの集合である．

データセット Q を作成するためには，まず， x_i を識別モデル $f(\cdot)$ と説明モデル $g(\cdot)$ に入力し，識別結果 $y_{1,i} = f(x_i) \in \{\text{悪性}, \text{良性}\}$ と説明結果 $y_{2,i} = g(x_i, f(\cdot))$ を生成する．本研究においては，説明結果 $y_{2,i}$ は特徴量ベクトル x_i と同じく m 次元ベクトルとして与えられる．具体的には，説明結果 $y_{2,i}$ の各要素は，特徴量ベクトル x_i の各要素に対応する特徴量 $x_{i,j}$ の重要度である．特徴量 $x_{i,j}$ の重要度の算出には任意の XAI を使用することができる．通常，説明結果 $y_{2,i}$ は，識別結果 $y_{1,i}$ の判断根拠を人間が解釈するために参照される．これに対し本研究では，識別結果と説明結果の組 $(y_{1,i}, y_{2,i})$ を，異議判定モデルの学習における特徴量として用いる．

次に，識別結果 $y_{1,i}$ を真の識別ラベル y_i^* と比較し，異議ラベル $y_{3,i} = I(y_{1,i} \neq y_i^*)$ を生成する．ただし， $I(\cdot)$ は括弧内が真のときに 1 (異議あり)，偽のときに 0 (異議なし) となる指示関数である．すなわち，識別モデルによる識別結果 $y_{1,i}$ と真の識別ラベル y_i^* が一致しなければ「異

議あり」，一致すれば「異議なし」となるラベルを生成する．以上の手順により，異議判定モデルを学習するために用いるデータセット $Q = \{(y_{1,i}, y_{2,i}), y_{3,i}\}_{i=1,2,\dots,n}$ が作成される．

異議判定モデル $h(\cdot)$ は，データセット Q を用い，特徴量 $(y_{1,i}, y_{2,i})$ に対する異議判定モデルの出力 $h(y_{1,i}, y_{2,i})$ が異議ラベル $y_{3,i}$ に一致するように学習される．異議判定モデルの学習には任意の教師あり機械学習アルゴリズムを使用することができる．なお，異議判定の精度を高めるための工夫として，データセット Q に含まれるデータのうち，異議判定が生じている可能性が高い（つまり識別結果に誤検知や見逃しが生じている可能性が高い）データのみを用いて異議判定モデルの学習を行う．

4. 評価

4.1 データセット

本節では，悪性サイト検知およびマルウェア検知に提案手法を適用した際の有効性を評価する．異なる種類の悪性活動の検知を評価対象にすることで提案手法の汎用性を確認する．

4.1.1 悪性サイト検知のデータセット

文献 [17] では，ドメイン名の利用状況の時系列変動に着目した特徴量設計を行うことで，99%の精度で悪性ドメイン名と良性ドメイン名を識別できる DomainProfiler と呼ばれる識別モデルを提案している．文献 [17] は，悪性サイト検知に関する広範なサーベイ論文 [13, 14] で取り上げられている．文献 [17] において DomainProfiler の学習に用いられたデータセット（以下，DomainProfiler データセット）は非公開であるが，本研究では，文献 [17] の著者らから DomainProfiler データセットの提供を受け，提案手法の評価に使用した．

本実験で利用する DomainProfiler データセットは，227,283 個のドメイン名（悪性ドメイン名 135,720 個，良性ドメイン名 91,563 個）で構成される．これらのドメイン名は，DomainProfiler [17] の訓練データのラベリング負荷を下げる富士らによる研究 [18] で評価に利用されているものと同一の条件で取得されている．具体的には，悪性ドメイン名は，2018 年 5 月 29 日時点での公開ブラックリスト hpHosts [19] に掲載されていたドメイン名から選定されている．このブラックリストには，攻撃サイト，マルウェア配布サイト，フィッシングサイトとして利用されていることが確認された悪性ドメイン名が含まれる．良性ドメイン名は，2018 年 5 月 29 日時点で人気 Web サイトランキングである Alexa Top Sites [20] に掲載されていた人気上位 10 万件のドメイン名から選定されている．なお，悪性ドメイン名に良性ドメイン名が誤って混入する可能性，及び，良性ドメイン名に悪性ドメイン名が混入する可能性を極力排除するため，評価時点において，VirusTotal [21] や

複数の商用セキュリティデータを参照して検証済みのドメイン名のみがデータセットに採用されている。また、データセットの各ドメイン名には 128 個の特徴量が付与されている。具体的には、文献 [18] と同様に作成した 115 個の特徴量（時系列変動を表現する TVP 特徴量 80 個、関連 IP アドレスを表現する rIP 特徴量 18 個、関連ドメイン名を表現する rDomain 特徴量 17 個）に加え、文献 [22] でドメイン名のカテゴリを特定するのに有用であった特徴量 13 個が付与されている。

4.1.2 マルウェア検知のデータセット

Endgame 社は、マルウェア検知の研究に利用できるベンチマークデータを充実させる目的で、EMBER (Endgame Malware BEenchmark for Research) と呼ばれるデータセット（以下、EMBER データセット）を公開している [23]。本研究では、EMBER データセットを用い、マルウェア検知に提案手法を適用した際の有効性を評価する。なお、EMBER データセットは、マルウェア検知に関する広範なサーベイ論文 [11] で取り上げられている。また、サーベイ論文 [11] で紹介されている複数の公開データセットのうち最も新しいものである。これらの理由から、提案手法の評価に EMBER データセットを用いることとする。

EMBER データセットは、Windows ポータブル実行ファイル (PE ファイル) から特徴量を抽出して作成されたものであり、VirusTotal によってスキャンされた悪性 PE ファイルと良性 PE ファイルの両方のサンプルが含まれている。各サンプルには、ファイルの SHA256 のハッシュ値、そのファイルが最初に観測された月、ラベル、及び PE ファイルから得られた特徴量が含まれている。EMBER2017 データセットには、2017 年以前にスキャンされた 110 万個のサンプルが含まれ、EMBER2018 データセットには、2018 年以前にスキャンされた 100 万個のサンプルが含まれている。本実験では、EMBER2018 データセットを用いる。EMBER2018 データセットに含まれる悪性 PE ファイルと良性 PE ファイルのサンプルは各々 40 万個であり、その他に悪性／良性のラベルが付与されていないサンプルが 20 万個ある。本研究では、悪性／良性のラベルが付与された 80 万個のサンプルを使用する。また、各サンプルの特徴量は 2,381 個である。

4.2 各種モデルの学習

4.2.1 識別モデルの学習

識別モデル $f(\cdot)$ を学習するための訓練データとして、DomainProfler データセットからは 181,826 個、EMBER データセットからは 160,000 個のサンプルを無作為に抽出したものを使用した。識別モデルの学習には任意の機械学習アルゴリズムを用いることができるが、本研究では簡便な実装により十分な精度を得られるランダムフォレストを採用した。ランダムフォレストは、複数の決定木の予測結

果を統合し最終的な予測結果を決定する機械学習アルゴリズムであり、多くの識別問題において高い精度を達成することが知られている [24]。なお、本研究では、統合する決定木の本数を 300 本とした。

4.2.2 説明モデルの学習

説明モデル $g(\cdot)$ としては、その説明結果が異議判定モデルの特徴量 (XAI 特徴量) として利用できるものであれば任意のアルゴリズムを用いることができる。ただし、大量の識別結果に対する XAI 特徴量を生成するためには、個々の識別結果に対する説明結果を現実的な計算量で出力できる必要がある。そこで本研究では、計算量を抑制するために、LIME により出力される各特徴量の重要度を XAI 特徴量として用いた。LIME の実装としては [25] を使用し、各種パラメータ値はデフォルトの設定とした。なお、LIME の拡張手法である SHAP を用いれば、特徴量間の相関を加味した詳細な説明結果を取得できるが、特徴量の数が増えると計算量が膨大になるという欠点があり、XAI 特徴量の生成手法としては不向きである。

4.2.3 異議判定モデルの学習

異議判定モデルは識別モデルが出力した識別結果の正誤を判定するために用いられる。したがって、どのような場合に誤識別が生じやすいのかということについて事前知識があれば、異議判定の精度を高められると期待できる。そのような事前知識は、例えば、学習済みの識別モデルが、その学習に用いた訓練データに対し、どのような場合に誤識別するのかという傾向を調べることで把握できる。実際、4.2.1 節において学習された識別モデルの場合、識別結果の確信度に明らかな特徴があった。具体的には、DomainProfler データセットについては、悪性ドメイン名と識別する確信度が 0.5 以上、0.9 以下の場合に誤識別する（すなわち、悪性ドメイン名と誤検知する）傾向があった。また、EMBER データセットについては、悪性 PE ファイルと識別する確信度が 0.1 以上、0.5 以下の場合に誤識別する（すなわち、悪性 PE ファイルを見逃す）傾向があった。そこで本研究では、この事前知識を踏まえ、識別モデルによる識別結果の確信度が上記の条件を満たす場合にのみ、異議判定モデルによる異議判定を行う。なお、誤識別の生じやすさの傾向は、識別モデルのモデル化、学習に使用する機械学習アルゴリズムや特徴量、訓練データに含まれるサンプル数やラベルの偏りなどに影響して変化するものと考えられるため、検知タスクに応じて都度調整する必要がある。

また、異議判定モデルの学習に用いるデータセット Q を作成する元となるデータセット P の対象も、同様の条件を満たすサンプルに限定する。DomainProfler データセットの場合、識別モデルの学習に用いなかった 45,457 個のサンプルを識別モデル $f(\cdot)$ に入力し、各サンプルに対する識別結果の確信度を算出した。その結果、398 個のサンプルが上

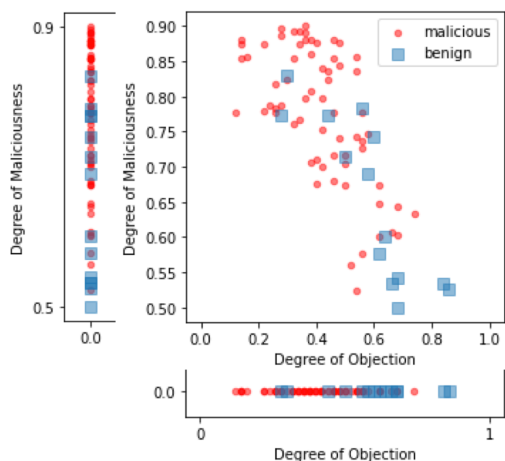


図 3 DomainProfiler データセットにおける異議判定結果

記の条件を満たした。そのうち、318 個のサンプルをデータセット $\mathcal{P}$ として使用し、残りの 80 個のサンプルはテストデータとして使用した。この 80 個のサンプルはすべて悪性ドメイン名として識別されたものであるが、その中に実際には良性ドメイン名であったものが 12 個含まれている。なお、上述の識別モデルの学習に用いなかった 45,457 個のサンプルに対する識別モデルの正解率 (accuracy) は 0.9976 であった。

EMBER データセットの場合、データセット $\mathcal{P}$ のサイズが DomainProfiler データセットと同規模になるように調整し、識別モデルの学習に用いなかった 2,000 個のサンプルを識別モデル $f(\cdot)$ に入力し、各サンプルに対する識別結果の確信度を算出した。その結果、402 個のサンプルが上記の条件を満たした。そのうち、321 個のサンプルをデータセット $\mathcal{P}$ として使用し、残りの 81 個のサンプルはテストデータとして使用した。この 81 個のサンプルはすべて良性 PE ファイルとして識別されたものであるが、その中に実際には悪性 PE ファイルであったものが 18 個含まれている。なお、上述の識別モデルの学習に用いなかった 2,000 個のサンプルに対する識別モデルの正解率 (accuracy) は 0.9565 であった。

本研究では、以上のようにデータセット $\mathcal{P}$ を作成した上で、3.2 節に示した手順に従いデータセット $\mathcal{Q}$ を作成し、異議判定モデルを学習した。なお、異議判定モデルの学習には任意の機械学習アルゴリズムを用いることができるが、本研究では、識別モデルの学習と同様、ランダムフォレストを利用した。また、ランダムフォレストを構成する決定木の本数は 50 本とした。

4.3 異議判定の精度評価

DomainProfiler データセット、EMBER データセットについて、上記のテストデータを用いて異議判定モデルの有効性を評価した結果を図 3、4 にそれぞれ示す。図中の各

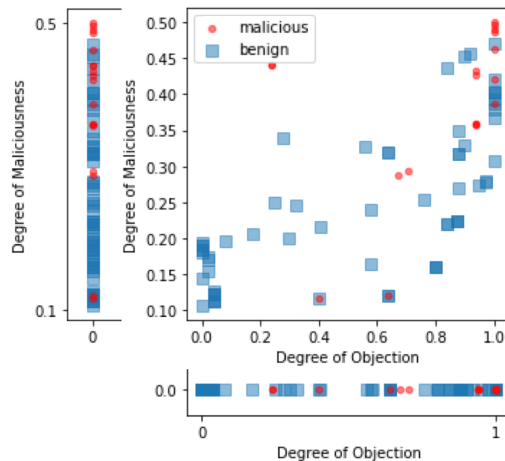


図 4 EMBER データセットにおける異議判定結果

点は、上記のテストデータに含まれる各サンプルに対応しており、赤点が悪性サンプル、青点が良性サンプルを表す。散布図の縦軸は入力されたサンプルを識別モデルにより識別した際の出力結果を表している。出力結果は 0 から 1 の範囲の連続値で与えられており、数値が高いほど悪性と識別する確信度が高くなる。また、散布図の横軸は入力されたサンプルを異議判定モデルにより異議判定した際の出力結果を表している。出力結果は 0 から 1 の範囲の連続値で与えられており、数値が高いほど異議ありと判定する確信度が高くなる。なお、散布図の左側に示された一列に並ぶ点の分布は、散布図中の各点を縦軸上に射影したものである。同様に、散布図の下側に示された一列に並ぶ点の分布は、散布図中の各点を横軸上に射影したものである。

上述の通り、DomainProfiler データセットに対する結果 (図 3) の場合、テストデータに含まれるサンプル数は 80 個である。この 80 個のサンプルは識別モデルによってすべて悪性ドメイン名と識別されたものである。しかしその中には、実際は良性ドメイン名であるにもかかわらず、悪性ドメイン名として誤識別されたものが 14 個含まれる。また、EMBER データセットに対する結果 (図 4) の場合、テストデータに含まれるサンプル数は 81 個である。この 81 個のサンプルは識別モデルによってすべて良性 PE ファイルと識別されたものである。しかしその中には、実際は悪性 PE ファイルであるにもかかわらず、良性 PE ファイルとして誤識別されたものが 18 個含まれている。以下では、これらの誤識別されたサンプルを特定する「労力」を評価する。

まず、異議ありと判定する確信度が 0.5 以上のサンプルを「異議あり」、0.5 未満のサンプルを「異議なし」と判定し、異議ありと判定されたサンプルについてののみアナリストによる検証を行う場合を想定した評価を行う。DomainProfiler データセットの場合、異議ありと判定されたサンプルは 31 個であり、そのうち、良性ドメイン名であったものは 11 個であった。したがって、異議ありと判定されたサンプルに

含まれる良性ドメイン名の割合（異議ありと判定したときの正解率）は 0.35 (= 11/31) である。また、異議なしと判定されたサンプルは 49 個であり、そのうち、良性ドメイン名であったものは 3 個であった。したがって、異議なしと判定されたサンプルに含まれる良性ドメイン名の割合（異議なしと判定したときの不正解率）は 0.06 (= 3/49) である。一方、EMBER データセットの場合、異議ありと判定されたサンプルは 52 個であり、そのうち、実際に悪性 PE ファイルであったものは 15 個であった。したがって、異議ありと判定されたサンプルに含まれる悪性 PE ファイルの割合は 0.28 (= 15/52) である。また、異議なしと判定されたサンプルは 29 個であり、そのうち、悪性 PE ファイルであったものは 3 個であった。したがって、異議なしと判定されたサンプルに含まれる悪性 PE ファイルの割合は 0.10 (= 3/29) である。以上より、異議なしと判定されたサンプルは検証の対象とせず、異議ありと判定されたサンプルのみを検証の対象とすれば、検証すべきサンプル数を削減でき、誤識別されたサンプルを特定できる確率も高まることがわかる。識別結果をアナリストが検証する際の労力は検証するサンプル数に比例するため、提案手法は有効であるといえる。

ここで、異議判定モデルからの出力（異議ありと判定する確信度）を参照せずに、識別モデルからの出力（悪性と識別する確信度）のみを参照した上で、アナリストによる検証の対象を決める場合について考察する。この場合、悪性と識別する確信度が 0.5 に近いものから順に検証の対象とするのが自然である。一方、異議判定モデルからの出力を参照しなければ、図 3, 4 の散布図を得ることはできず、左側に示された一列に並ぶ点の分布を参照することになる。ただし、本図においては、赤点（悪性）と青点（良性）が区別されて表示されているが、アナリストはこの情報を事前に知ることができない。このことを考慮すると、悪性と識別する確信度が 0.5 に近いものから順に、いくつかのサンプルを検証の対象とするべきかを決めるのは困難であることがわかる。これに対し、異議判定モデルからの出力を参照するならば、異議ありと判定する確信度が 0.5 以上のサンプルのみを検証の対象とするという明確な指針が得られる。上述のように、この指針に基づくことによって誤識別されたサンプルを効率的に特定できるため、提案手法はアナリストの労力を削減する上で有効であるといえる。

4.4 ケーススタディ

異議判定モデルにより異議あり／異議なしと判定されたサンプルについて、実用を意識したケーススタディを示す。

4.4.1 DomainProfiler データセットにおける事例

図 5, 図 6 はそれぞれ、DomainProfiler データセットにおける異なる二つのサンプル（ドメイン名）に対する LIME の説明結果を示したものである。図中左の「Prediction

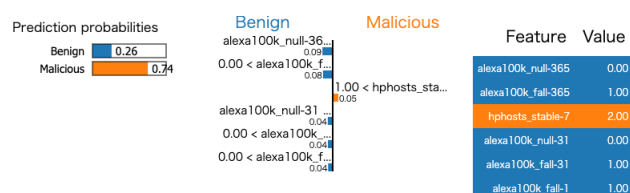


図 5 DomainProfiler データセットにおける「異議あり」の実例

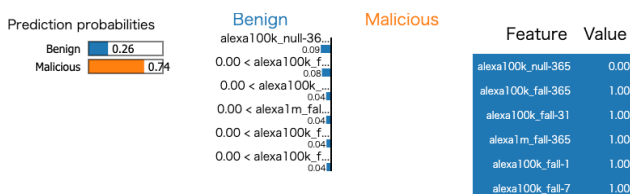


図 6 DomainProfiler データセットにおける「異議なし」の実例

probabilities」は識別モデルによる識別結果を示したものであり、良性（Benign）と識別する確信度と悪性（Malicious）と識別する確信度を示している。この情報より、この二つのサンプルは、いずれも悪性ドメイン名として識別する確信度が高く、その値には違いがないことがわかる。また、図中央のヒストグラムと図中右の表は、重要度が上位となる特徴量を示したものである。これらの情報を総合すると、この二つのサンプルは「重要度が高い特徴量の多くは良性ドメイン名として識別することを支持するが、結果的には悪性ドメイン名として識別された」と解釈するのが自然である。しかし、この解釈は論理的に整合しない。そのため、いずれのサンプルについても誤識別が疑われ、アナリストによる検証の対象にすべきと考えられる。

一方、前節に示した実験において、この二つのサンプルに対する異議判定は、図 5 のサンプルは異議あり、図 6 のサンプルは異議なしと判定された。実際、図 5 のサンプルは良性ドメイン名、図 6 のサンプルは悪性ドメイン名であり、異議判定の結果は妥当である。このように、識別結果と説明結果を照らし合わせ、識別結果の正誤を区別することは人間にとっては直感的ではなく、困難である場合がある。本研究で提案する異議判定モデルは、説明結果に生じるわずかな差異を学習することで、人間にとっては気が付きにくい情報であっても、悪性活動の誤検知や見逃しの効率的な特定に有効活用できているといえる。

4.4.2 EMBER データセットにおける事例

図 7, 図 8 はそれぞれ、EMBER データセットにおける異なる二つのサンプル（PE ファイル）に対する LIME の説明結果を示したものである。この図に示された情報を総合すると、この二つのサンプルは「重要度が上位の特徴量は、重要度の値は低いものの、良性 PE ファイルとして識別することを支持していることから、結果的に良性 PE ファイルとして識別された」と解釈するのが自然である。この解釈は論理的に整合しており、いずれのサンプルも誤識別は

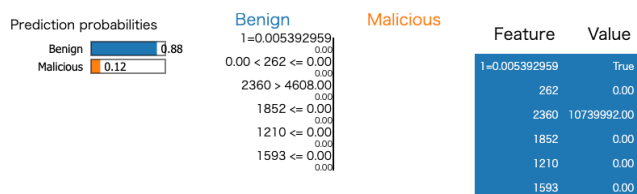


図 7 EMBER データセットにおける「異議あり」の実例

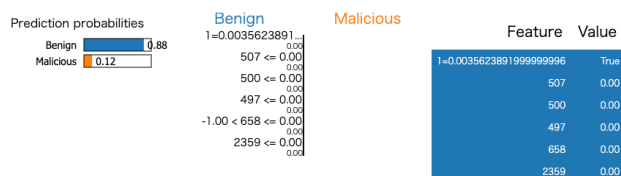


図 8 EMBER データセットにおける「異議なし」の実例

強くは疑われない。しかし、重要度が上位の特徴量であっても重要度の値が低くなっていることから、誤識別の可能性も残される。そのため、アナリストによる検証の対象とすべきかどうかの判断は難しい。

一方、前節に示した実験において、この二つのサンプルに対する異議判定は、図 7 のサンプルについては異議あり、図 8 のサンプルについては異議なしと判定された。実際、図 7 のサンプル悪性 PF ファイル、図 8 のサンプルは良性 PE ファイルであり、異議判定の結果は妥当である。したがって、DomainProfiler データセットにおける実例と同様、EMBER データセットにおいても、誤識別された悪性活動の特定に異議判定モデルは有効であるといえる。

5. おわりに

本研究では、サイバー空間における悪性活動を機械学習で検知する際の誤検知や見逃しを効率的に特定するために、XAI の出力を特徴量として利用する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。今後の課題としては、検知の対象となる悪性活動の特性が時間変化する場合（コンセプトドリフト）への対応がある。

謝辞 本研究の一部は、日本学術振興会における科学研究費補助金基盤研究（C）（課題番号 20K11800）による支援を受けている。ここに記し謝意を表す。

参考文献

- [1] Sundaramurthy, S. C. et al.: A Human Capital Model for Mitigating Security Analyst Burnout, *Proc. SOUPS* (2015).
- [2] Kokulu, F. B. et al.: Matched and Mismatched SOC's: A Qualitative Study on Security Operations Center Issues, *Proc. ACM CCS* (2019).
- [3] Vielberth, M. et al.: Security Operations Center: A Systematic Study and Open Challenges, *IEEE Access*,

- Vol. 8, pp. 227756–227779 (2020).
- [4] Buczak, A. L. and Guven, E.: A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 2, pp. 1153–1176 (2016).
- [5] Shaukat, K. et al.: A Survey on Machine Learning Techniques for Cyber Security in the Last Decade, *IEEE Access*, Vol. 8, pp. 222310–222354 (2020).
- [6] Ribeiro, M. T., Singh, S. and Guestrin, C.: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, *Proc. KDD*, p. 1135–1144 (2016).
- [7] Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, *Proc. NeurIPS*, pp. 4765–4774 (2017).
- [8] Lage, I. et al.: Human-in-the-Loop Interpretability Prior, *Proc. NeurIPS*, Vol. 31 (2018).
- [9] Egele, M. et al.: A Survey on Automated Dynamic Malware-Analysis Techniques and Tools, *ACM Comput. Surv.*, Vol. 44, No. 2 (2008).
- [10] Ye, Y., Li, T., Adjeroh, D. and Iyengar, S. S.: A Survey on Malware Detection Using Data Mining Techniques, *ACM Comput. Surv.*, Vol. 50, No. 3 (2017).
- [11] Aslan, c. A. and Samet, R.: A Comprehensive Review on Malware Detection Approaches, *IEEE Access*, Vol. 8, pp. 6249–6271 (2020).
- [12] Gardiner, J. and Nagaraja, S.: On the Security of Machine Learning in Malware C&C Detection: A Survey, *ACM Comput. Surv.*, Vol. 49, No. 3 (2016).
- [13] Zhauniarovich, Y., Khalil, I., Yu, T. and Dacier, M.: A Survey on Malicious Domains Detection through DNS Data Analysis, *ACM Comput. Surv.*, Vol. 51, No. 4 (2018).
- [14] Sahoo, D. et al.: Malicious URL Detection using Machine Learning: A Survey (2019). arXiv:1701.07179.
- [15] Vidal, T., Pacheco, T. and Schiffer, M.: Born-again Tree Ensembles, *Proc. ICML*, pp. 9846–9856 (2020).
- [16] Hara, S. and Hayashi, K.: Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach, *Proc. AISTATS*, pp. 77–85 (2018).
- [17] Chiba, D. et al.: DomainProfiler: Discovering Domain Names Abused in Future, *Proc. IEEE/IFIP DSN*, pp. 491–502 (2016).
- [18] Fukushi, N. et al.: Exploration into Gray Area: Toward Efficient Labeling for Detecting Malicious Domain Names, *IEICE Transactions on Communications*, Vol. E103.B, No. 4, pp. 375–388 (2020).
- [19] hpHosts, <http://www.hosts-file.net/>. [Online; accessed 29-May-2018].
- [20] Alexa Top Sites, <http://www.alexa.com/topsites>. [Online; accessed 29-May-2018].
- [21] VirusTotal, <https://www.virustotal.com/>. [Online; accessed 29-May-2018].
- [22] Chiba, D. et al.: DomainChroma: Building actionable threat intelligence from malicious domain names, *Computers & Security*, Vol. 77, pp. 138–161 (2018).
- [23] Anderson, H. S. and Roth, P.: EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models (2018). arXiv:1804.04637.
- [24] Fernández-Delgado, M. et al.: Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, *Journal of Machine Learning Research*, Vol. 15, No. 90, pp. 3133–3181 (2014).
- [25] LIME Library, <https://github.com/marcotcr/lime>. [Online; accessed 19-August-2021].