

表層情報を用いたマルウェア検知精度を維持する 機械学習モデル更新手法の検討

栗原史弥¹ 松木隆宏¹ 寺田真敏¹

概要: 機械学習モデルを利用したマルウェア検知手法では、コンセプトドリフトと呼ばれる、機械学習モデルをトレーニングした時と予測する時のデータのずれが発生した場合に精度が落ちてしまうことが知られている。精度維持のため、ずれの発生したデータにラベルを付与し、再学習するというアプローチもあるが、学習データの作成に多大なコストがかかってしまうという課題がある。本稿では、この課題を解決するために、Fuzzy Hash 値を用いた機械学習モデル更新手法を提案する。提案方式は、マルウェア検知で高い精度を示した表層情報を利用することで、コストを抑えた機械学習モデル更新を実現すると共に、FFRI Dataset を用いた検証を通してその有効性を示す。

Feasibility study of Updating Machine Learning Models to Maintain Malware Detection Accuracy Using Surface Information

FUMIYA KURIHARA^{†1} TAKAHIRO MATSUKI^{†1}
MASATO TERADA^{†1}

1. はじめに

情報セキュリティ 10 大脅威 2016 年版[1]の個人 2 位、組織 7 位、2021 年版[2]の組織 1 位にはランサムウェアによる脅威がランクインしており、国内の個人や組織にとって、長きにわたってランサムウェアなどのマルウェアは大きな脅威となっている。

従来のマルウェア検知手法は、シグネチャによるパターンマッチング検知やヒューリスティック検知であったが、亜種のマルウェアを検知できなかったり、誤検知率が高かったりと検知手法としては必ずしも十分ではなかった。これらの課題に対処する新たな手法として、機械学習を利用した検知手法の研究が進められているが、機械学習にはコンセプトドリフトと呼ばれる、機械学習モデルのトレーニング時と予測時のデータのずれが発生した場合、精度が落ちてしまうという課題がある。Mtrends-2021[3]の報告によれば、2020 年にマルウェアの侵害を受けた環境で観測されたマルウェアファミリー 294 種類のうち 144 種類は 2020 年に新たに発見されたものであるとしている。新たなマルウェアの出現はトレーニング時と予測時のデータのずれ、すなわちコンセプトドリフトを引き起こし、マルウェア検知率低下につながる。検知率低下の対策として、機械学習モデルの再学習があるが、マルウェアかクリーンウェアかを詳細に調査した上でラベル付けする必要がある。対策として利用できるまでに時間がかかってしまう。具体的には、マルウェアを仮想環境で実行しプロセス、API 呼び出し、外部との通信などをツールによって追跡する動的解析や、

マルウェアを実行せず難読化を解除しソースコードを読む静的解析、ハッシュ値や PE ヘッダやセクションなどの情報を収集する表層解析から情報を得て、ラベルを付ける必要がある。

本稿では、動的解析、静的解析に比べてマルウェアの情報を収集しやすい表層解析から得られる PE ヘッダ、セクション、インポート関数、エクスポート関数、文字列情報、Fuzzy Hash 値などの表層情報、特に Fuzzy Hash 値に着目してマルウェアを検知する。さらに、コンセプトドリフトによる検知率の低下を防ぐため、表層情報として得られた複数の Fuzzy Hash 値のそれぞれの類似度に基づくコストを抑えた機械学習モデル更新手法について述べる。

2. 関連研究

茂木[4]は、FFRI Dataset 2018[5]を用いて Fuzzy Hash 値を除いた表層情報や Fuzzy Hash 値を除いた表層情報と Fuzzy Hash 値の組合せがマルウェア検知に有効で、特に Fuzzy Hash 値は誤検知率の低下に有効であることを示した。しかし、マルウェア検知率の推移に関して時間軸の視点からは検討しておらず、Fuzzy Hash 値の利用をマルウェア検知にのみ限定している。また、マルウェアに関するデータセットの分布を見ると、マルウェアは一部の大きなクラスに密集し、クリーンウェアは多数の小さなクラスに散らばる傾向がある。愛甲[6]は、データセットにこのようなスケールフリー性が内在することから、それらの時系列変化に注目しコンセプトドリフトによる変化に対応し、検知率を

¹ 東京電機大学
Tokyo Denki University

一定に保てることを示したが、マルウェア検知とマルウェア検知率維持の実装を Fuzzy Hash 値の 1 つである ssdeep の検証に留まっている。

本稿では、関連研究を踏まえ、機械学習での Fuzzy Hash 値のさらなる活用を考え、類似性判定そのものについて Fuzzy Hash 値の特徴を活かした 2 つのアプローチについて検討すると共に、Fuzzy Hash として知られている ssdeep, tlsh, impfuzzy による Fuzzy Hash 値を用いたマルウェア検知率の時間的な推移の検証、Fuzzy Hash 値を機械学習モデルの更新に利用することを目的とした。

3. 提案手法

本章では、表層情報を用いたマルウェア検知率の低下を防ぐ機械学習モデル更新手法について述べる。提案手法は、(1) 表層情報として Fuzzy Hash として知られている ssdeep, tlsh, impfuzzy による Fuzzy Hash 値を用い、(2) 類似性判定そのものについて Fuzzy Hash 値の特徴を活かした 2 つのアプローチを使って、(3) マルウェアは一部の大きなクラスに密集し、クリーンウェアは多数の小さなクラスに散らばる傾向があるという Fuzzy Hash 値に見られるスケールフリー性を利用して機械学習モデルを更新する。

以降、類似性判定そのものについて Fuzzy Hash 値の特徴を活かした 2 つのアプローチの観点から、機械学習モデルの更新手法を説明する。

3.1 共通部分文字列による機械学習モデル更新手法

Fuzzy Hash 値から抽出した共通部分文字列は、各データに含まれる類似した部分的特徴を表し、多くのデータに同じ部分的特徴が含まれているならば、それらは大きなクラスを形成することになる。共通部分文字列による機械学習モデル更新手法(以降、共通部分文字列による手法)は、これらクラスを前述のスケールフリー性の傾向からマルウェアと判定すると共に、その判定結果を基に毎月機械学習モデル更新するアプローチである。

共通部分文字列による手法の流れは、次の通りである(図 1)。なお、ここでは FFRI Dataset を収集した検体群として説明する。

- (1) 検体群(FFRI Dataset)から収集月を基に 1 ヶ月分の Fuzzy Hash 値を収集する。
- (2) 項番(1)を Suffix Tree に追加し、Fuzzy Hash の種類によって特定の文字数で共通部分文字列を抽出する。
- (3) 項番(2)の中で項番(1)の 3 つ以上のデータに含まれる共通部分文字列をマルウェアの特徴とする。項番(1)～(3)で抽出した共通部分文字列、すなわち、部分的特徴が存在することと前述のスケールフリー性の傾向からマルウェアであると判定する。
- (4) 項番(3)の特徴を持つマルウェアと判定した項番(1)の

データをマルウェアの Fuzzy Hash 値とする。

- (5) 検体群(FFRI Dataset)から FEXRD[7]によって抽出した表層情報(Fuzzy Hash 値を除く PE ヘッダ、セクション、インポート関数、エクスポート関数、文字列情報などのベクトル列)と項番(3)を結び付け、それらにマルウェアラベルを付与する。マルウェアラベルについては、機械学習モデルを作成、更新するためクリーンウェアなら 0、マルウェアなら 1 を付与する。
- (6) 項番(5)を機械学習モデルに学習させ更新する。その後、項番(1)～(6)を毎月分繰り返す。

なお、本手法では、先行研究を参考に項番(2)で抽出する文字数の既定値を、ssdeep は 30 文字、tlsh は 20 文字、impfuzzy は 30 文字とした。また、項番(2)(3)の Suffix Tree を用いた共通部分文字列の抽出例として、next, knee, done の 3 単語に対して 3 単語に含まれる 2 文字の共通部分文字列“ne”を抽出する流れを図 2 に示す。

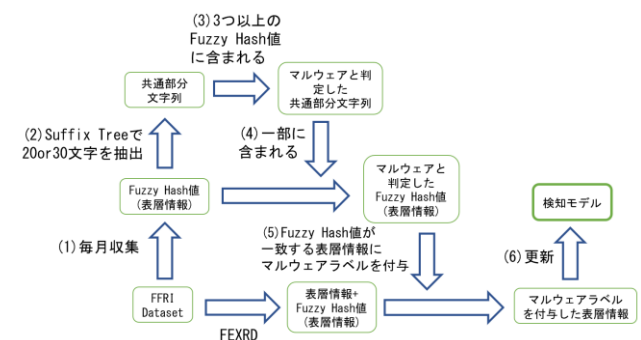


図 1 共通部分文字列による手法の流れ

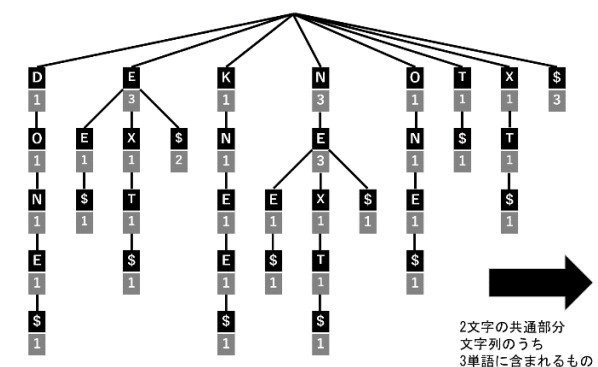


図 2 Suffix Tree で next, knee, done に対して 3 単語に含まれる 2 文字の共通部分文字列の抽出例

3.2 レーベンシュタイン距離による機械学習モデル更新手法

レーベンシュタイン距離は、2 つの文字列がどの程度異なっているかを示し、文字の挿入・削除・置換によって、一方の文字列をもう一方の文字列に変形するのにかかる最小の回数である。共通部分文字列による手法では、Fuzzy Hash 値の部分的な特徴の一致でマルウェアの判定をして

いたが、レーベンシュタイン距離による機械学習モデル更新手法(以降、レーベンシュタイン距離による手法)では、レーベンシュタイン距離から Fuzzy Hash 値の類似度を算出し、Fuzzy Hash 値の全体的な類似度と前述のスケールフリー性の傾向からマルウェアと判定すると共に、その判定結果を基に毎月機械学習モデル更新するアプローチである。

レーベンシュタイン距離による手法)の流れは次の通りである(図 3)。

- (1) 検体群(FFRI Dataset)から収集月を基に 1 ヶ月分の Fuzzy Hash 値を収集する。
- (2) 項番(1)から 2 つ取り出し、レーベンシュタイン距離を求める。
- (3) 項番(2)を式 1 に当てはめ類似度を算出する。
- (4) 項番(2)(3)をすべての組み合わせで繰り返す。
- (5) 項番(3)で類似度が 80 を超えた組み合わせを、項番(1)～(4)で算出した類似度の高い Fuzzy Hash 値の組み合わせの存在と前述のスケールフリー性の傾向からマルウェアであると判定する。また、マルウェアと判定した項番(1)のデータをマルウェアの Fuzzy Hash 値とする。
- (6) 検体群(FFRI Dataset)から FEXRD によって抽出した表層情報(Fuzzy Hash 値を除く PE ヘッダ、セクション、インポート関数、エクスポート関数、文字列情報などのベクトル列)と項番(5)を結び付け、それらにマルウェアラベルを付与する。
- (7) 項番(6)を機械学習モデルに学習させ更新する。その後、項番(1)～(7)を毎月分繰り返す。

$$\text{類似度} = 100 - \left(\frac{\text{レーベンシュタイン距離}}{\text{長い方の文字列長}} \times 100 \right)$$

式(1)

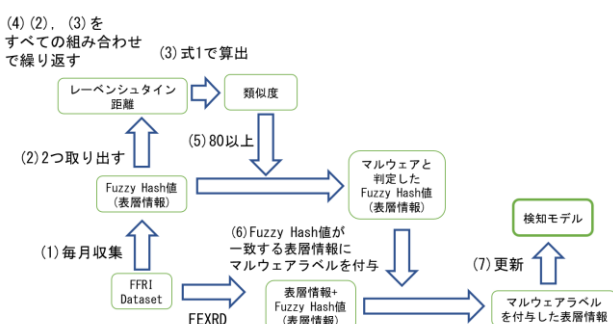


図 3 レーベンシュタイン距離による手法の流れ

4. 評価実験

本章では、提案手法である共通部分文字列とレーベンシュタイン距離による手法において、ssdeep, tlsh, impfuzzy による Fuzzy Hash 値をそれぞれ単独で機械学習モデル更新に適用することで、Fuzzy Hash 値ごとの違いと提案手法の有効性を評価する実験について述べる。

4.1 評価実験の目的

(1) 評価実験 1：機械学習モデルの期間

表層情報を用いたマルウェア検知において、テストデータの期間を変化させコンセプトドリフトが発生した場合の精度劣化、機械学習モデルの期間の長さが与える影響について、マルウェア検知率とクリーンウェア誤検知率の推移によって評価する。

(2) 評価実験 2：共通部分文字列による手法

共通部分文字列による手法において、ssdeep, tlsh, impfuzzy による Fuzzy Hash 値をそれぞれ単独で機械学習モデル更新に適用することで、Fuzzy Hash 値ごとの違いと提案手法の有効性をマルウェア検知率の維持、クリーンウェア誤検知率という視点から評価する。

さらに、抽出する共通部分文字列の文字数を変え、抽出する部分的特徴の長さが与える影響についても評価する。

(3) 評価実験 3：レーベンシュタイン距離による手法

レーベンシュタイン距離による手法において、ssdeep, tlsh, impfuzzy による Fuzzy Hash 値をそれぞれ単独で機械学習モデル更新に適用することで、Fuzzy Hash 値ごとの違いと提案手法の有効性をマルウェア検知率の維持、クリーンウェア誤検知率という視点から評価する。

4.2 利用するデータセット

機械学習モデルの作成に FFRI Dataset 2020[8]のクリーンウェア(良性データ) 7.5 万件、マルウェア(悪性データ) 7.5 万件、テストデータに FFRI Dataset 2021[9]の良性データ 7.5 万件、悪性データ 7.5 万件を利用した。なお、表層情報は FFRI Dataset 2018 から提供されているが、FFRI Dataset 2020 以降ではファイルサイズの大きなデータも解析対象になっているため、FFRI Dataset 2020 以降を利用することとした。

4.3 機械学習モデルの作成

FFRI Dataset 2020 に含まれるマルウェア及びクリーンウェアの表層情報を FEXRD によって抽出、機械学習モデルを作成した。分類には scikit-learn[10]を用いて Random Forest で実装し、パラメータは random_state=0、残りのパラメータはデフォルトとした。

4.4 評価実験の手順

(1) 評価実験 1：機械学習モデルの期間

機械学習モデルの期間を次の 3 つに分け、FFRI Dataset 2021 に含まれる表層情報を FEXRD によって抽出し、毎月のクリーンウェア誤検知率である Precision とマルウェア検知率である Recall の推移を確認する。

- 2019 年 1 月
- 2019 年上半期
- 2019 年全て

(2) 評価実験 2：共通部分文字列による手法

機械学習モデル(期間は 2019 年全て)を基に、FFRI Dataset 2021 に含まれる表層情報を FEXRD によって抽出し、共通部分文字列による手法(既定値：ssdeep は 30 文字、tlsh は 20 文字、impfuzzy は 30 文字)を、次の 4 つの条件下で、クリーンウェア誤検知率である Precision とマルウェア検知率である Recall の推移を確認する。

- 機械学習モデル更新なし
- ssdeep による機械学習モデルの毎月更新
- tlsh による機械学習モデルの毎月更新
- impfuzzy による機械学習モデルの毎月更新

さらに、評価実験 2 では、共通部分文字列の文字数を半分(短縮値：ssdeep は 15 文字、tlsh は 10 文字、impfuzzy は 15 文字)にした場合についても確認する。

(3) 評価実験 3：レーベンシュタイン距離による手法

機械学習モデル(期間は 2019 年全て)を基に、FFRI Dataset 2021 に含まれる表層情報を FEXRD によって抽出し、レーベンシュタイン距離による手法を、評価実験 2 の同一の条件下で、クリーンウェア誤検知率である Precision とマルウェア検知率である Recall の推移を確認する。

5. 評価結果

5.1 評価実験 1：機械学習モデルの期間

評価実験 1 の結果を図 4、図 5 に示す。

(1) クリーンウェア誤検知率(Precision)

Precision は、月ごとにある程度値の増減がある。また、機械学習モデルの期間を変えても同じような推移をしており、期間が短いほど値は低下し、月ごとの増減の幅も大きい。

(2) マルウェア検知率(Recall)

Recall は、Precision に比べて月ごとの値の増減幅が大きい。機械学習モデルの期間による値の差が大きい月もあれば小さい月もあり、期間が短いほど値は低下し、月ごとの増減の幅も大きい。

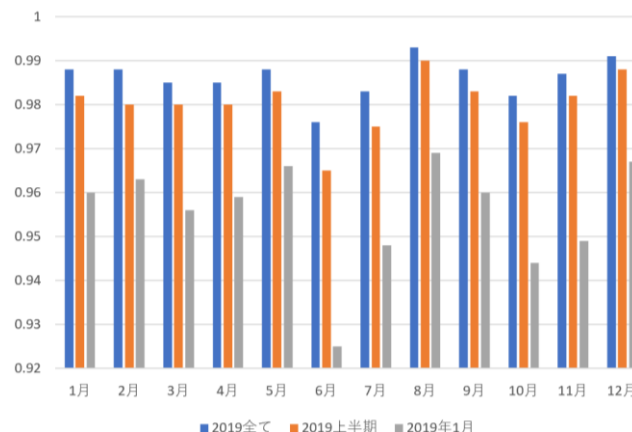


図 4 評価実験 1：クリーンウェア誤検知率の推移

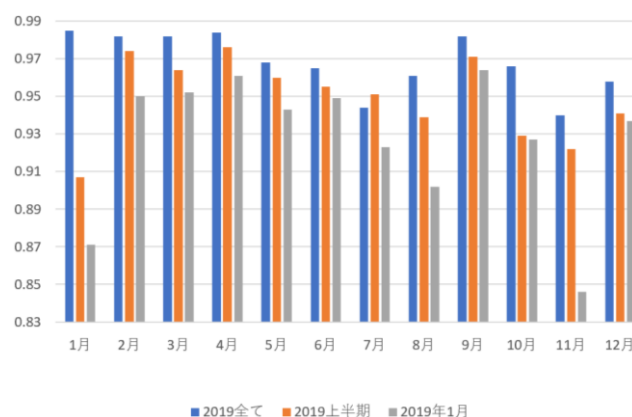


図 5 評価実験 1：マルウェア検知率の推移

5.2 評価実験 2：共通部分文字列による手法

5.2.1 共通部分文字列の文字数を既定値とした場合

評価結果実験 2 の結果を図 6、図 7 に示す。

(1) クリーンウェア誤検知率(Precision)

Precision は、機械学習モデルを更新しても機械学習モデル更新なしと比べて値にほぼ変化がなく高い値を維持している。また、Fuzzy Hash 値ごとの差もほぼない。

(2) マルウェア検知率(Recall)

Recall は、機械学習モデル更新なしの場合は 7 月、11 月は減少幅が大きいのにに対し、機械学習モデルを更新することで、Fuzzy Hash 値のいずれも高い値を維持している。Fuzzy Hash 値ごとに比較すると、impfuzzy は他の 2 つに比べ若干平均値が低く、機械学習モデル更新なしの減少幅が大きい 7 月、11 月は特に差が確認できる。

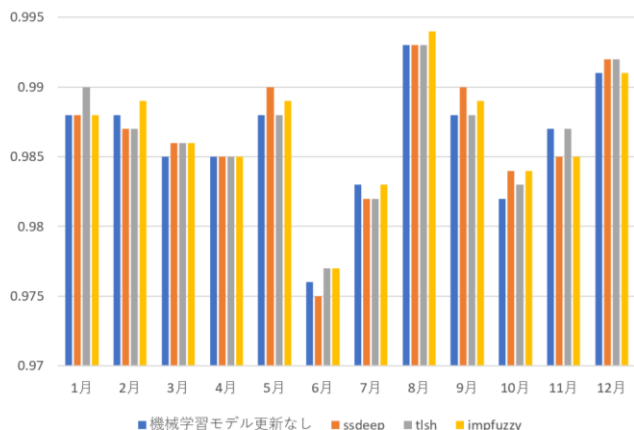


図 6 評価実験 2：クリーンウェア誤検知率の推移

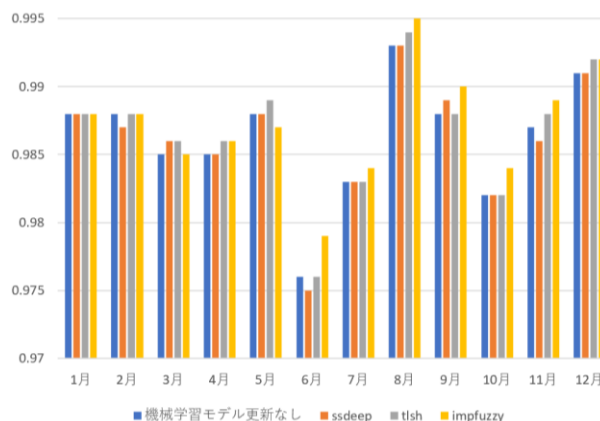


図 8 評価実験 2：クリーンウェア誤検知率の推移
共通部分文字列の文字数を半分にした場合

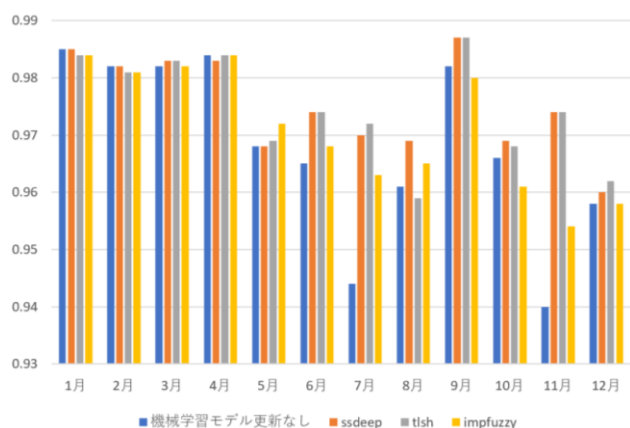


図 7 評価実験 2：マルウェア検知率の推移

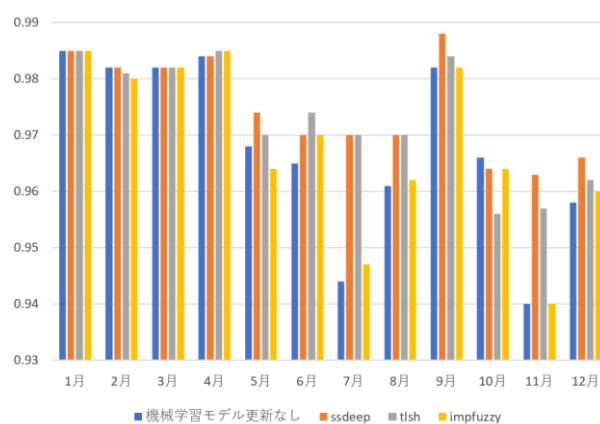


図 9 評価実験 2：マルウェア検知率の推移
共通部分文字列の文字数を半分にした場合

5.2.2 共通部分文字列の文字数を短縮値(既定値の半分)にした場合

評価実験 2 において、共通部分文字列の文字数を既定値の半分にした場合の結果を図 8、図 9 に示す。

(1) クリーンウェア誤検知率(Precision)

Precision は、既定値と同様に機械学習モデルを更新しても機械学習モデル更新なしと比べて値にほぼ変化がなく高い値を維持している。Fuzzy Hash 値ごとに比較すると 6 月以降 impfuzzy の値が他の 2 つに比べ若干高い。

(2) マルウェア検知率(Recall)

Recall は、減少幅が大きい 7 月、11 月も ssdeep と tlsh の 2 つは既定値と同様に高い値を維持している。Fuzzy Hash 値ごとに比較すると、impfuzzy は機械学習モデル更新なしの場合と値にほとんど差がない。また、既定値と比べて、11 月の値が軒並み下がり、全体的に impfuzzy の値が下がった。

5.3 評価実験 3：レーベンシュタイン距離による手法

評価実験 3 の結果を図 10 図 11 に示す。

(1) クリーンウェア誤検知率(Precision)

Precision は、評価実験 2 と同様に、機械学習モデルを更新しても機械学習モデル更新なしと比べて値にほぼ変化がなく高い値を維持している。Fuzzy Hash 値ごとの差もほぼない。

(2) マルウェア検知率(Recall)

Recall は、減少幅が大きい 7 月、11 月でも Fuzzy Hash 値のいずれも高い値を維持し、Fuzzy Hash 値ごとに比較すると、impfuzzy が他の 2 つに比べて高い値である。また、評価実験 2 に比べて、impfuzzy の値が高い値を維持している。

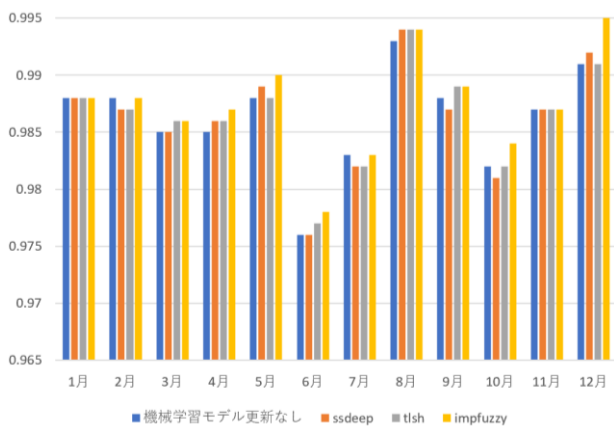


図 10 評価実験 3：クリーンウェア誤検知率の推移

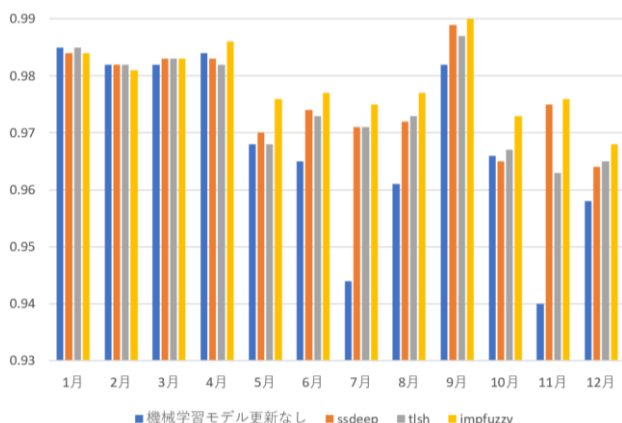


図 11 評価実験 3：マルウェア検知率の推移

5.4 考察

(1) コンセプトドリフトと学習データの期間が与える影響について

- クリーンウェア誤検知率の低下については、全ての評価実験において同様の变化で変化量も大きくないため、コンセプトドリフトの影響の可能性は低いと考えており、表層情報を基にした機械学習モデルのパラメーターチューニングなどによって改善の余地はあると考えている。
- コンセプトドリフトの影響によってマルウェア検知率が大きく落ちている月があり、1章で述べたマルウェアの変化が検知率の低下につながったと考えている。
- 学習データの期間はクリーンウェア誤検知率、マルウェア検知率共に長い方が精度は高く、ブレも少ないものの、期間が長くてもコンセプトドリフトの影響は受ける。

(2) 共通部分文字列による手法について

- 共通部分文字列による手法は、多少の差はあるもののクリーンウェア誤検知率にはほとんど影響を与えていない。コンセプトドリフトの影響によるマルウ

ェア検知率低下を抑え、高い水準で維持できているため部分的特徴を抽出する本手法はマルウェア検知率維持に有効であると考えられる。

- Fuzzy Hash 値ごとの比較では、`impfuzzy` の検知率が低く、共通部分文字列の長さを短くすると検知率が機械学習モデル更新なしとほぼ変わらない精度まで落ちる。これは、マルウェア・クリーンウェアの Import API から計算する `impfuzzy` の値に対し、部分的特徴を抽出する本手法ではマルウェアの特徴をうまく捉えられなかったことが原因と推測する。

(3) レーベンシュタイン距離による手法について

- レーベンシュタイン距離による手法は、共通部分文字列による手法と同様にクリーンウェア誤検知率にはほとんど影響を与えていない。マルウェア検知率についても同様に高い水準を維持しているため本手法もマルウェア検知率維持に有効であると考えている。
- Fuzzy Hash 値ごとの比較では、共通部分文字列による手法と違い、`impfuzzy` の値が 3 つの中では最も高く、これは、全体的な類似度を算出する当手法が `impfuzzy` の値からマルウェアの特徴をうまく捉えられたことによると推測する。

(4) 手法による差について

- 共通部分文字列による手法は部分的特徴を抽出する手法、レーベンシュタイン距離による手法は全体的な類似度を算出する手法であり、その差が `impfuzzy` のマルウェア検知率に表れていることがわかる。
- 10 月以降の後半にかけてレーベンシュタイン距離による手法の精度の方が若干高くなっていることから、手法による有効性の差を明確にするにはこの後の推移を見る必要がある。

(5) 提案手法の有効性と課題について

- 提案手法は、検体群にスケールフリー性があるという仮定の下で、Fuzzy Hash 値があればマルウェアラベル付与ができるため、再学習にあたって検体を詳細に解析する必要がなくラベル付けのコストなど課題解決に有効であると考えられる。
- ゼロデイ攻撃など、初期段階で未知のマルウェア検知をするには、提案手法で想定している機械学習モデルの更新間隔をより短くしなければならないと考えており、短期間で大量のマルウェアならびにクリーンウェアデータを収集する方法についても検討する必要がある。

6. まとめ

本稿では、機械学習での Fuzzy Hash 値のさらなる活用を考え、マルウェアは一部の大きなクラスタに密集し、クリーンウェアは多数の小さなクラスタに散らばる傾向があるという Fuzzy Hash 値に見られるスケールフリー性に着目した 2 つの類似性判定アプローチを使った機械学習モデルの更新手法を提案した。また、評価実験を通して提案手法がマルウェア検知率の維持に有効であることを示した。

今後は、複数の Fuzzy Hash 値を組み合わせることによる検知精度の改善、Fuzzy Hash 値以外の表層情報である PE ヘッド、セクション、インポート関数、エクスポート関数、文字列情報などの活用を検討すると共に、数年という長い期間の視点からマルウェアの変化と検知、動的解析や静的解析との組み合わせるによる適用範囲の課題にも取り組んでいきたいと考えている。

参考文献

- [1] 情報処理推進機構, “情報セキュリティ 10 大脅威 2016”, <https://www.ipa.go.jp/security/vuln/10threats2016.html>, 2021 年 12 月 30 日参照
- [2] 情報処理推進機構, “情報セキュリティ 10 大脅威 2021”, <https://www.ipa.go.jp/security/vuln/10threats2021.html>, 2021 年 12 月 30 日参照
- [3] FIREEYE, “M-TRENDS2021 最新のサイバー・トレンドと攻撃に関する知見”, <https://www.fireeye.jp/current-threats/annual-threat-report/mtrends.html>, 2021 年 12 月 30 日参照
- [4] 茂木裕貴, “PE 表層情報を用いた機械学習による静的マルウェア検知”, 暗号と情報セキュリティシンポジウム 2019, 2019
- [5] 研究用データセット MWS Datasets について, <https://www.iwsec.org/mws/datasets.html>, 2021 年 12 月 30 日参照
- [6] 愛甲健二, “Concept Drift と Scale-free の性質を持つデータセットに対する検知の自動化手法”, コンピュータセキュリティシンポジウム 2018, 2018
- [7] 株式会社 FFRI セキュリティ, “FEXRD”, <https://github.com/FFRI/FEXRD>, 2021 年 12 月 30 日参照
- [8] 株式会社 FFRI, “FFRI Dataset 2020 のご紹介”, https://www.iwsec.org/mws/2020/files/FFRI_Dataset_2020.pdf, 2021 年 12 月 30 日参照
- [9] 株式会社 FFRI セキュリティ, “FFRI Dataset 2021 のご紹介”, http://www.iwsec.org/mws/2021/files/FFRI_Dataset_2021.pdf, 2020 年 12 月 30 日参照
- [10] scikit-learn, <https://scikit-learn.org/stable/>, 2020 年 12 月 30 日参照