静的特性アンサンブルを用いたマルウェアの分類

ダオ・ヴァン・トゥアン[1,a)]　佐藤　浩[1,b)]　久保　正男[1,c)]　中村　康弘[1,d)]

概要：近年，マルウェアの脅威が著しく増大しつつある．悪意のあるプログラムの数や巧妙さが増しているため，従来のシグネチャベースの技術では，新しいマルウェアの亜種を検出することができなくなっている．検知技術の進化に連れてマルウェアの検出率は向上しているが，それぞれのマルウェアをファミリー別に分類することは，依然として困難である．従来の分析手法は多くの時間的・空間的リソースを要するが，機械学習は少ないリソースでこの問題を解決できる．本研究では，標準的な機械学習アルゴリズムを用いたマルウェア分類のために，レジスタとオペコードを含むアンサンブルの静的特性を提供する．そして，特徴空間に次元削減を適用することによって，実世界のマルウェアをより高い精度で分類することができた．さらに，適切な特徴を選択することが，マルウェアの分類タスクに大きく影響を与えることが分かった．

キーワード：マルウェア分類，機械学習，逐次特徴選択

Malware Classification Using Ensemble of Static Characteristics

TUAN DAO VAN[1,a)]　Hiroshi Sato[1,b)]　Masao Kubo[1,c)]　Yasuhiro Nakamura[1,d)]

Abstract: The threat of malware has increased significantly in recent years. Due to the increasing number and sophistication of malicious programs, traditional signature-based techniques can no longer detect new malware variants. Although malware detection rates have improved as detection technologies have evolved, it does not remain easy to classify each malware by family. Traditional analysis methods require many temporal and spatial resources, while machine learning can solve this problem with fewer resources. In this study, we provide an ensemble static characteristics set including opcode and register for malware classification using standard machine learning algorithms. Then, we could classify real-world malware with higher accuracy by applying dimensionality reduction to the feature space. Furthermore, we found that the selection of appropriate features has a significant impact on the malware classification task.

Keywords: Malware Classification, Machine Learning, Sequential Feature Selection

1. はじめに

マルウェアは近年，指数関数的な増加を遂げている．Kaspersky の検知システムは，2021 年に 1 日平均 38 万個の新しいマルウェアを発見しており，2020 年と比較して 1 日あたり 2 万個の増加分となっている [1]． 2021 年の SonicWall のレポートによると，検出されたマルウェアのうち，268,362 個が 2020 年以前に見られたことのないものがある [2]．

その主な理由は，TheFatRat [3]，Arbitrium-RAT [4] などのオープンソースの無料マルウェア作成ツールにアクセスすることで，容易に新しい亜種のマルウェアを作成することができるからである．さらに，マルウェアの作者は，検知を回避するために，悪意のあるコンポーネントにポリモーフィズムを利用している [5]．つまり，同じマルウェアファミリーに属し，同じ形態の悪意ある動作をするマルウェアは，常に様々な手法で修正されたり，難読化された

[1] 防衛大学校理工学研究科
Graduate School of Science and Engineering, National Defense Academy of Japan
a) ed21006@nda.ac.jp
b) hsato@nda.ac.jp
c) masaok@nda.ac.jp
d) yas@nda.ac.jp

りする．その結果，従来のシグネチャベースの手法では，このマルウェアへの対処が追いつかない．

解析者は，マルウェアを正しいファミリーに分類するために，まずマルウェアの特性と挙動を理解する必要がある．悪意のあるコードの特性を見つけるために一般的に使用される基本的な方法には，静的解析と動的解析の 2 つがある．静的解析では，悪意のあるコードを実行せずに探索するが，アセンブリを読み取るための高度な解析や，非常に複雑で暗号化された悪意のあるコードをデバッグする前にアンパックする手法が必要となる．さらに，マルウェアの作成者は，高度なプログラミングやパック，難読化ツールを使用してマルウェアを生成するため，解析はより困難になる． その結果，この方法は多くの場合，時間と労力を必要とする．

動的解析では，仮想環境を通じて悪意のあるコードの挙動を捕捉し，ログを取る．この方法は，プロセスの作成，ファイル，レジストリの操作を正しく明らかにし，メモリ，変数，レジスタの実際の値を含むため，静的解析よりも性能が高く，静的アプローチのように解析に時間がかからないことが多い．しかし，ファイル，レジストリキー，プロセスをチェックして仮想環境を検出し，犠牲者がターゲットでないことを検出すると，直ちにすべての活動を停止する悪質なコードが多くなってきている．さらに，いくつかのマルウェアは，実行のために大量なリソースを必要とする．

静的解析と動的解析の両方の限界に対処するため，セキュリティベンダーや研究者は，近年，機械学習やニューラルネットワークアプローチを採用することが一般的である．研究者は，静的な特性から様々な機械学習ベースの手法でマルウェアを分類することができる．

機械学習では，入力データがシステムの性能を決定する重要な要因の一つである．これまでの一部の研究者においては，API コール，API 引数，命令列，文字列情報などの高レベルの特性に着目する一方で，オペコードやレジスタなどの低レベルの特性を利用する研究を欠落させている．

また，高レベルの特性については，自然言語処理による分類が進んでいる研究が多くある．しかし，暗号化だけでなく複数の難読化手法を用いるマルウェアにはこの手法が適用できず，得られるデータはノイズが多く，シーケンスも乱れてしまうという問題があった [6]．

低レベルの特徴量としては，オペコードの並びから機械学習用の特徴量を作成する N-gram が一般的である [7]．レジスタを用いた代表的な手法として VSA（Value Set Analysis）があり，メタモーフィックマルウェアの検出において，検出率最大 100 %の精度で有効であることが証明された [8, 9]．

なお，マルウェア自身が，隠れているかどうかに関わらず，シーケンスを変更したり，ノイズを追加したりしても，元のコードの有害な機能は保持される．また，マルウェアがレジスタの再割り当て技術を適用しても，レジスタ間の相関関係は変化しない．このため，マルウェアファミリーごとに，オペコードとレジスタの両方に一定の関係があることになる．

本研究では，低レベルの特徴に着目する．低レベルの特徴を用いると，ファミリーごとの微小な差異を認識できるだけなく，高レベルの特徴を用いるよりも処理時間が短くなるためである．悪意のあるコードを逆コンパイルした際に得られる典型的な ASM ファイルよりオペコードとレジスタのデータを両方抽出し，逐次特徴選択アルゴリズムにより，強い相関で重要度の高い特徴群を得ることで，分類結果をさらに改善することができた．

2. 関連研究

Yenboah ら [7] は，異なる n-gram サイズの オペコードシーケンスから生成されるアンサンブル特徴量を採用している．著者らは，あらかじめ定義された重みのセットに対してグリッド検索を適用し，アンサンブル特徴セットに対する最適な重みの組み合わせを求め，2,000 サンプルのバランスデータセットに対し，Random Forest を用いて最高の検出精度 98.1%を達成した．

Rad ら [12] は，マルウェアの統計的特徴であるオペコードに基づく分類法を提案及び検証している．決定木を用いて学習及び評価することは，高い性能を発揮する一方，データ不足によるオーバーフィッティングの可能性があると考えられる．

Li ら [10] は，予備実験により，単一のレジスタに着目するだけでは良好な検出結果が得られないことを証明した．著者らは，レジスタを一定の順序で利用する．EAX, EBX, ECX, EDX, EBP, ESP, ESI, EDI の順で利用し，単純な CNN 層と 3 つの基本 LSTM 層で構成される．これらは，Accuracy，Precision，Recall，F1 など，すべての検出方法において，オペコードシーケンスよりも高い性能を達成している． しかし，機械学習アルゴリズムを用いた場合，最も高い検出精度は 0.796 であり，また，レジスタはマイクロアーキテクチャレベルの特徴であり，次元空間に隠蔽される前の情報を直接反映できないため，単純な機械学習アルゴリズムでは悪質行為の検出は困難であると著者は述べている．また，この論文では，なぜ上記のような固定された順番で実験を行ったかについては説明されていない．

Shiasi ら [8] は，実行ファイルを通してレジスタ値の分布と変化を追跡するために Value Set Analysis（VSA）のアイデアを適用し，実験により著者の提案手法が多様なデータセットのサンプルを低い偽陽性で識別することに成功したことを示し，マルウェアと良性ソフトウェアの区別において，平均 95%以上の精度を達成した．しかし，制御された環境でマルウェアバイナリを実行・監視する場合にのみ適用可能である．また，マルウェアの悪意ある活動に重要

な 6 つの DLL をベースにしているが，レジスタグループの数が比較的多いため，計算量が多くなってしまうという短所があった．

3. 逐次特徴選択アルゴリズム

逐次特徴選択アルゴリズムとは，初期の d 次元特徴空間を k 次元特徴部分空間（k <d）に縮小するために用いられる貪欲な探索アルゴリズムの一群である．逐次特徴選択アルゴリズムを適用して，入力の次元を適切な数の特徴に削減する．特徴量の選択の目的は 2 つある．計算効率の向上と，無関係な特徴やノイズを除去してモデルの汎化誤差を低減することである [13].

逐次特徴選択アルゴリズムでは，分類器の性能に基づいて，目的の大きさ k の特徴部分集合に達するまで，一つずつ特徴を削除または追加する SFS – Sequence Forward Selection(変数増加法)，及び SBS – Sequential Backward Selection(変数減少法) を使用する．SFS 及び SBS の拡張アルゴリズム Sequential Forward Floating Selecion (SFFS)，Sequential Backward Floating Selection (SBFS) にも使われている．Floating 手法はより多くの特徴量のサブセットの組み合わせをサンプリングできるように，一度含まれた（または除外された）特徴量を除外または包含のステップが追加されている．本研究ではすべての特徴選択手法を試し，最も高精度となる手法について検討する．

4. 提案手法

本研究の提案手法を図 1 に示す．

まず，逆アセンブラ IDA Pro と objdump コマンドの逆アセンブル機能を用いて，既知のマルウェアを逆アセンブルし，アセンブリソースファイルを取得する．そして，ASM ファイルから `MOV`，`PUSH`，`CALL`，`POP` などの 19 種類の共通オペコードと 8 種類のレジスタで合計 27 特徴を抽出する．統計量を抽出した後，逐次特徴選択アルゴリズムを用いて，すべての特徴量（27 個）の集合から，適切な特徴量の部分集合（k 個）を選択することで精度向上を図る．

本研究では k-NN，Nearest Centroid，SVM といった機械学習の典型的な分類器アルゴリズムを使用する．本手法を評価するために，10-fold Cross-Validation を利用する．これは，10 個のサブサンプルのうち 1 個を検証データとして持ち出し，残りの 9 個のサブサンプルを学習データとして使用する．このプロセスを 10 回繰り返し，10 個のサブサンプルそれぞれを検証データとして使用する．10 回の結果の平均がその手法の品質となる．

5. 実験と結果

5.1 実験データ

本研究はマルウェア 5 ファミリー，総サンプル数 8942 のデータセット [11] を使用した．それぞれのマルウェアファミリーの数を表 1 に示す．Winwebsec ファミリーのサンプル数が全体の半分近くを占めており，アンバランスなデータセットであることがわかる．そこで，Accuracy による性能評価に加え，F-score によるモデルの評価も行う．

表 1　マルウェアデータセット [11]
Table 1　Malware dataset [11]

Malware family	Sample 数
Locker	302
Mediyes	1450
Winwebsec	4400
Zbot	2100
Zeroaccess	690

表 2　各逐次特徴選択手法による精度の結果
Table 2　Comparison of Sequential Feature Selection method with each classifier

特徴選択手法 ＼ アルゴリズム	k-NN	Nearest Centroid	SVM
SFS	**98.88**	82.22	98.49
SBS	98.77	**90.78**	**98.60**
SFFS	98.49	82.06	98.38
SBFS	98.49	90.89	**98.60**

5.2 各逐次特徴選択の比較

4 つの逐次特徴選択アルゴリズムの比較を表 2 に示す．特徴選択手法の SFS とし，k-NN 分類器での一番良い精度が得られた．選択された特徴量の数に対する精度の変化を図 2 に示す．図 2 によると，27 の特徴量より 12 の特徴が選択されることで一番良い結果が得られた．この時，選択されたオペコードとレジスタの組み合わせは ‘`mov`’, ‘`push`’, ‘`xor`’, ‘`sub`’, ‘`inc`’, ‘`dec`’, ‘`imul`’, ‘`or`’, ‘`shr`’, ‘`shl`’, ‘`edx`’, ‘`esp`’ であった．

5.3 実験結果

本実験ではオペコード及びレジスタそれぞれを単独で機械学習が処理する入力データとして用いた場合，両方の場合，そして最後に特徴選択を用いた場合の，Accuracy と F-score の両方を比較する．

実験結果を見ると，オペコード及びレジスタのそれぞれで高い精度及び F-score が得られた．k-NN と SVM 分類器で 90%以上の高いパフォーマンスを発揮した．オペコードとレジスタを組み合わせることでより良い結果となることが期待されたが，実験結果より，Nearest Centriod 分類器以外に及ばなかった．つまり，非選択的なマルウェアの特徴を取り入れるだけでは，望ましい結果が得られない可能性があることが分かった．そのため，低レベルの特徴の組み合わせから特徴を選択することで，機械学習

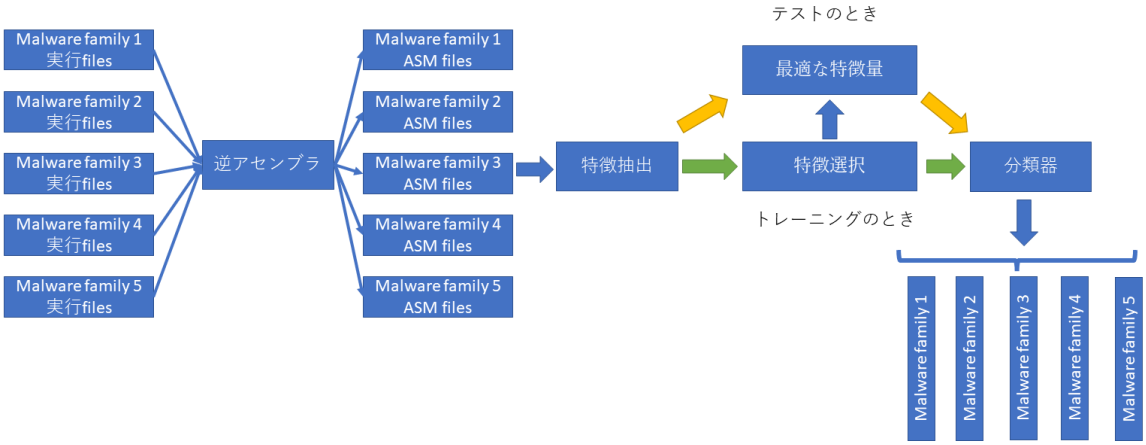

図 1　提案手法

Fig. 1　Overview of proposed method

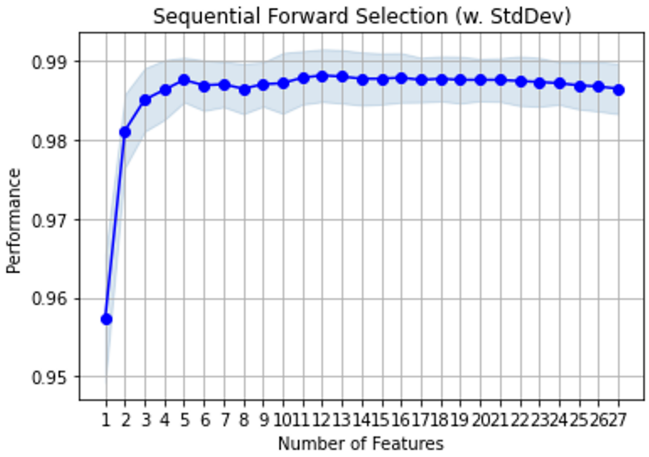

図 2　変数増加手法

Fig. 2　Sequential Forward Selection

がより良い結果で分類できるようになる．本手法では，各分類器の中，k-NN が一番良い結果を得た．Accuracy は特徴選択なしより 0.11%改善され， F-score は 0.67%改善された．Accuracy と F-score の最大値はそれぞれ 98.88%と 96.88%である．

6.　おわりに

本研究では，低レベルの特徴を用いたマルウェアの分類を行った．オペコードとレジスタを単純に組み合わせることで特徴量を増やしても，精度が上がるとは限らないことが分かった．本研究では，逐次特徴選択アルゴリズムを適用することで，より良い特徴量の組み合わせが得られた結果，分類性能が改善された．実験では，98.88% の Accuracy 及び 96.68%の F-score が得られ，高レベルの特徴を持った既存研究と匹敵する結果になる．さらに，逐次特徴選択アルゴリズムより得られた特徴の最適な組み合わせを考慮することで，マルウェアに関する高度な分析のヒントとして役に立つ可能性がある．

本研究ではシンプルな手法で高い分類精度を得た．しかし，本手法を使うためには大量なデータが必要であることと，標的型攻撃またはゼロデイマルウェアで作成された悪意のあるコードは，通常のデバッグツールでは正しく逆アセンブルすることができないこともある．マルウェアの作者が，コードをより複雑に改造したり，様々なエンコーディングなどを利用することで，データにノイズが入ってしまうと，統計的な手法に大きく影響を与える．今後は，ノイズを防ぐために他の手法との連携などが考えられる．

表 3　各分類器によるマルウェア分類の精度

Table 3　Accuracy of each malware classifier

アルゴリズム	精度 (%)			
	オペコード	レジスタ	オペコードとレジスタ	オペコードとレジスタ (特徴選択)
k-NN	98.77	98.21	98.71	**98.88**
Nearest Centroid	80.55	70.88	81.16	**90.78**
SVM	98.55	95.19	98.38	**98.60**

表 4　各分類器によるマルウェア分類の F-score

Table 4　F-score of each malware classifier

アルゴリズム	F-score(%)			
	オペコード	レジスタ	オペコードとレジスタ	オペコード と レジスタ (特徴選択)
k-NN	96.10	94.54	96.01	**96.68**
Nearest Centroid	66.43	45.44	67.01	**67.85**
SVM	94.86	84.61	94.40	**95.30**

参考文献

[1] Kaspersky Lab, “Digital life deserves complete protection: Kaspersky announces early access to new and reimagined consumer product portfolio”, Press release, Aug. 04, 2022, https://www.kaspersky.com/about/press-releases/2022_digital-life-deserves-complete-protection-kaspersky-announces-early-access-to-new-and-reimagined-consumer-product-portfolio

[2] SonicWall, “2021 Cyber Threat Report ”, https://www.sonicwall.com/resources/white-papers/2021-sonicwall-cyber-threat-report

[3] https://github.com/screetsec/TheFatRat

[4] https://github.com/ToR-0/Arbitrium-RAT

[5] D. Gibert, C. Mateu, J. Planes and R. Vicens, “Classification of Malware by Using Structural Entropy on Convolutional Neural Networks”, The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18), Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32, No.1, pp.7759–7764, 2018.

[6] J. Singh and J. Singh, “Challenges of Malware Analysis: Obfuscation Techniques”, International Journal of Information Security Science, Vol.7, No. 3, pp. 100–110, 2019.

[7] P. N. Yenboah, S. K. Amuquandoh, and H. Balle, “Malware Detection Using Ensemble N-gram Opcode Sequences”, International Journal of Interactive Mobile Technologies (iJIM), Vol. 15, No. 24, pp. 19–31, 2021.

[8] M. Shiasi, A. Sami, and Z. Salehi, “Dynamic malware detection using registers values set analysis”, 9th International ISC Conference on Information Security and Cryptology, pp. 54–59, Sep. 2012.

[9] F. Leder, B. Steinbock, and P. Martini, “Classification and Detection of Metamorphic Malware using Value Set Analysis”, 4th International Conference on Malicious and Unwanted Software (MALWARE), pp. 39–46, Otc. 2009.

[10] F. Li, C. Yan, and Z. Zhu, “A Deep Malware Detection Method Based on General-Purpose registers Features”, 19th International Conference, Faro, Portugal, pp. 221-235, Jun. 2019.

[11] Anh Pham Tuan, An Tran Hung Phuong, Nguyen Vu Thanh, Toan Nguyen Van, “Malware Detection PE-Based Analysis Using Deep Learning Algorithm Dataset”, figshare, 2018, https://doi.org/10.6084/m9.figshare.6635642.v1.

[12] B. B. Rad, M. Masrom, S. Ibrahim, and S. Ibrahim, “Morphed Virus Family Classification Based on opcodes Statistical Feature Using Decision Tree”, International Conference on Informatics Engineering & Information Science, pp. 123-131, Nov. 2011.

[13] 中村 嘉彦，北坂 孝幸，水野 慎士，古川 和宏，後藤 秀実，藤原 道隆，三澤 一成，伊藤 雅昭，縄野 繁，森 健策，“特徴選択による 3 次元腹部 X 線 CT 像からのリンパ節自動検出手法の精度向上”，第 32 回日本医用画像工学会大会予稿集，pp. 1–14, 2013.