Vol.2015-DPS-162 No.49
Vol.2015-CSEC-68 No.49
2015/3/6

マハラノビス距離を用いた静的解析によるマルウェアの検出

岩本 舞[1]　小島 俊輔[2]　中嶋 卓雄[3]

概要：近年のマルウェアには多態性を持つものや，多数の亜種が存在するものがある．そのため，マルウェアの収集率が低下し，従来のシグネチャ型の検知が難しくなってきている．そこで我々の研究では，機械学習に着目した．本稿では，正常な EXE ファイルの統計量を用い，マハラノビス距離によってマルウェアを検出する手法を提案する．この手法では学習用のマルウェアが不要であり，未知のマルウェアの検出が期待できる．実験の結果，本手法の有用性が確認できた．

キーワード：マルウェア検出，マハラノビス距離，PE ファイルフォーマット，情報エントロピー，静的解析

Malware Detection based on Statistical Analysis using Mahalanobis-distance

MAI IWAMOTO[1]　SHUNSUKE OSHIMA[2]　TAKUO NAKASHIMA[3]

Abstract: Recently, there is many polymorphic and/or subspecific malwares. It makes difficult to collect malwares and detect them by the Signature based detection. So, we focused on machine learning. In this paper, we propose malware detection method using Mahalanobis-distance that calculated from some statistic values of benign executable files. This method does not require malwares for learning. Our method can be expected that detect unknown malwares. As a result of experiments, the method was effective.

Keywords: malware detection, Mahalanobis-distance, PE file format, information entropy, statistical analysis

1. はじめに

通常，マルウェアの検出は，既知のマルウェアの特徴的なコードを集めたデータベースを照会し，類似するコードを含む EXE ファイルをマルウェアと判定する，いわゆるシグネチャ方式で行われる．この手法では，あらかじめ EXE ファイルを収集・解析しておき，マルウェアを発見してデータベースに登録しておく必要がある．EXE ファイルの解析には，コードを実行して挙動を調査する動的解析と，コードを逆アセンブルし手動で挙動を調べる静的解析が用いられる．動的解析では，マルウェアの挙動を調べられるという利点があるが，コードを実行するためにコンピュータがマルウェアに感染する危険性がある．一方，静的解析では感染の危険がないが，逆アセンブルの結果からコードの挙動を調べることは困難である．

そこで，コードの挙動を調べることなくマルウェアを発見する手法が研究されている．たとえば，正常な EXE ファイルとマルウェアの情報エントロピーの違いを用いる手法 [1]，PE ヘッダ [2] の特徴を用いる手法 [3][4]，既知のマルウェアについて静的解析で得られる統計量を学習し，同じ統計量をもつ EXE ファイルをマルウェアとして検出する手法 [5][6] がある．機械学習の手法には，大きく分けて

[1] 熊本高等専門学校 技術・教育支援センター
Center for Technical and Educational Support, National Institute of Technology, Kumamoto College
[2] 熊本高等専門学校 ICT 活用学習支援センター
ICT Center for Learning Support, National Institute of Technology, Kumamoto College
[3] 東海大学 基盤工学部 電気電子情報工学科
Dep. of Electronics Engineering and Computer Science, Tokai University

教師あり学習と教師なし学習があり，教師あり学習では，教師となる既知のマルウェアが必要になる．しかし近年，マルウェアは多態性を持つうえ寿命が短く，さらに標的型といったハニーポットで収集できないタイプが存在するなど，教師となるマルウェアを収集することが困難な状況にある．そこで我々は，教師を必要としない静的解析手法に着目した．

本稿では，静的解析で得られた情報を確率変数とするマハラノビス距離 [7] を用いて，マルウェアを検出する．マハラノビス距離による検出には正常な EXE ファイルの分散共分散行列を用いるため，学習用のマルウェアが不要であるという特長がある．

マルウェアは，シグネチャ型の検知を回避するため，また収集された場合に挙動を解析されにくくするために難読化またはパックされている場合が多い．難読化またはパックされたコードは，通常のコードに比べ情報エントロピーが高くなる傾向にあるため，情報エントロピーはマルウェアの検出に有効である．しかしながら，難読化またはパックされていないマルウェア，難読化またはパックされた通常の EXE ファイルも存在することから，他にもマルウェア検出のための統計量が必要となる．そこでセクション数や実行書込可能セクションの有無といった他の統計量にも着目し，複数の統計量を組み合わせることによってマルウェアを検出することを試みた．

2 章では関連研究について述べる．4 章で提案手法を，5 章で実験方法，6 章で実験結果を述べ，本稿の手法が有用であることを示す．

2. 関連研究

文献 [1] では，難読化またはパックされたマルウェアを検出するために，EXE ファイルのセクションデータを固定長で区切ってブロックとし，各ブロック内での情報エントロピーを計算して平均値と最大値を算出している．実験結果では，難読化またはパックされた EXE ファイルと通常の EXE ファイルの情報エントロピーが完全に分離しており，難読化またはパックされたマルウェアの検出が可能である．文献 [6] では，ロジスティック回帰分析によってどのような統計情報がマルウェアの検出に有効であるかを調査している．また調査の結果，マルウェアの検出に有効だと判定された統計情報を判別分析に適用し，マルウェアの検出を行っている．文献 [5] では，実行ファイルの PE ヘッダから静的に得られるあらゆる情報を非線形 SVM で学習し，検出率を向上させた統計量について分析している．また，評価に 4 種類の異なる特徴を持つデータセットを用い，データセットによる検出性能の差についても検証している．

文献 [1] は難読化またはパックに着目した検出手法であり，難読化またはパックされていないマルウェアは検出できない．また文献 [6][5] は，いずれも教師あり学習を用いた検出手法であり，事前に収集したマルウェアによる学習が必要である．しかし，近年のマルウェアは先述の通り収集が難しい．そこで我々はマハラノビス距離による教師なし学習に着目し，研究を行った．

3. PE ファイルフォーマットの概要

PE ファイルフォーマット [2] とは，Microsoft Windows 環境でバイナリコードを実行するためのファイル形式である．PE ファイルフォーマットでは，ファイルはプログラムの情報を示すヘッダとバイナリコードに分けられる．**図 1** に概要を示す．なお，かっこ内はフィールドのサイズ (Bytes) を示す．MS-DOS 2.0 Section には，ファイルを MS-DOS で起動した場合に必要な情報やプログラムが格納される．COFF File Header には対応している CPU の種類やセクション数が格納される．Optional header には，PE ファイルが 32bit OS 用 (PE32) か，64bit OS 用 (PE32+) かを示すマジックナンバーや，ImageBase，セクションの最小単位が格納される．続く Section Header には各セクションのセクション名，バーチャルメモリやファイル上での位置およびサイズ，セクションの特徴を示すフラグ (実行可能，書込可能，初期化済等) が格納され，その後に各セクションの Raw Data が続く．

4. 提案手法

ここでは，実行可能領域とそれ以外の領域で情報エントロピーを算出する手法および複数の統計量を組み合わたマハラノビス距離によってマルウェアの検出を行う手法を提案する．

4.1 情報エントロピー

従来手法 [1] ではすべての Raw Data を一律に処理してい

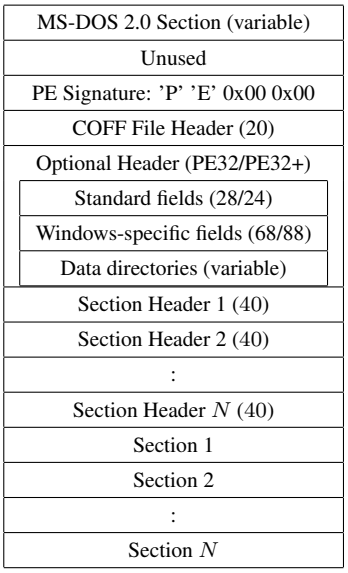


図 1 PE ファイルフォーマットの概要

るが，ここでは各セクションを実行可能なセクション (以下，**コードセクション**と記す) とそれ以外 (以下，**データセクション**と記す) に分類し，それぞれあらかじめ定めておいたサイズ (以下，**ブロック長**と記す) に区切り，各ブロックから 1Byte ずつ取り出したバイナリデータ (以下，**シンボル**と記す) を確率変数として，情報エントロピーを算出する．ここで，i 番目のブロックにおけるシンボル b_j(0x00-0xFF) の出現確率を $p_{i,j}$ とすると，i 番目のブロックの情報エントロピー h_i は以下の式で求められる．

$$h_i = -\sum_{j=0}^{255} p_{i,j} \log_2 p_{i,j} \tag{1}$$

ある EXE ファイルのコードセクションまたはデータセクションにおける h_i の集合を H とする．提案手法では，コードセクション・データセクションそれぞれの H の平均値または最大値を確率変数として使用する．ここで，平均値 H_{ave}，最大値 H_{max} は以下の式で表される．

$$H_{\text{ave}} = \text{ave}(H) \tag{2}$$

$$H_{\text{max}} = \max(H) \tag{3}$$

なお，ブロックがセクションをまたぐ場合や，最後に定められたサイズを満たさないブロックができた場合も，他のブロック同様 H_{ave} または H_{max} の算出に加える．

今回はホワイトリストの情報エントロピーについて平均及び標準偏差をとり，正規分布の 2σ(2.275%) をしきい値として使用した．

4.2 マハラノビス距離

本稿では，複数の統計量を確率変数とするマハラノビス距離によってマルウェアの検出を試みる．マハラノビス距離とは，2 つのベクトル間の統計学的な距離で，マハラノビス距離が近いほど 2 つのベクトルは類似している．正常な EXE ファイルの統計量 を 1 つのベクトルとし，もう一方を検査したい EXE ファイル のベクトルとすると，EXE ファイル が異常であるほど距離が長くなる．提案手法は，マハラノビス距離の違いによりマルウェアを検出する．

N 個 ($N \geq 1$) を要素とするベクトル W_i と X から，マハラノビス距離を算出する．ここで，W_i は正常な EXE ファイルリスト W における i 番目のファイルの統計量ベクトル，X はマルウェアであるかどうか判別したい EXE ファイルの統計量ベクトルである．Σ を W_i の分散共分散行列，$\overline{W}$ を W における各統計量の平均ベクトルとすると，X の W に対するマハラノビス距離 d は式 (4) で表される．

$$d = \sqrt{(X - \overline{W})^T \Sigma^{-1} (X - \overline{W})} \tag{4}$$

ここで，A^T は行列 A の転置行列，A^{-1} は A の逆行列である．

あらかじめ定めておいたしきい値 λ について $d > \lambda$ を満たすとき，X はマルウェアであると判定する．マハラノビス平方距離は自由度 N の χ^2 分布に従うことが知られている．今回は正規分布の 2σ に相当する確率 (2.275%) の χ^2 値をしきい値 λ として実験を行った．すなわち，しきい値 λ は以下の式で求められる．

$$\lambda = \sqrt{\chi^2(P(Z > 2), N)} \tag{5}$$

なお，今回の実験では，複数の統計量を用いてマハラノビス距離を算出した場合の有用性を確かめるため，すでに従来の研究で有効であるとされている以下の統計量 [8] を使用する．

セクション数　PE ヘッダに記載されたセクション数 (NumberOfSections) .

実行書込可能フラグ　実行可能かつ書き込み可能なセクションが存在すれば 1，存在しなければ 0 とする．

5. 実験方法

5.1 データセット

実験では，一般的な用途で使用している PC1(Windows7), PC2(Windows8) の計 2 台の PC から，Program Files(x86) フォルダ内の 10KBytes-500KBytes の EXE ファイルを抽出し，正常な EXE ファイルとみなして使用した．Program Files(x86) 内のファイルをに限定した理由は，現在のマルウェアの多くは 32bit・64bit OS 双方で利用可能な 32bit で作成されているためである．また EXE ファイルのサイズを 10KBytes-500KBytes と限定したのは，後述するマルウェア EXE ファイルのサイズが 10KBytes-500KBytes 程度だったためである．PC1 から抽出された 386 個 (以下，**ホワイトリスト**と記す) は，マハラノビス距離で用いる分散共分散行列 Σ の算出に用いた．また，PC2 から抽出された 166 個 (以下，**正常データセット**と記す) および 2011 年から 2014 年の間に収集された D3M データセット [9] に含まれる EXE ファイル 41 個 (以下，**マルウェアデータセット**と記す) を検証に用いた．

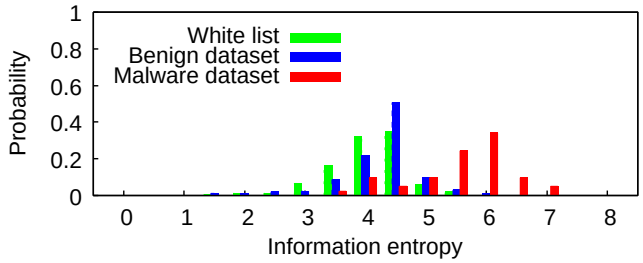
5.2 評価基準

今回の実験では，マルウェアであることを正しく検知した True-Positive(以下 TP)，マルウェアでないものをマルウェアと検知した False-Positive(以下 FP)，マルウェアを検知しなかった False-Negative(以下 FN) を基準とする F 値で検出手法を評価する．

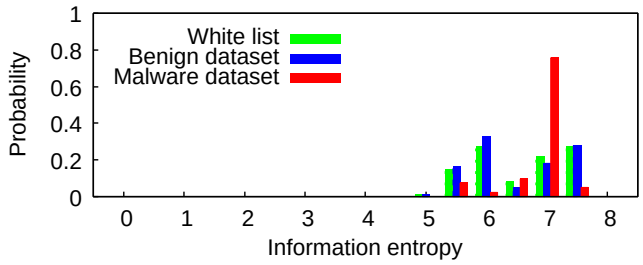
ここで，FP，FN を客観的に評価するための一般的な尺度として，再現率 Recall (以下 R)，適合率 Precision(以下 P)，F-measure(以下 F 値) を使用する．R，P，F 値は，それぞれ式 (6), (7), (8) のように定義される．

$$R = \frac{tp}{tp + fn} \tag{6}$$

$$P = \frac{tp}{tp + fp} \tag{7}$$



(a) 平均値



(b) 最大値

図 2 情報エントロピーの分布 (従来手法)

$$F\text{ 値} = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} \tag{8}$$

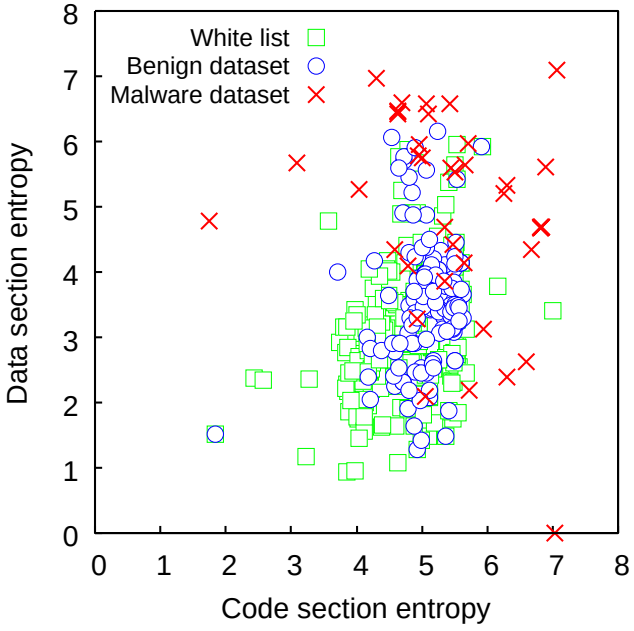
ここで，tp，fn，fp はそれぞれ TP，FN，FP の数である．R，P，F 値は 0 以上 1 以下の値をとり，1 に近いほど検出手法が正確であったことを意味する．そこで，本研究ではマルウェア検出の性能評価に F 値を用いる．

6. 実験結果

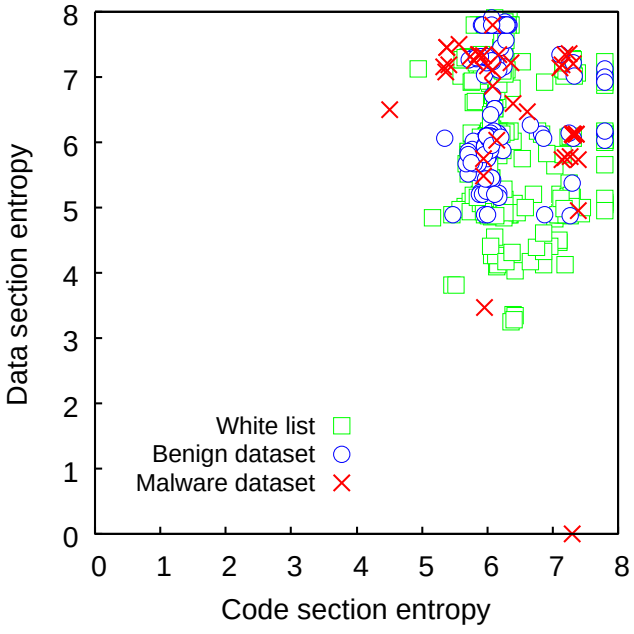
まず，提案するコードセクション・データセクションごとの情報エントロピーの有用性および複数の確率変数を用いたマハラノビス距離の有用性を示す．その後，2 つの手法をを組み合わせた結果を示すことで，提案手法の有用性を示す．

6.1 情報エントロピー

ここでは，従来手法 [1] と提案手法での情報エントロピーの分布の違いについて述べる．以下の実験ではブロック長を従来手法で提案されている 256Bytes とした．**図 2** は，従来手法で算出した情報エントロピーの分布図である．図では，情報エントロピーを 0.5 ごとに区分し集計した，各データセット内でのファイルの割合を示している．**図 2(a)** に，従来手法 (平均値) の分布を示す．マルウェアデータセットは全体的に情報エントロピーが高い傾向にあるものの，一部でホワイトリストおよび正常データセットと分布が重なる．また**図 2(b)** に，従来手法 (最大値) をの分布を示す．ホワイトリストとマルウェアデータセットの分布が重なっており，区別することができない．[1] では，情報エントロピーの最大値は通常の EXE ファイルと難読化またはパックマルウェアで区別できるとされていたが，現在は傾向が変わっており，区別できなくなっていることが



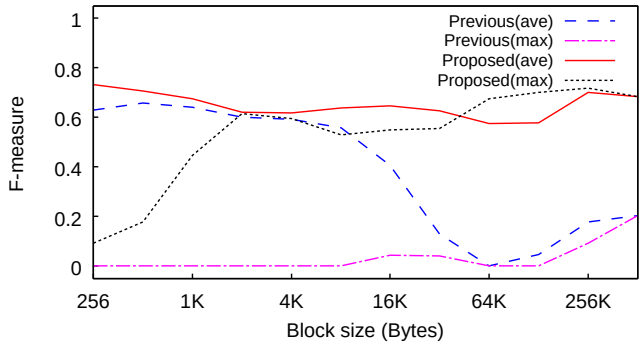
(a) 平均値



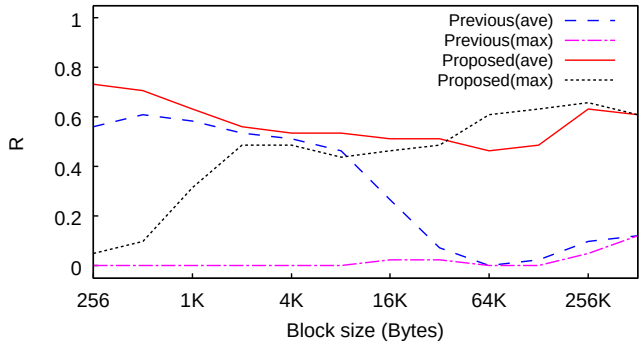
(b) 最大値

図 3 情報エントロピーの分布 (提案手法)

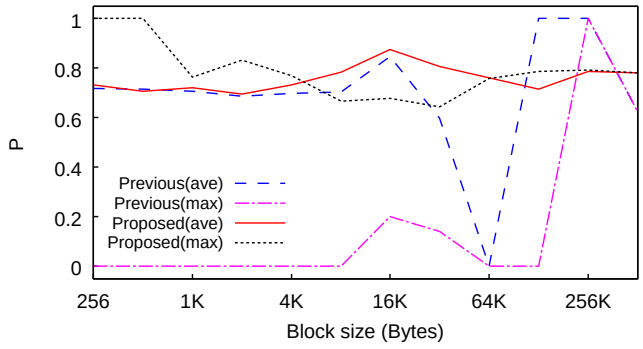
分かる．**図 3** は，提案手法で算出した情報エントロピーの分布図である．横軸をコードセクション，縦軸をデータセクションの情報エントロピーとし，ホワイトリスト，正常データセット，マルウェアデータセットの 3 種類をを 2 次元マップ上にプロットした．**図 3(a)** は，提案手法における平均値の分布である．多くのマルウェアがホワイトリストおよび正常データセットの分布の中心から外れており，コードセクション・データセクションごとの情報エントロピーが検出に有効であることが分かる．**図 3(b)** に，最大値の分布を示す．最大値を使用すると，コードセクション・データセクション双方の情報エントロピーが高い方向に偏



(a) F 値



(b) R

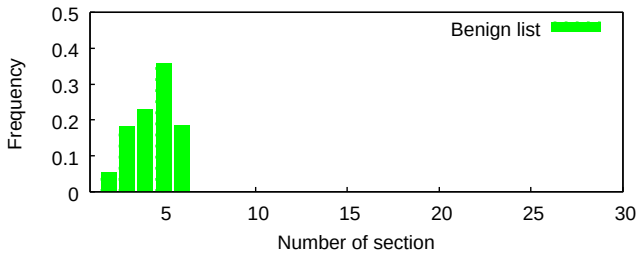


(c) P

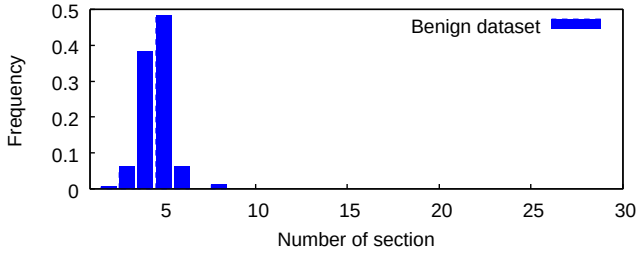
図 4 ブロック長による F 値，R，P の変化 (情報エントロピー)

り，各データセットが重なりあってしまった．

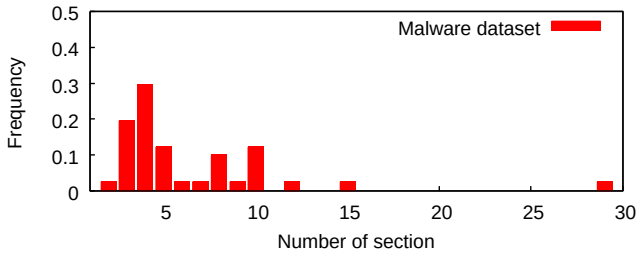
図 4 に，ブロック長をパラメタとする F 値，R，P の変化を示す．ここで横軸はブロック長，縦軸はそれぞれ F 値，R，P である．今回の実験では，ブロック長を 2^n Bytes とし，n を 8-19(256Bytes-512KBytes) まで変化させて実験を行った．なお，各データセットの EXE ファイルはすべて 500KBytes 以下のため，ブロック長 512KBytes の場合の情報エントロピーは全体の情報エントロピーと同値であり，平均値と最大値が同じ値になる．従来手法 (平均値) を確率変数とした場合，F 値は 512Bytes で最大の 0.65 となり，ブロック長を大きくすると低下する．従来手法 (最大値) を確率変数とした場合は R，P 双方が低く，有用性が見られなかった．提案手法 (平均値) を確率変数とした場合，F 値は 256Bytes で最大の 0.73 となる．ブロック長の変化はほとんど F 値に影響せず，0.6-0.75 の間に収まっているが，R はブロック長が小さいところで，P はブロック長が



(a) ホワイトリスト



(b) 正常データセット



(c) マルウェアデータセット

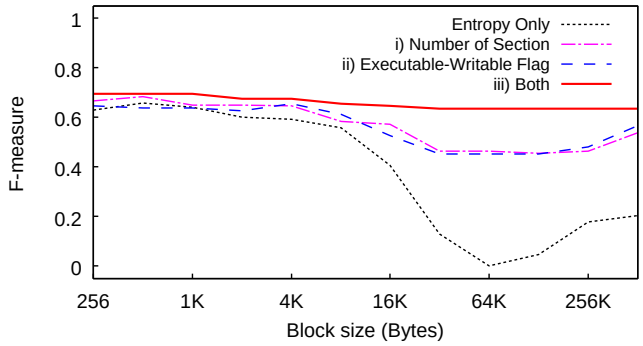
図 5 EXE ファイルに含まれるセクション数の分布

16KBytes 程度のところで大きくなっている．提案手法 (最大値) を確率変数とした場合，F 値は 2KBytes 程度から向上し始める．

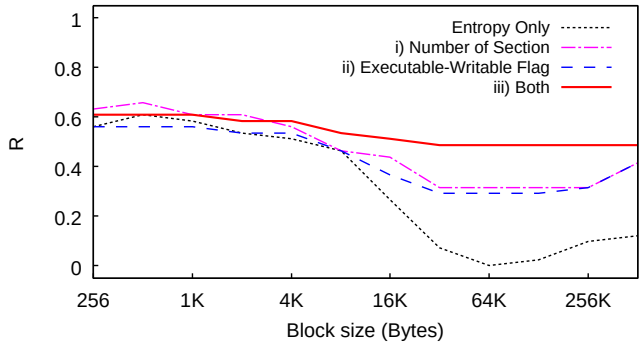
6.2 マハラノビス距離

ここではまず，セクション数および実行書込可能フラグについて，実験に用いたデータセットでも有効であることを示すため，正常データセットと マルウェアデータセットによる予備実験を行った．**図 5** は，データセットごとのセクション数の分布である．横軸はセクション数，縦軸は各データセット内でその数のセクションを持つファイルの割合である．ホワイトリストおよび正常データセットのセクション数は多くても 6 程度だが，マルウェアではセクション数が多いものが存在する．セクション数が 6 より多いファイルは，ホワイトリストおよび正常データセットでは 552 個中 2 個 (0.36%) だったのに対し，マルウェアデータセットでは 41 個中 14 個 (34.1%) であった．また実行書込可能フラグについて，実際にデータセットに含まれる EXE ファイルを調査したところ，実行書込可能なセクションを持つファイルの数は，ホワイトリストおよび正常データセットでは 552 個中 3 個 (0.54%) だったのに対し，マルウェアデータセットでは 41 個中 12 個 (29.2%) であった．

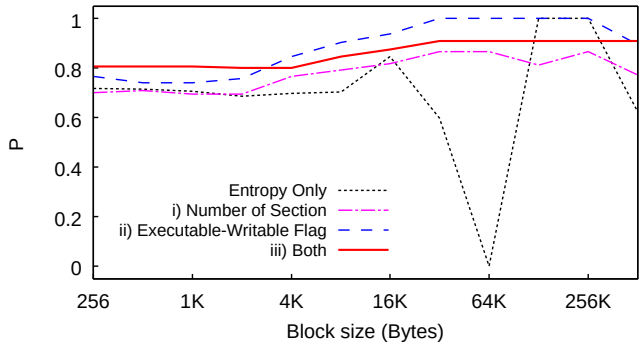
図 6 に，情報エントロピー (従来手法) に i) セクション



(a) F 値

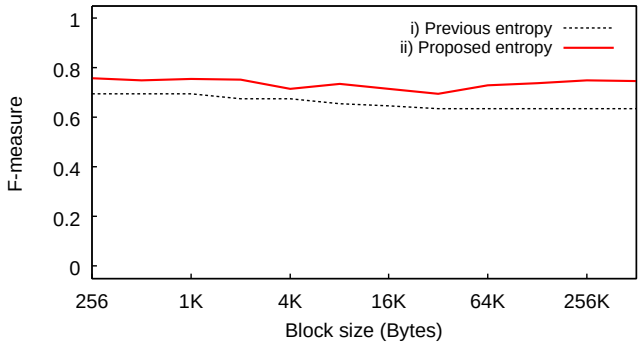


(b) R

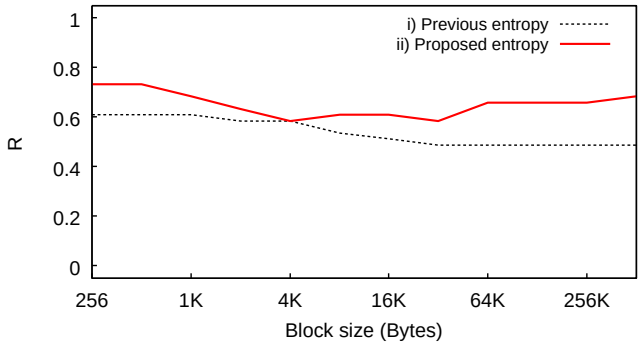


(c) P

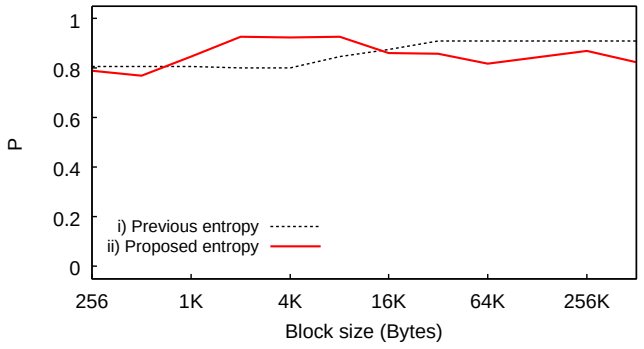
図 6 ブロック長による F 値，R，P の変化 (情報エントロピー (従来手法)，セクション数，実行書込可能フラグ，双方)



(a) F 値



(b) R



(c) P

図 7 ブロック長による F 値，R，P の変化 (情報エントロピー，セクション数，実行書込可能フラグ)

数，ii) 実行書込可能フラグ，iii) セクション数・実行書込可能フラグ双方を加えて確率変数とした場合の，ブロック長をパラメタとする F 値，R，P の変化を示す．いずれの場合も情報エントロピーだけを使用した場合に比べ F 値が向上していることが分かる．

また，情報エントロピーがあまり有効でない場合 (ブロック長 $>$ 4KBytes) でも，セクション数や実行書込可能フラグを組み合わせることで R が向上しており，さらに双方を確率変数とした場合には，情報エントロピーが有効な場合とほぼ同じ R を保っていることが分かる．これは，複数の統計量を確率変数とするマハラノビス距離を使用することにより，いずれかの確率変数が有効でなくなったとしても，安定した検出が可能であることを意味する．

6.3 提案手法の有用性

図 7 に，i) 従来手法の情報エントロピー，ii) 提案手法の情報エントロピーと，セクション数・実行書込可能フラグを確率変数とした場合のブロック長をパラメタとする F 値，R，P の変化を示す．セクション数・実行書込可能フラグを提案するコードセクション・データセクションごとの情報エントロピーと組み合わせた場合には，従来手法の情報エントロピーと組み合わせた場合に比べ，F 値が高くなっている．特に R が向上しており，より多くのマルウェアを発見できたことが分かる．よって，我々が提案するコードセクション・データセクションごとの情報エントロピーの平均値を用いる手法および複数の統計量を確率変数とするマハラノビス距離を用いる手法は，それぞれが有効なだけでなく，組み合わせた場合にはより高い検出率となることが確認できた．

7. まとめ

実験の結果，正常な EXE ファイルとマルウェアではコー

ドセクション・データセクションごとの情報エントロピーの分布が異なること，複数の統計量によるマハラノビス距離を用いることによって F 値が向上することが分かり，双方を組み合わせた手法では F 値は 0.76 となった．マハラノビス距離では，使用する確率変数を自由に設定できる．今回は情報エントロピー，セクション数，実行書込可能フラグの計 3 種類の確率変数を用いたが，検出できないマルウェアも存在した．他にもマルウェアの検出に有効な確率変数は存在するため，今後は他にどのような確率変数がマルウェアの検出に有効なのかを検証し，さらなる検出率の向上を目指す．また，現在のデータセットはホワイトリストおよび正常データセットが 552 個，マルウェアデータセットが 41 個と少ないため，今後はデータセットを増やして検証したい．

参考文献

[1] Lyda, R. and Hamrock, J.: Using Entropy Analysis to Find Encrypted and Packed Malware, *IEEE Security and Privacy*, Vol. 5, No. 2, pp. 40–45 (online), DOI: 10.1109/MSP.2007.48 (2007).
[2] Microsoft: Microsoft PE and COFF Specification, http://msdn.microsoft.com/library/windows/hardware/gg463119.aspx (2014).
[3] Liao, Y.: PE-Header-Based Malware Study and Detection, http://www.techrepublic.com/resource-library/whitepapers/pe-header-based-malware-study-and-detection/ (2012).
[4] Devi, D. and Nandi, S.: PE File Features in Detection of Packed Executables, *International Journal of Computer Theory and Engineering*, Vol. 4, No. 3, pp. 476–478 (online), DOI: 10.7763/IJCTE.2012.V4.512 (2012).
[5] 笹生憲, 村上純一, 松木隆宏 and 森達哉: 機械学習によるマルウェア検出　リローデッド, **コンピュータセキュリティシンポジウム** *2014* **論文集**, Vol. 2014, No. 2, pp. 827–834 (2014).
[6] 田中恭之, 有川隼 and 畑田充弘: 統計的手法を用いたマルウェア判定の実験結果, **コンピュータセキュリティシンポジウム** *2014* **論文集**, Vol. 2014, No. 2, pp. 821–826 (2014).
[7] Mahalanobis, P. C.: On the generalised distance in statistics, *Proceedings National Institute of Science, India*, Vol. 2, No. 1, pp. 49–55 (1936).
[8] Yonts, J.: Attributes of Malicious Files, http://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979 (2014).
[9] 秋山満昭, 神薗雅紀, 松木隆宏 and 畑田光弘: マルウェア対策のための研究用データセット 〜MWS Datasets 2014〜, **情報処理学会研究報告**. *SPT,* **セキュリティ心理学とトラスト**, Vol. 2014, No. 19, pp. 1–7 (online), available from ⟨http://ci.nii.ac.jp/naid/110009804706/⟩ (2014).