LPC ケプストラム分析を利用したマルウェアの感染検知

岩野 透† 吉浦 裕† 畑田 充弘‡ 市野 将嗣†

†電気通信大学大学院情報理工学研究科
182-8585 東京都調布市調布ヶ丘 1-5-1

iwano-toru@uec.ac.jp, ichino@inf.uec.ac.jp, yoshiura@hc.uec.ac.jp

‡NTT コミュニケーションズ株式会社
108-8118 東京都港区芝浦 3-4-1 グランパークタワー16F

m.hatada@ntt.com

あらまし トラヒックデータを利用したマルウェア感染検知において，検知精度の向上が課題となっている．そこで，本稿ではLPCケプストラム分析を利用したマルウェア感染検知手法を提案する．LPCケプストラム分析を特徴量抽出に適用することで，特徴量の時間的な変化に着目した感染検知を行う．この提案手法の有効性を評価するため，感染時の通信データとしてD3Mを，正常時の通信データとしてあるイントラネットを流れるトラヒックデータを用いた。TPRとTNRにより、提案手法と先行研究の検知精度との比較を行い，特徴量の有効性を示す．

Applying LPC Cepstrum analysis on malware infection detection

Toru Iwano† Hiroshi Yoshiura† Mitsuhiro Hatada‡ Masatsugu Ichino†
†Graduate School of Informatics and Engineering,
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-si, Tokyo, 182-8585, JAPAN
iwano-toru@uec.ac.jp, ichino@inf.uec.ac.jp, yoshiura@hc.uec.ac.jp

‡NTT Communications Corporation
Gran Park Tower 16F, 3-4-1 Shibaura, Minato-ku, Tokyo, 108-8118 JAPAN

m.hatada@ntt.com

Abstract We propose a method to detect malware infection using LPC Cepstrum analysis. LPC Cepstrum analysis enables variations of temporal features to be applied to malware detection, can also possibly improve the accuracy. To evaluate this method, we used D3M as infected traffic data, intranet traffic as normal traffic. Then we calculate the accuracy by TPR and TNR, and compare to previous works.

1 はじめに

近年，インターネットを利用したサービスが拡大し，様々な場面で必要不可欠な存在となっている，その一方で，マルウェアへの感染による犯罪被害が社会問題となっている．

このようなマルウェア感染への主な対策として，マルウェアのバイナリパターンの特徴を利用するシグネチャ型の検知手法がある．しかし，シグネチャ型の検知手法は事前にマルウェア検体を解析する時間と労力が必要であり，短期間で大量に出現する未知のマルウェアには対応できない．そのため，マルウェアへの感染を前提とし，感染後に被害を拡大させないためのマルウェア対策手法(感染検知)が重要となっている．

マルウェアに感染した PC は感染後に特有な通信を行うと考えられる．具体的には C&C サーバ

との通信やポートスキャン・インターネットへの疎通確認などである．これらの通信は未知のマルウェアも共通して実行する挙動と考えられるため，本研究ではトラヒックデータに着目した感染検知手法を検討する．しかし，トラヒックデータに着目した検知手法は誤検知が多いという課題がある．そこで検知精度の向上を目的として，本研究ではトラヒックデータからの特徴量の抽出手法にLPCケプストラム分析を適用することを提案する．

以下，2.で既存研究に利用されている特徴量を整理し，3.で本研究における特徴量手法を述べ，4.で提案手法，既存手法における検知精度の評価実験の内容を説明し，5.でこの結果を述べる．6.で利用した攻撃通信データの調査結果を踏まえた提案手法の検知結果についての考察を行う．

2　先行研究における利用特徴量

本章では，既存のマルウェア感染検知やネットワーク異常検知に関する研究で用いられているパケットのヘッダ情報を利用した特徴量について整理する．

宮本らはパケットのヘッダ情報の種類毎のパケット数を特徴量とし，SVM を利用した異常検出手法を提案している [1]．利用したヘッダ情報は TCP パケットのフラグ情報と TCP・UDP・IP のポート番号情報・パケットサイズである．種類毎にそれぞれパケット数を 1 分毎の区間で算出し，SVM を利用して異常検知を行うものである．

川元らはヘッダ情報から得られた 36 種類の特徴量について感染検知における有効性の評価を行っている[2]．この実験では TPR・TNR の観点から評価を行い，4 種類の特徴量が感染検知に有効であることを示している．

これらの先行研究では，ヘッダ情報から算出できる統計的な数値が特徴量として利用されていた．しかし，マルウェアの感染後の通信挙動は日々変化しており，実用を考えると先行研究では十分な検知精度が得られていない．

より幅広い感染後の通信挙動に対しても高い検知性能となる検知手法を検討する必要がある．

3　本研究における提案手法

本章では提案手法であるトラヒックデータからの時間波形の作成方法と，時間波形への LPC ケプストラム分析の適用について説明する．

3.1　パケットスロット

本研究では，トラヒックデータからの特徴抽出にパケットスロットを用いた．パケットスロットとは，トラヒックデータを特定のパケット数で区切った範囲のことを示す．

特徴抽出の別の単位としては時間間隔を用いる手法やフロー（送受信 IP アドレス，送受信ポート番号，プロトコルの 5 項目が同一のパケット群）を用いる手法がある．しかし，時間間隔を用いる場合には，トラヒックの混雑状況に応じて特徴量抽出を行う適切な間隔が変動してしまう場合があり，フロー毎での特徴量抽出は通信が終了するまで特徴量の抽出を行うことができないため，早期検知には不向きである．一方パケットスロットを用いる方法は，パケットスロット幅を調整することで特徴量の取得時間を短縮できるので，早期検知が期待できる．そのため，本研究では感染後の早期検知に適すると考えられるパケットスロットを利用する．

3.2　特徴量の時間波形化

本研究ではパケットスロットから取得した特徴量を時間波形として利用する．具体的には取得したパケットスロットの特徴量の値を時系列順にプロットすることでグラフ化する．この特徴量の推移を時間波形として利用する．時間波形の一例を図 2 に示す．

3.3　LPC ケプストラム分析

本節では LPC ケプストラム分析の概要とその利点について説明する．

3.3.1　LPC ケプストラム

LPC ケプストラムは LPC(線形予測符号) 分析により予測した波形から短時間振幅スペクトルを算出し，このスペクトルの対数の逆フーリエ変換によって生成される[3]. LPC ケプストラム分析を行うことで時間波形からスペクトル包絡とスペクトル微細構造を近似的に分離して抽出できる．

LPC 分析で予測した時間波形$\mathrm{x}(t)$を二つの信号で表すことを考える．このとき，$\mathrm{x}(t)$は周期信号$g(t)$とインパルス応答$h(t)$の畳込みで表すと(1)式のようになる．

$$\mathrm{X}(\omega) = G(\omega)H(\omega) \tag{1}$$

ここで, $X(\omega), G(\omega), H(\omega)$は$x(t)$，$g(t)$，$h(t)$のフーリエ変換である．(1)式の対数変換を逆フーリエ変換すると(2)式となる．

$$c(\tau) = \mathcal{F}^{-1}\log|G(\omega)| + \mathcal{F}^{-1}\log|H(\omega)| \quad (2)$$

この(2)式が LPC ケプストラムであり，第一項はスペクトル微細構造を，第二項はスペクトル包絡を表す．スペクトル包絡とスペクトル微細構造はフィルタで分離することが可能であり，本研究ではスペクトル包絡を特徴量として利用する．トラヒックデータからスペクトル包絡作成までの流れを図 1 に示す．

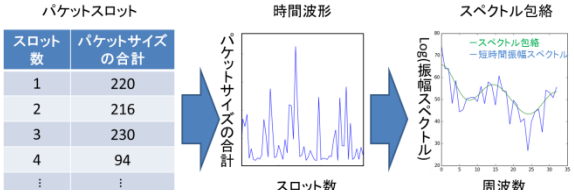

図 1 トラヒックデータへの LPC ケプストラム分析の適用概要図

3.3.2 LPC ケプストラム分析による利点

スペクトル包絡は時間波形の振幅情報を表している．振幅情報にはトラヒックデータの時間的な変化が含まれると考えられるので，トラヒックデータから通信の特徴となる時間的な変化を抽出し利用することが可能になる．

また，LPC ケプストラム分析はスペクトル包絡のピーク特性を強調して抽出する．これにより，トラヒックデータの時間的な変化を強調して抽出でき，感染時と正常時の通信の時間的な変化を明確に捉えられる可能性がある．

以上より，特徴量の時間的な変化を捉えるために LPC ケプストラム分析によって抽出されるスペクトル包絡の情報を特徴量に利用する．

4 検知精度の評価実験

提案手法の検知精度を評価するため，以下で検知精度の評価実験を行なった．

4.1 評価実験の概要

提案手法における感染検知の流れを図 2 にまとめた．また，検知精度の比較のために先行研究についても併せて実施する．先行研究は特徴量の抽出後，LPC ケプストラム分析を行わず，パケットスロットの値を特徴量として利用するものとする．

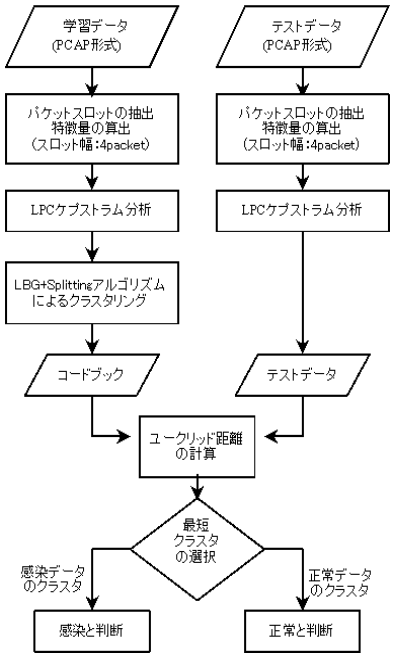

図 2 提案手法におけるフローチャート

4.2 実験データ

感染時の通信データとしてMWS Datasets 2014のD3M [4]を利用した．また，正常時の通信データとしてあるイントラネットを流れるトラヒックデータを利用した．感染時・正常時共に2012年のトラヒックデータを学習データとして利用し，テストデータには2012・2013・2014年のトラヒックデータを利用した．なお，2012年のトラヒックデータは検知精度を正確に評価するために，学習データとテストデータに分割し利用している．

4.3 評価特徴量

先行研究[2]で利用されていた特徴量の中から，13種類の特徴量を選択し表1にまとめた．この特徴量は時間波形を作成した際にLPCケプストラム分析が適用できることを基準に選択したものである．

4.4 感染検知の手法

ベクトル量子化を用いて，感染時通信のスペクトル包絡の情報から学習を行なった感染コードブックと，正常時通信のスペクトル包絡の情報から学習を行った正常コードブックを予め作成する．ベクトル量子化のアルゴリズムには，LBG+Splittingアル

ゴリズム[5]を用い，レベル数（コードブックの数）は2，4，8の3種類とした．

そして，予め感染時通信か正常時通信かのラベル付けされた各特徴量のスペクトル包絡の情報を与え，テストデータと感染，正常コードブックとのユークリッド距離を計算し，感染(正常)コードブックとの距離の方が小さければ感染（正常）と識別する．

表 1　利用特徴量の一覧表

番号	特徴量の説明
1	パケットサイズの総数[byte]
2	パケットサイズの平均[byte]
3	パケットサイズの最小[byte]
4	パケットサイズの最大[byte]
5	パケットサイズの標準偏差[byte]
6	到着間隔の平均[ミリ秒]
7	到着間隔の最小[ミリ秒]
8	到着間隔の最大[ミリ秒]
9	到着間隔の標準偏差[ミリ秒]
10	SYNパケット数[packet]
11	FINパケット数[packet]
12	PSHパケット数[packet]
13	ACKパケット数[packet]

4.5　評価の指標について

本実験では検知精度の評価指標として TPR と TNR を利用した．TPR とは感染データを感染データと正しく分類できた割合，TNR とは正常データを正常と正しく分類できた割合のことを指す．

5　実験結果

利用した特徴量毎の TPR を表 2 に，TNR を表 3 にまとめた．ここで表記している検知精度はベクトル量子化レベル数 2,4,8 の時からそれぞれ算出した検知精度の平均値である．

表 2,3 では各手法における TPR と TNR で最も検知精度の良い特徴量を太字で表記した．これは，各手法を実運用するとき，検知に利用される特徴量の候補となるからである．

表 2　各実験における特徴量毎の TPR

特徴量	提案手法 TNR (%)			先行研究 TNR (%)		
	2012	2013	2014	2012	2013	2014
1	48.1	49.0	46.9	30.8	51.7	71.8
2	43.0	45.5	47.9	30.8	51.6	71.8
3	51.5	50.9	51.4	12.5	16.1	17.7
4	46.6	50.7	49.8	40.7	64.2	80.5
5	45.3	47.1	48.2	37.3	61.6	67.7
6	68.0	67.9	61.1	**75.7**	**77.9**	**91.1**
7	**93.1**	**92.8**	**90.0**	54.1	53.9	52.1
8	80.0	79.3	77.0	57.7	56.9	53.6
9	72.5	71.3	67.7	57.4	56.7	53.5
10	40.8	42.1	32.6	4.0	3.4	1.3
11	59.3	58.6	50.9	3.0	3.5	1.3
12	75.7	75.2	70.0	39.6	50.0	22.0
13	62.8	62.0	56.2	50.2	50.4	49.8

表 3　各実験における特徴量毎の TNR

特徴量	提案手法 TPR (%)			先行研究 TPR (%)		
	2012	2013	2014	2012	2013	2014
1	71.2	74.1	65.4	69.4	93.4	82.5
2	78.8	66.7	53.7	69.4	93.6	82.5
3	57.7	38.9	36.5	**84.7**	**49.5**	**95.8**
4	73.1	74.1	58.2	64.4	91.5	76.8
5	73.9	72.2	65.5	66.5	93.1	76.4
6	51.9	46.3	31.8	42.8	82.1	13.5
7	36.5	3.7	39.8	49.4	49.0	46.3
8	25.0	16.7	22.1	45.3	48.0	44.7
9	44.2	38.9	29.2	51.5	49.6	45.5
10	89.1	50.0	47.8	20.3	1.9	27.1
11	82.6	43.8	54.5	19.7	2.1	17.2
12	**87.0**	**100**	**76.2**	42.4	50.0	49.0
13	23.9	68.8	28.5	34.6	90.3	5.0

表 2 より 2012,2013 年度の TPR において提案手法の検知性能が先行研究の検知性能よりも高いことが分かった．また表 3 より，2012,2013 年度の TNR において提案手法の検知性能が先行研究の検知性能よりも高いことが分かった．

6　考察

本章では 5 章で得られた実験結果を，時間波形の概形とスペクトル包絡の概形の観点から考察し，提案手法の評価を行う．

6.1　時間波形の概形による考察

本節では TPR・TNR それぞれについて時間波形の概形と通信データの挙動について考察する．

6.1.1　TPR における特徴量の有効性

提案手法の TPR で最も優秀な検知精度だった PSH パケット数について考察を行う．先行研究での本特徴量の TPR が 2012 年は 42.4%,2013 年は 50.0%,2014 年は 49.0%だったのに対し，提案手法での本特徴量の TPR が 2012 年は 87.0%,2013 年は 100%,2014 年は 76.2%と大幅に向上した．この原因調査のため，攻撃通信データを解析した．

6.1.1.1　PSH パケット数の時間波形

初めに，感染時のテストデータと感染時・正常時の学習データについて，PSH パケットのヒストグラムを作成し，図 3 にまとめた．ヒストグラムでは横軸に 1 スロット当たりの特徴量の値が，縦軸に出現頻度が表記される．そのため，データの分布を視覚的に確認することができる．

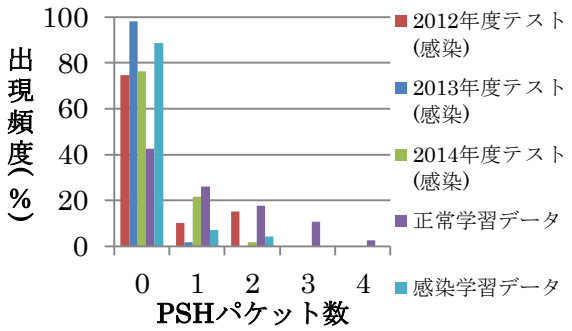

図 3 PSH パケット数のヒストグラム

図 3 より出現頻度に違いこそあるものの，PSH パケット数 0 が最も出現頻度が高いことが分かる．このため，先行研究のような時間的な変化を重視しない特徴量抽出手法では正常と感染で差異が明確化されず，マルウェアを正しく識別することができなかったと考えられる．

次に，学習データにおける PSH パケット数の時間波形の一部を図 4,5 で示す．図 4 より，正常時のトラヒックデータは連続的な時間波形であることが分かる．また図 5 より感染時のトラヒックデータは前半部分では規則的な波形となり，後半は散発的なインパルス波形となっていることが分かる．

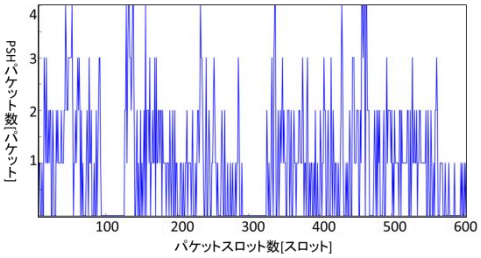

図 4 PSH パケット数の時間波形(正常)

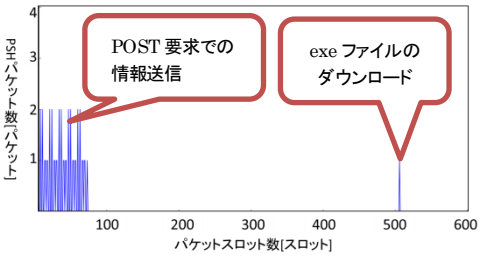

図 5 PSH パケット数の時間波形(感染)

図 4,5 の波形の違いを分析するために通信データの挙動について分析を行なった．その結果，正常時のトラヒックデータでは http request 要求や SSL 通信・FTP 通信などに PSH フラグが利用されていた．これらの通信はユーザが日常的に利用する通信であり，特に http request 要求は Web ページの画像ファイルの表示に利用されている．そのため，正常時のトラヒックデータは連続的な時間波形になったと考えられる．

それに対して，マルウェアの攻撃通信データでは http request 要求にのみ PSH フラグが利用されていた．この request 要求では他のマルウェア検体のダウンロードや情報の送受信などが行われていた．これらの通信はマルウェア検体が定期的・

散発的に行う挙動であるため，感染時のトラヒックデータは図 5 の様な定期的，散発的な波形になったと考えられる．

以上より，正常時・感染時の通信には時間的な変化に異なる特徴があることがわかった． LPC ケプストラム分析を適用することでこの特徴を捉えることができ，正常時と感染時のトラヒックデータを分離できたものと考えられる．

6.1.1.2 検体種類ごとの検知精度

6.1.1.1 より学習データでは正常と感染の特徴を分離できたが，テストデータのマルウェア検体は学習データとは異なる．そこで，使用した攻撃通信データについて TrendMicro[6]と Kaspersky[7]を利用して調査を行なった．学習データのマルウェア名称(接頭語・ファミリー名)と時間波形分類を表 4 に，テストデータのマルウェア名称(接頭語・ファミリー名)と時間波形分類・検知精度を表 5 にそれぞれまとめた．なお，時間波形分類については時間波形の傾向によって 3 種類に分類したものであり，詳細は後述する．

表 4　感染時通信の学習データ検体内訳

接頭語	ファミリー名	時間波形分類
トロイの木馬	Generic	iii
	IRCBRUTE	ii
	FakeAV	i

表 5　感染時通信のテストデータ検体内訳

年度	接頭語	ファミリー名	波形分類	検知精度[%]
2012	トロイの木馬	Ransom	ii	88.9
		Generic	i	100
2013	バックドア	Win32	i	100
		Win32	i	100
		Win32	i	100
		Win32	i	100
		Win32	i	100
	トロイの木馬	Dropper	i	100
2014	トロイの木馬	VILSEL	i	100.0
		VILSEL	i	100.0
		Kryptk	ii	83.3
		Kryptk	iii	59.5
	スパイウェア	Fareit	ii	100

表 4,5 より，今回利用した学習データはトロイの木馬型の検体のみだったと分かる．また，時間波形の分類によって検知精度が異なり，接頭語やファミリー名とは関係がないことが分かる．

この原因調査のため，攻撃通信データの時間波形を調査した．その結果，時間波形は 3 種類に分類され、攻撃通信データの挙動が異なることが判明した．以下で時間波形の分類と挙動を説明する．

i.　散発的なリクエスト要求を行う検体

これらの検体は散発的に http request 要求を行なっていた．http request 要求では php や js ファイルへの情報送信，偽造ファイルのダウンロードが行われており，解析したマルウェア検体が他の攻撃の踏み台に利用されたと考えられる．時間波形の例（2013 年の Dropper 検体の時間波形）を図 6 に示す．この波形分類は学習データにも含まれ，時間波形を学習できたので，高い TPR になったと考えられる．

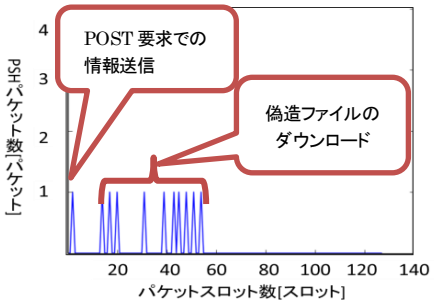

図 6 散発的な http request の時間波形

ii.　定期的な http requset 要求を行う検体

これらの検体は定期的に http request 要求を行なっていた．http request 要求は送信先 IP アドレスの異なる同じ名称の php ファイルに送信されていた．これは攻撃側が長時間の情報収集をするため通信先を切り替えて利用してたことが考えられる．時間波形の例（2012 年の Ransom 検体の時間波形）を図 7 に示す．この波形分類は学習データにも含まれていたため，高い TPR となったが，定期的な通信の準備段階などでは正しく識別ができなかった．

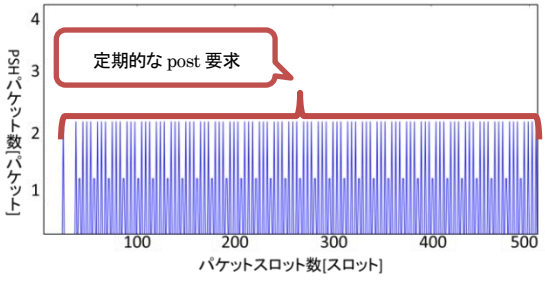

図 7 定期的な http request の時間波形

iii. ランダムな http request 要求を行う通信

これらの検体はランダムに http request 要求を行なっていた. 複数の IP アドレスと同時並行で通信を行なっており，これによって正常時の通信と似た波形となったと考えられる．これらの検体の時間波形の例(2014 年の Kryptk)を図 8 に示す．この波形分類は学習データにも含まれていたが，波形に一定の規則性がなく，学習した時間波形の概形とテストデータの時間波形の概形が同一でなかったため，検知精度が悪かったと考えられる．

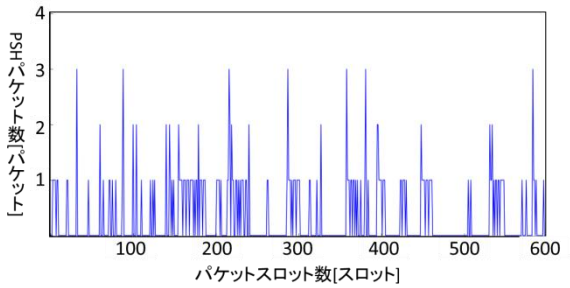

図 8 連続的な http request の時間波形

以上より，PSH パケット数はデータの分布は感染と正常で差異が少ないが，時間的な変化の規則性が大きく異なることが分かる．PSH パケット数を時間波形として捉えることで，ｉ・ii のような時間波形に分類されるマルウェアを検知できたと考えられる．

6.1.2 TNR における特徴量の有効性

提案手法の TNR で最も優秀な検知精度だった到着間隔の最小値について考察を行う．先行研究での本特徴量の TNR が 2012 年は 54.1%,2013 年は 53.9%,2014 年は 52.1%だったのに対し，提案手法での本特徴量の TNR が 2012 年は 93.1%,2013 年は 92.8%,2014 年は 90.0%と大幅に向上した．この原因調査のため，通信データの解析を行った．

6.1.2.1 到着間隔の最小値の時間波形

初めに, 各データにおける PSH パケットのヒストグラムを図 9 にまとめた.

図 9 より出現頻度に違いこそあるものの，到着間隔の最小値が 1 ミリ秒程度の時が最も出現頻度が高いことが分かる．このため，先行研究のような時間的な変化を重視しない特徴量抽出手法では正常と感染で差異が明確化されず，正常時通信を正しく識別することができなかったと考えられる．

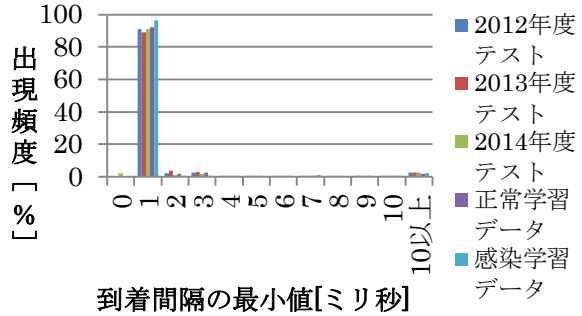

図 9 到着間隔の最小値のヒストグラム

次に，学習データにおける到着間隔の最小値の時間波形の一部を図 10,11 で示す．また，テストデータの時間波形を図 12 で示す．

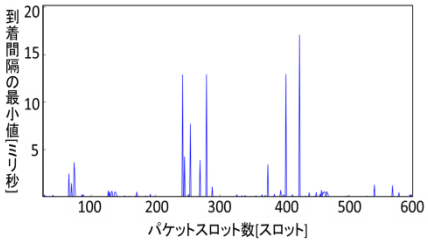

図 10 到着間隔の最小値の時間波形(正常)

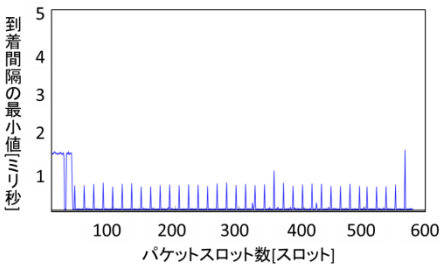

図 11 到着間隔の最小値の時間波形(感染)

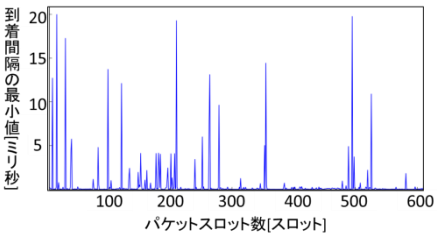

図 12 テストデータの到着間隔の最小値の時間波形(正常)

これより感染時と正常時の時間波形には差異があることが分かる．これは，正常時の通信がユーザの行動次第で変化するのに対し，マルウェアの通信は攻撃者によって定義されたプログラムに従い，自動的に行われるためだと考えられる．

以上より，到着間隔の最小値はデータの分布は感染と正常で差異が少ないが，時間的な変化のパターンが大きく異なることが分かる．到着間隔の最小値を時間波形として捉えることで，この時間的な変化を捉えることができたと考えられる．

6.2　スペクトル包絡での特徴量の有効性

本節では TPR・TNR それぞれについてスペクトル包絡の概形に基づいた考察を行う．

6.1 で考察した特徴量について，提案手法のコードブックをスペクトル包絡として図 13,14 に示す．

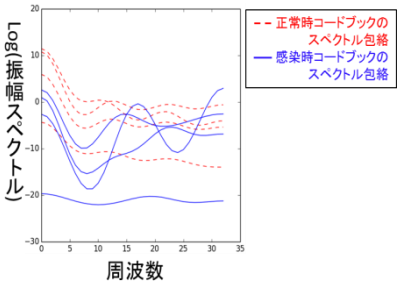

図 13　PSH パケット数のコードブック

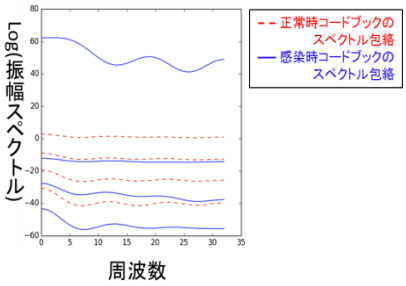

図 14　到着間隔の最小値のコードブック

図 13 ではスペクトル包絡のホルマント情報（曲線の極大値）など，スペクトル包絡の概形が正常時と感染時で異なっていることが分かる．また，図 14 では正常と感染のスペクトル包絡が重なり無く分離していることが分かる．これより，正常・感染におけるスペクトル包絡が分離されていると考えられる．さらに、スペクトル包絡は通信の時間的な変化を表すので，正常時と感染時の通信の時間的な変化の差を捉えることができたと考えられる．

6.3　本検知手法の発展性

本検知手法では高い TPR・TNR を得ることができた．特に高い TNR を利用して正常時通信の識別に本研究は有効と考えられる。その特性を活かし，始めに正常時通信を識別し，該当しないトラヒックデータへの感染識別を行う二段階検知手法への利用が期待できる．

7　まとめ

本論文では LPC ケプストラム分析を利用した特徴量抽出手法を提案し，この検知精度を評価した．結果として，TPR では PSH パケット数で，TNR では到着間隔の最小値で先行研究よりも検知精度が向上した．今後はさらなる検知精度の向上を目的として，複数の特徴量を組み合わせた検知手法や二段階検知手法についての検討していく．

参考文献

[1] 宮本貴朗，小島篤博，泉正夫，福永邦雄：SVM を用いたネットワークトラヒックからの異常検出，電子情報通信学会論文誌，Vol.B，通信 J87-B(4)，pp593-598，2004

[2] 川元研治，市田達也，市野将嗣，畑田充弘，小松尚久，”マルウェア感染検知のための経年変化を考慮した特徴量評価に関する一考察”，”マルウェア対策研究人材育成ワークショップ 2011(MWS2011)，2011

[3] 吉井貞熙：ディジタル音声処理，東海大学出版会，1985

[4] 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.

[5] Linde Y, Buzo A. and Gray R, ”An Algorithm for Vector Quantization”, IEEE Trans, Commun, Vol.28 No,1 pp84-95,1980

[6] トレンドマイクロ セキュリティデータベース
http://www.trendmicro.co.jp/jp/index.html

[7] Kaspersky セキュリティデータベース
http://www.viruslistjp.com/