Computer Security Symposium 2014
22 - 24 October 2014

統計的手法を用いたマルウェア判定の実験結果

田中恭之[†1,†2,a)]　有川隼[†1]　畑田充弘[†1]

概要：マルウェアが爆発的に増加する中でシグネチャによらない軽量なマルウェア判定方法が望まれている．本稿では，マルウェア判定に有効と考えられるファイルの静的な情報から独立変数を定義し，統計的手法を用いてマルウェアらしさを判定する実験を行った結果及び考察を示す．統計手法としてはロジスティック回帰分析と判別分析を用いた．実験対象検体に対し，有効な独立変数を探り，フィットするモデルを構築することができた．また，統計的手法の有効性・課題を確認することができた．

キーワード：マルウェア，統計的手法，静的解析，MWS Datasets 2014，D3M

An Experimental Result of Malware Detection using Statistical Techniques

YASUYUKI TANAKA[†1,†2,a)]　JUN ARIKAWA[†1]　MITSUHIRO HATADA[†1]

Abstract: In malware increases explosively, malware judgment a lightweight way that does not depend on signature is desired. In this paper, we define the independent variable from the static information in the file that we are considered to be valid to the malware judgment. Next, we show the results and discussion conducted experiments to determine the malware by using a statistical method. We've used the discriminant analysis and logistic regression analysis as a statistical method. We could construct a good model to find a valid independent variables. We confirmed the effectiveness and issues statistical methods.

keywords: malware, statistical method, static analysis, MWS Dataset 2014, D3M

1. はじめに

従来のシグネチャをベースとしたマルウェア検知手法は限界と言われている．これは，現状のアンチウイルスソフトが，マルウェアの感染などを検知できるのは 45%程度でありそれ以外は検知できないといった報道[1]からも窺い知れる．我々は，攻撃手法が巧妙になる中で，多層防御の観点から,出口対策として有効な URL フィルタリングに適用できる URL ブラックリストを,サンドボックス解析結果から作成する検討を進めている[2]．入手した解析対象をすべてサンドボックス解析することは効率が悪く，マルウェアらしさが高い検体から優先的に解析したいニーズがある．また，マルウェア自体が爆発的に増加するなか，今後すべてをサンドボックス解析対象とすることが困難になることも予想される．

本稿では，マルウェア判定に有効と考えられるファイルの静的な情報から独立変数を定義し，統計的手法を用いてマルウェアらしさを判定する実験を行った結果及び考察を示す．統計手法としてはロジスティック回帰分析と判別分析を用いた．有効な独立変数を探り，フィットするモデルを構築することができた．このモデルを用いて，マルウェアらしさが高いものを優先的に解析対象とすることで解析効率を向上できる．また，一般的にもシグネチャベースではない静的かつ軽量な判定手法は，マルウェアが爆発的に増加する中，既存の動的解析や静的解析手法と組み合わせて用いることの有効性が期待できる．尚，統計的検定は無作為抽出を前提としているので，今回の実験結果は，実験対象検体内での結果であることに注意する必要がある．

†1 NTT コミュニケーションズ株式会社
NTT Communications Crop., Gran Park Tower 16F, 3-4-1, Shibaura, Minato-ku, Tokyo, 108-8118, Japan
†2 情報セキュリティ大学院大学
IISEC, Yokohama, Kanagawa 221–0835, Japan
a) yasuyuki.tanaka@ntt.com

2. 関連研究

文献[3]では，検査対象ファイルをいくつかの区間に分割し，区間毎の情報エントロピーを計算し，ファイル全体での統計から通常のファイルかマルウェアかを識別する．マルウェアの多くはパッキングされているため高い情報エントロピーを持つという性質を利用している．

文献[4]では，検査対象ファイルに含まれる可読な文字列を抽出し，教師あり学習を用いる識別手法である SVM（サポートベクターマシン）を適用することによって，通常のファイルかマルウェアかを判定する．文献[5]ではさらに，検査対象のファイルの先頭バイトに特化して可読文字を抽出し同様に SVM を適用し，[3]や[4]と比較を行っている．

文献[6][7]では，統計的手法である，ロジスティック回帰

分析を用い検査対象のファイルがマルウェアか否かの判定を試みている．

3. 予備実験

独自にハニーポットで取得し，取得時に 4 社のアンチウイルスソフトすべてでマルウェアと判定されなかった検体を 1000 個用意*し，12 時間毎のアンチウイルスソフトの判定状況を図 1 にグラフ化した. 24 時間までは検知率が著しく向上するがそれ以降は非常に緩やかになっていることがわかる．84 時間後の時点では 1000 個中 186 個が少なくとも 1 社のアンチウイルスソフトでマルウェア判定された．また[1]の報道から推測すると，実際にはもっと多くの割合でマルウェアが存在すると考えられる．

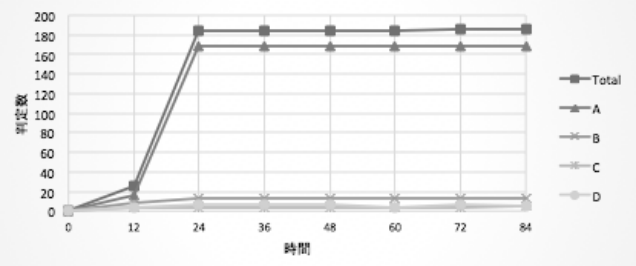

図 1 アンチウイルスソフト判定推移

4. 判定方式

4.1 方針

関連研究から対象ファイルの情報エントロピーの量は，それ単品でも，マルウェアか否かを判定する上でかなり良い指標であること，また，統計手法（ロジスティック回帰分析）もマルウェア判定に有効性があることがわかる．統計手法を用いる上で，有効な独立変数の組み合わせが重要になるため，本稿では，情報エントロピーの量に加えて，文献[6][7]では用いられていないものを含めた有効と考えられるいくつかの独立変数を定義して用い，有効性を評価した結果を示す．また，手法としてロジスティック回帰分析に加えて判別分析を実施した結果も示す.

4.2 統計手法

ロジスティック回帰分析は，独立変数を複数個とる回帰分析である重回帰分析の一種であり，重回帰分析が従属変数として量的変数を取ることと比較して，従属変数として質的変数を取ることができる特徴がある．マルウェアか否かを判定する場合 2 値の質的変数を出力ととればよい．重回帰分析を用いる場合，適切な独立変数を適切な個数投入しないと，多重共線性等が発生し正しい分析結果が出ないことから，モデルがどの程度フィットしたかが重要になるので，この点の結果についても触れる．

判別分析では，同様にマルウェアか否かを判定できるが，ロジスティック回帰分析で得られるようなそれぞれの独立変数の影響度を示すオッズ値等の情報は得られない．

4.3 独立変数の定義

マルウェアか否かの 2 値を従属変数とし，独立変数の候補としてファイルから静的に得られる情報を[8][9]を参考に以下に定義した．

- 変数：Date

一部のマルウェアはタイムスタンプ（PE ファイルヘッダの IMAGE_FILE_HEADER.TimeDateStamp の値）を意図的に過去や未来に改変しているものがある．ここでは，1999 年以前, 2015 年以降をマルウェアの可能性が高いとして Date=1 とし，それ以外を Date=0 とした．

- 変数：Packer

Packer により Pack されているファイルの方がマルウェアである確率が高いと考え，PEiD†により Packer が判別できたものを Packer=1，それ以外を Packer=0 とした．

- 変数：UPX

マルウェアの場合, Packer のなかでも特に UPX を用いられることが多いという推測から，PEiD で Packer が特定できなくても UNIX の file コマンドの結果で UPX でパックされていると推測可能な場合, UPX=1 とし, それ以外を UPX=0 とした．尚，本実験では変数 : Packer から値 UPX は除外せずに両変数の相関を確認することにした．

- 変数：EP

疑わしいエントリーポイントを示す. PE オプショナルヘッダの IMAGE_OPTIONAL_HEADER.AddressOfEntryPoint に示されるエントリーポイントを含むセクション名は通常，”.text”等の名称で，実行されるコードのバイナリデータが記録されていると定義されるコードセクションである．この名称は任意に付与でき，名称によりセクション属性が決まるわけではないが，一般的に用いられる”.text”,”.code” 等ではない場合疑わしいとして EP=1 とし，それ以外を EP=0 とした．

- 変数：TLS

TLS (thread local storage) エントリがある場合 TLS=1，それ以外を TLS=0 とする．TLS callback 手法を用いると，TLS スレッド毎に固有な記憶領域を利用し，エントリーポイントより前で任意のコードを実行させることができる．悪意のあるソフトウェアで用いられる確率が高いと推測し，このエントリの存在有無を変数として追加することにした．

- 変数：Resource

Resource エントリがある場合 Resource=1，それ以外を Resource=0 とする．Resource エントリにはプログラムで用いられるリソースに関連する情報が記述されており，悪意の無いソフトウェアでも用いられるが，多くのマルウェアで用いられることが知られており，他の変数と組み合わせると有効である可能性が高いと推測したため定義した．

- 変数：IAT

---

* 判定日：2014/08/14

† http://woodmann.com/BobSoft/Pages/Programs/PEiD

IAT（import address table）に登録された関数のうち，マルウェアが良く用いると考えられる関数の個数を，量的変数 IAT として定義する．該当関数は予め登録しておきヒットした関数の数を IAT の値として用いる．ファイルがパックされている場合，パック前の IAT はパック後の IAT と異なることから一般に解析価値は薄いが，ファイルを観察していると，パック後の関数に関しても特徴がみられる可能性が高いと考えられたため量的変数として定義した．

- 変数：Section

異常と定義したセクションの個数を，量的変数 Section として定義する．異常セクションが無い場合を Section=0 とする．セクション異常の定義として，PE セクションテーブルの IMAGE_SECTION_HEADER.SizeOfRawData の値がゼロである，もしくは，各セクションのエントロピーが極端に高いか低い場合とした．情報理論では対象データのエントロピー値は 0～8 を取り，8 に近いほどランダムであることを示す．マルウェアはパックによる暗号化や難読化がされている確率が高いと考えられることからこの指標を定義した．

- 変数：VerInfo

PE ファイルのバージョン情報には，作成者や所属会社の情報，コピーライト，ファイルバージョン，内部名称等の様々な付加情報が記載されている．ここでは，内容の精査はせずに，バージョン情報が存在しない場合，マルウェアの可能性が高いと推測し，質的変数 VerInfo=1 とし，それ以外を VerInfo=0 とした．

- 変数：FileEntropy

ファイル全体のエントロピー値を量的変数 FileEntropy として定義した．変数 Section ではセクション毎に判定した 2 値の累計としているが，FileEntropy では 0～8 までの量的な値とする．

5. 実験

5.1 対象検体

- マルウェア検体

MWS 2014 Datasets(D3M 2010,2011,2012,2013,2014)に含まれるマルウェア[10]及び，ハニーポットで取得しアンチウイルスソフトでマルウェア判定された検体‡．合計 433 個．なお 3 章の予備実験で用いた検体とは環境の都合上異なるものである．

- 正常検体

Windows 端末上に存在した Microsoft 社提供の実行ファイル及びフリーソフトとしてダウンロードした一般のアプリケーションの実行ファイル．合計 339 個．

5.2 従属変数-独立変数間の相関係数

従属変数を変数 Mal とし，マルウェア検体の場合 Mal=1，正常検体の場合 Mal=0 とする．従属変数と各独立変数の相関係数を表 1 に示す．尚，本稿では値は R§を用いて算出している．変数 EP，VerInfo に弱い相関が見られ（>0.2），変数 FileEntropy にやや強い相関(>0.4)が見られた．TLS について負の相関となっていて，変数を定義した想定と逆になっていた．そこでデータを確認すると，マルウェア検体で，TLS=1 となるものはほとんどなく，正常検体で TLS=1 検体が多数存在した．TLS=1 の該当アプリケーションを確認したところ，C/C++コンパイラの各種ソフトウェアに含まれる実行ファイルやデバッガが該当し一般のアプリケーションとは用途が異なるものであった．また，他の変数についてはほとんど相関が見られなかった．

表 1 従属変数との相関係数

独立変数名	相関係数	独立変数名	相関係数
Date	0.1145	Resource	0.1743
Packer	0.0067	IAT	0.0263
UPX	0.1323	Section	0.0236
EP	0.2384	VerInfo	0.2228
TLS	-0.3568	FileEntropy	0.4900

5.3 独立変数-独立変数間の相関係数

次に独立変数間で相関が強いものを調査した．相関係数が 0.4 以上だったものについて表 2 に示す．また独立変数のうち質的変数で値が 1（悪性要因となりえる）の度数を表 3 に示す．

表 2 独立変数間の相関係数

項番	独立変数	相関係数
1	Date, EP	0.5409
2	EP, UPX	0.5818
3	Resource, VerInfo	-0.5115
4	Section, VerInfo	0.4514
5	TLS, Resource	-0.6643
6	TLS, Scection	0.4973

表 3 質的変数で値 1 の度数

Date	Packer	UPX	EP	TLS	Resource	VerInfo
31	70	25	79	70	667	290

表 2 の項番 1,2 について，表 3 より UPX と Date の値 1 の割合が少なすぎて全体としてよいモデルが作りにくいと考えられる．項番 2 について UPX=1 のデータのほとんどが EP=1，つまり UPX でパックされたものはエントリポイント異常となっていた．これは UPX Packer の特徴であると考えられる．項番 3,4 について母数を正常検体のみ（Mal=0）とした場合，さらに強い負の相関が出た．その一方で，マルウェア検体のみ（Mal=1）とすると相関が弱まった．項番 5,6 については前節より TLS=1 はほぼ正常検体であったことから，正常検体についての特徴であり母数を正常検体のみとすると相関係数は 0.8 程度と大きくなった．項番 3,5

‡ 判定日：2014/7/15

§ http://www.r-project.org/

から，変数 Resource と他の変数間には負の相関関係があり，原因の推測はできなかったが，正常検体を特定する特徴となりうる可能性はある．

5.4 ロジスティック回帰分析への適用

5.2 節の考察から従属変数 Mal と一番相関係数が高い変数である FileEntropy のみで構築したものをモデル 1 とし表 4 に示す．次に，4.3 節の考察から独立変数間で多重共線性を引き起こす影響は少ないと考え，TLS を除いた全変数を入れたものをモデル 2 として表 5 に示す．

表 4 モデル 1

	偏回帰係数	有意確率	有意水準
切片	-10.7525	<2e-16	***
FileEntropy	1.6686	<2e-16	***
NagelkerkeR2		0.3480	
-2 対数尤度		826.55	
AIC		830.55	
N		772	

表 5 モデル 2

	偏回帰係数	有意確率	有意水準
切片	-15.76401	<2e-16	***
Date	-0.16556	0.85666	
Packer	-0.21159	0.56702	
UPX	-1.72831	0.12401	
EP	2.59862	0.00449	**
Resource	2.46110	1.03e-08	***
IAT	-0.01431	0.71998	
Section	-0.70718	4.61e-11	***
VerInfo	2.88719	2.77e-15	***
FileEntropy	2.06487	<2e-16	***
NagelkerkeR2		0.5238	
-2 対数尤度		675.95	
AIC		695.95	
N		772	

表 6 モデル 3

	偏回帰係数	有意確率	有意水準
切片	-15.5226	<2e-16	***
EP	1.9288	0.00104	**
Resource	2.4759	4.55e-09	***
Section	-0.6743	2.87e-11	***
VerInfo	2.8857	1.19e-15	***
FileEntropy	2.0170	<2e-16	***
NagelkerkeR2		0.5209	
-2 対数尤度		678.76	
AIC		690.76	
N		772	

モデル 2 では 1 と比べ，モデルが対象データの何割を説明できるかを示す NagelkerkeR2 の値が約 34%から 52%に向上していることがわかる． 表 6 に示すモデル 3 はモデル 2 から AIC 値の上昇を抑えつつ有意水準の高い変数に絞ったモデルである．モデル 3 を採用することとして，各独立変数の偏回帰係数の対数をとり算出したオッズ値を表 7 に示す．オッズ値は，質的変数であれば 0 である場合に比べて 1 の場合，何倍，マルウェアであると判定されやすくなることを示している．

表 7 各変数のオッズ値

EP	FileEntropy	Resouce	Section	VerInfo
6.88	7.51	1.18	0.509	10.8

次にモデル 3 において，対象検体ごとに，マルウェアか否かの確率値を算出したグラフを図 2 各検体の確率値分布に示す．横軸は検体 ID で検体 ID:0 から 432 までがマルウェア検体，検体 ID:433 から 771 までが正常検体である．縦軸がマルウェアらしさをあらわす確率値である．

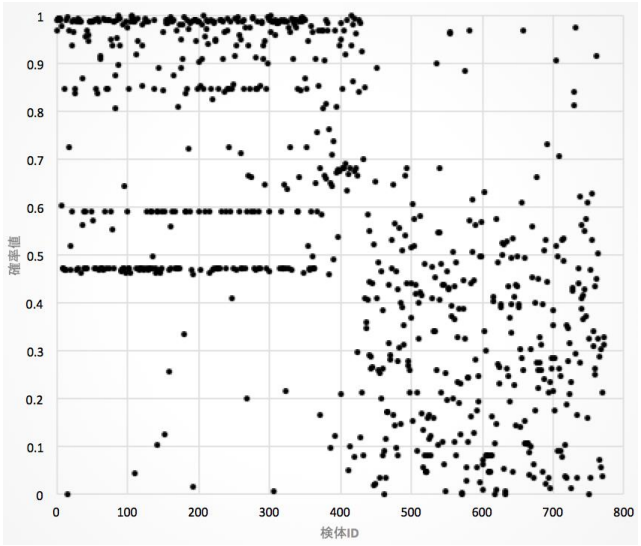

図 2 各検体の確率値分布

表 8 に図 2 で確率値をロジスティック回帰の結果をそのまま表す閾値 0.5 でマルウェア検体，正常検体に分類した結果を示す．グレー部分が False Positive 及び False Negative になり，それぞれ，18.6%，26.1%であった．

表 8 閾値 0.5 とした場合の FP と FN

	確率値 0.5 を閾値		合計
	Negative	Positive	
マルウェア検体	113(26.1%)	320	433
正常検体	276	63(18.6%)	339

表 9 に図 2 で確率値を閾値 0.8 でマルウェア検体，正常検体に分類した結果を示す．同様にグレー部分が False Positive 及び False Negative になり，特に False Positive が 3.8%に抑えられた．これは実運用で誤検出を減らしたい場合に適している．

表 9 閾値 0.8 とした場合の FP と FN

	確率値 0.8 を閾値		
	Negative	Positive	合計
マルウェア検体	194(44.8%)	239	433
正常検体	326	13(3.8%)	339

表 10 に図 2 で確率値を閾値 0.45 でマルウェア検体，正常検体に分類した結果を示す．False Positive が 5.1%に抑えられた．これは我々の要望のように効率的にマルウェア検体を解析したい場合に有効である．

表 10 閾値 0.45 とした場合の FP と FN

	確率値 0.45 を閾値		
	Negative	Positive	合計
マルウェア検体	22(5.1%)	411	433
正常検体	254	85(25.1%)	339

5.5 判別分析への適用

判別分析は，事前に与えられたデータが，どのグループに属していると予めわかっている場合，未知のデータが属するグループを推定する手法である[11]．ロジスティック回帰分析が出力として確率値を取るのに対し，判別分析ではどのグループか等の固定値が出力されることから，単純に分類したい場合に適している．判別分析には，①線形判別関数を用いて，値を直線的モデルに当てはめる方法と，②マハラノビスの距離を用いて，確率を 2 次曲線モデルに当てはめる非線形判別関数を用いる手法がある．本実験では①の方法を実現している表計算ソフト Excel の TREND 関数を用いて算出した．独立変数としてロジスティック回帰のモデル 1，モデル 2，モデル 3 の変数を用いた結果を表 11 表 12 表 13 に示す．モデル 2 が False Positive が 5.9%と良い結果になっているが，False Negative は他と比較しての良し悪しは評価できなかった．

表 11 判別分析結果　モデル 1

	判別分析結果		
	正常判定	マルウェア判定	合計
マルウェア検体	159(36.7%)	274	433
正常検体	286	53(15.6%)	339

表 12 判別分析結果　モデル 2

	判別分析結果		
	正常判定	マルウェア判定	合計
マルウェア検体	178(41.1%)	255	433
正常検体	319	20(5.9%)	339

表 13 判別分析結果　モデル 3

	判別分析結果		
	正常判定	マルウェア判定	合計
マルウェア検体	179(41.3%)	254	433
正常検体	271	68(20.1%)	339

6. 実験結果のまとめ

- 独立変数

3.2 節で定義した独立変数のうち, EP, Resource, Section, VerInfo，FileEntropy が特に有意であることがわかった．また UPX, Date は値 1 のデータが少なすぎたため効果がわからなかった．対象検体を増やして再考察する必要がある．ただ，UPX は EP と相関が強い傾向が見えたので，擬似相関等から，双方は使えない可能性がある．TLS の効果は今回のデータからは見ることができなった．

- ロジスティック回帰

従来研究で FileEntropy が効果的であることが示されているが，今回，EP，Resource，Section，VerInfo の新しく定義した変数を FileEntropy に追加することでさらにフィットしたモデルを構築することができた．またロジスティック回帰の特徴であるが，確率値の判定閾値を変えることにより False Positive，False Negative の割合を修正でき，利用シーンによって調整して用いることができることを示した．

- 判別分析

今回のような白か黒かを判定するような要件の場合，比較的容易に算出でき，実験でもロジスティック回帰と同等もしくは上回る結果が出るケースもあり，有用だと思われる. ただ, 評価基準の観点が False Positive 及び False Negative の割合しかなく，投入する独立変数によるモデルの評価が難しいことがわかった．

7. まとめ

本稿では，マルウェア判定に有効と考えられるファイルの静的な情報から独立変数を定義し，統計的手法を用いてマルウェアらしさを判定する実験を行った結果及び考察を示した．統計手法としてはロジスティック回帰分析と判別分析を用いた．有効な独立変数を見つけることができ，フィットするモデルを構築することができた．このモデルを用いて，マルウェアらしさが高いものを優先的に解析対象とすることで解析効率を向上できる．また，一般的にもシグネチャベースではない静的かつ軽量な判定手法は，マルウェアが爆発的に増加する中，既存の動的解析や静的解析手法と組み合わせて用いることの有効性が期待できる．一部の独立変数の有効性について今回用いた実験検体では判断できなかったので検体を増やして再評価する必要がある．また他にも独立変数として適切なものを探っていく必要もあると考えられる．

参考文献

[1] Guardian News, Antivirus software is dead, says security expert at Symantec, http://www.theguardi an.com/technology/2014/may/06/antivirus-software-fails-catch-attacks-security-expert-symantec （参照 2014/8/15）

[2] 畑田充弘, 稲積孝紀, 有川隼, 田中恭之, “サンドボックス解析結果に基づく URL ブラックリスト生成方式に関する事例

調査, ” 情報処理学会 研究報告コンピュータセキュリティ（CSEC） Vol. 2014-CSEC-66, No. 47, pp. 1 - 6, 2014.
[3] R. Lyda and J. Hamrock, “Using Entropy Analysis to Find Encrypted and Packed Malware,” Security & Pri- vacy, IEEE, Volume 5, Issue 2, 2007 pp. 40–45.
[4] Ye, Y., Chen, L., Wang, D., Li, T., Jiang, Q. and Zhao, M.: SBMDS: An interpretable string based malware de- tection system, Journal in Computer Virology, Vol.5, No.4, pp.283–293 (2009).
[5] 戸部和洋, 森達哉, 千葉大紀, 下田晃弘, 後藤滋樹, “実行ファイルに含まれる文字列の学習に基づくマルウェア検 出 方法, “ MWS2010.（2010 年 10 月）
[6] FFRI : 静的情報に基づいたマルウェア判定指標の検討, FFRI Monthly Research 2014 年 1 月（オンライン）（2014）,入手元（ http://www.ffri.jp/assets/files/monthly_research/MR201401_Consideration_for_indicators_of_malware_likeness_based_on_static_file_information_JPN.pdf）（参照 2014/08/14）
[7] FFRI: ロジスティック回帰分析による未知ファイル分類の有効性, FFRI Monthly Research 2014 年 2 月（オンライン）(2014), 入手元（http://www.ffri.jp/assets/files/monthly_research/MR201402_Effectiveness%20of%20unknown%20malware%20classification%20by%20logistic%20regression%20analysis_JPN.pdf）（参照 2014/08/14）
[8] SANS Institute : Attributes of Malicious Files, SANS Institute reading room (on line) (2014), available from (http://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979)
[9] Michael Ligh, Malware Analyst's Cookbook and DVD : Tools and Techniques for Fighting Malicious Code, 2010.
[10] 秋山満昭, 神薗雅紀, 松木隆宏, 畑田充弘, "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 研究報告コンピュータセキュリティ（CSEC） Vol. 2014-CSEC-66, No. 19, pp. 1 - 7, 2014.
[11] 小西貞則, 多変量解析入門――線形から非線形へ, 岩波書店, 2013.