社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－CSEC－42（20）
2008 / 7 / 25

ふるまいに着目した未知の亜種ウイルスの識別

三森 春佳† 阿部 公輝†

† 電気通信大学 情報工学専攻 〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †{mimori,abe}@cacao.cs.uec.ac.jp

あらまし 大量メール送信ウイルスに着目し，ウイルスを実際に動かしてそのふるまいを観測した．ウイルスにラベルを付けてふるまいを学習し，未知の亜種ウイルスを自動で識別を行った．8 種類各 3 個のウイルス計 24 個を用いて学習した結果，8 割強の割合で未知ウイルスの種類を正しく識別できた．ウイルスの種類が異なると識別率も異なるが，確認時期が近い亜種の識別率は高い傾向にあることが分かった．

キーワード 未知ウイルス，亜種，データマイニング

Detection of Unknown Computer Virus Variants Based on Computer Behavior

Haruka MIMORI† and Kôki ABE†

† Department of Computer Science,he University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi,Tokyo 182-8585 Japan
E-mail: †{mimori,abe}@cacao.cs.uec.ac.jp

Abstract We observed the behavior of computer viruses by monitoring the behavior of computers infected by the mass-mailing viruses. We employed machine learning methods to learn the behaviors of known virus variants to identify unknown variants. The learning results using eight kinds of viruses each with three variants revealed that the methods can correctly identify unknown virus variants with an accuracy of more than 80%. It was also found that the accuracy of identification differs among different kinds of viruses, but the accuracy of identifying virus variants discovered close in time tends to be high.

Key words Unknown viruses, variants, data mining

1. はじめに

コンピュータウイルスはインターネット上の重大な脅威である．ウイルスを検出する基本的な手法としてパターンマッチングがある．この手法はウイルスに特徴的なコードからシグネチャ（パターン）を作成し，このシグネチャと検査対象となるコードとのパターンマッチを行う．しかしシグネチャがすでに作成されているウイルスでないと検出できないため，未知のウイルスには対応できない．

近年，コンピュータウイルスを作成する目的が愉快犯的なものから金銭的な利益へと移行してきている．そのため長く潜伏できるよう，以前のような派手な活動は行わず，見つかりにくいようなウイルスが次々と現われてきている．シグネチャを用いるパターンマッチではこのようなウイルスは検出できない．シグネチャを頻繁に更新することは難しく，また，感染するたびに自らが変化していくタイプのウイルスなどはパターンマッチングでは検出できない．

このような未知のウイルスに対応するため，近年様々な手法が研究されている．静的ヒューリスティック法はコードからウイルスらしいコードを検出する手法である．ビヘイビア法（動的ヒューリスティック法）はウイルスを実際に動かし，その動作を観察することでウイルスかどうかを判定する [1]．ビヘイビア法はパターンマッチングでは検出できないウイルスに対応できる可能性がある．データマイニングを適用した研究 [2] は，ウイルスはそのふるまいに固有の特徴があり，環境が異なっても識別できることを実験的に示した．

ウイルスは一つ新しいものが登場するとそれに少し変更を加えた亜種が多数登場し，その亜種同士は基本的には似たような構造をとる．これらの亜種に対する対策は，新たな種類のウイルスに対する対策に比べ考えやすい．しかし構造が似ていても同じシグネチャで亜種ウイルスを検出することはできない．異なるウイルスでも構造が似ていればふるまいも似ていると考

えられる．ふるまいの特徴に着目すれば未知のウィルスを既知ウィルスの亜種と認識できる可能性がある．

本研究では，大量メール送信ウイルスにおいて，ウイルスのふるまいの特徴をデータマイニングで学習し，未知の亜種ウイルスを自動で識別する．特徴量選択手法とクラス分類手法により識別率がどのように変化するかを実験により調べる．以下，第2章で実験使用するウイルス，第3章でデータマイニングで用いる特徴量ランク付けとクラス分類，第4章で実験方法，第5章で実験結果とその考察を述べ，第6章でまとめる．

2. 実験で使用するウイルス

ウイルスはその感染方法や感染後の動作に様々なものがある．感染経路はメールの添付ファイルによるものやネットワークの脆弱性をついて侵入するもの，P2Pによる感染などがある．感染すると，ファイルを書き換えたり，特定のサイトへの接続を妨害したり，感染したホストの情報を送信するなどユーザーの不利益になる動作を行う．

ここでは，ウイルスを実際に動かすが，実ネットワークに拡散しないよう隔離された環境のなかで動作させる．実験ではメールの添付ファイルで感染するウイルスに着目する．この型のウイルスは，大量メール送信で感染し拡散するため，メールサーバ周りを用意するとその動作が確認でき，実験環境の構築が比較的容易である．

実験では，主にウイルスサイト[3]にアップロードされているものを用いるが，一部メール添付で送られてきたものも使用する．1種類のウイルスにつき3個の亜種を用意した．亜種の名前の末尾に付けたアルファベットは確認された順を表す．Aは最も早い時期に確認された亜種を表す．いずれのウイルスも特定の拡張子を持つファイルからメールアドレスを収集し，独自のSMTPエンジンを使用してメール送信を行う．以下に各ウイルスの特徴を述べる[4][5]．

（1） Bagz

Bagzはhostsファイルの書き換え，特定のサイトへの接続の妨害やアンチウイルスソフトのアップデートの妨害を行う．特定のアンチウイルスソフト関連のレジストリを削除する．使用した亜種はBagzC, BagzD, BagzF.

（2） Bagle

Bagleはバックドアを開け，特定のプロセスを終了することでアンチウイルスソフトのアップデート妨害を行う．ある決まった日時になると活動を停止する．使用した亜種はBagleA, BagleC, BagleE.

（3） Doombot

Doombotはhostsファイルの書き換え，特定のサイトやアンチウイルスソフトのアップデートの妨害を行う．またメールの他にIRC経由でも感染する．IRC通信路に接続するためTCPのランダムなポートをオープンにする．使用した亜種はDoombotB, DoombotF, DoombotG.

（4） Fanbot

Fanbotはhostsファイルの書き換えを行う．ウイルスが実行されると「ファイルが開けなかった」という旨の偽のエラーメッセージダイアログを表示する．IRC通信路に接続するためTCPの特定のポートをオープンにする．使用した亜種はFanbotA, FanbotC, FanbotD.

（5） Kipis

Kipisはメールの他，P2Pによる拡散も試みる．特定の拡張子を持つファイルからメールアドレスを収集し，独自のSMTPエンジンを使用してメール送信を行う．使用した亜種はKipisA, KipisB, KipisG.

（6） Klez

Klezはメール送信のほか，ファイル共有での感染も行う．Outlook Expressの脆弱性を利用しメールをプレビューしただけで感染する機能を持っている．使用した亜種はKlezA, KlezB, KlezC.

（7） Mimail

Mimailはトロイの木馬タイプのウイルスで，感染ホストの特定の情報を決まったメールアドレスに送信する．Outlook Expressの脆弱性を利用しメールをプレビューしただけで感染する機能を持っている．使用した亜種はMimailA, MimailB, MimalC.

（8） Mydoom

Mydoomはバックドアを開け，トロイの木馬をダウンロードしたり攻撃者が感染ホストに侵入可能な状態にしたりする．使用した亜種はMydoomA, MydoomM, MydoomQ.

3. 特徴量ランク付けとクラス分類の手法

本研究は，大量メール送信ウイルスにおいて，ウイルスのふるまいの特徴をデータマイニングで学習し，未知の亜種ウイルスを自動で識別する．データマイニングにおいて，特徴量をランク付けして用いる特徴量の数を絞ったあと，絞られた特徴量を基に分類器を作成する．ここでは実験で用いる3つの特徴量ランク付け手法（Chi Square, Gain Ratio, ReliefF）とクラス分類手法（決定木，Naive Bayes，Bayesian Networks）の概要を簡単に述べる．

3.1 特徴量ランク付け

Chi Squareは独立性を表すカイ二乗統計量を用いて重み付けを行う[7]．この値が大きいほどクラスとの独立性が低い，つまり関連性が高い．

Gain ratioは情報利得における偏りを改善した手法である[7]．情報利得とは分類集合に分類規則を追加した場合の情報量の差分である．

Relief Fは，あるサンプルに近い二つのサンプルから特徴量の重みを求める[8]．具体的には，あるサンプルと同じクラスで最も近いサンプルと，異なるクラスで最も近いサンプルをそれぞれ選ぶ．距離は全ての属性を用いたユークリッド距離で計算される．

これらの特徴量ランク付け手法にはWeka[6]のスタンダード版を使用する．

3.2 クラス分類

決定木は代表的な分類器で，その構造は木構造である．基本的には貪欲アルゴリズムを用いて，集合を再帰的に分割するこ

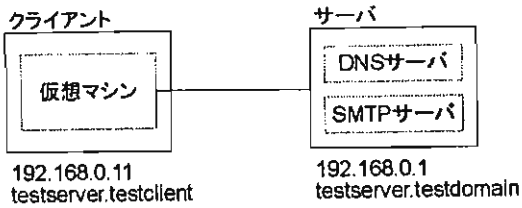

図 1 ウイルスを動作させる環境

Fig. 1 Experimental environment to observe virus activities

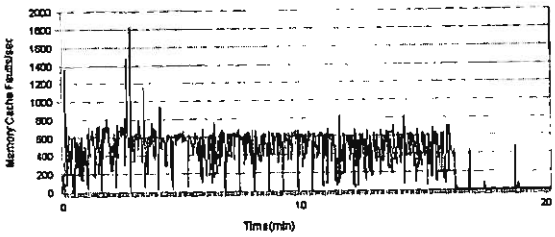

図 2 BagzC を動作させた際の Memory Cache_Fault/sec カウンタの時間変化

Fig. 2 Time variation of Memory Cache_Fault/sec counter when BagzC is executed.

とで木を生成する．生成後，枝刈りを行うことにより精度の向上を図る．本実験では一般的な決定木アルゴリズム C4.5 [9] を実装した Weka 版 J48 を使用する．

ベイズの定理により，あるクラスの事後確率は事前確率とそのクラスが与えられたときの特徴量を持つ確率とに比例する．Naive Bayes は事後確率の計算において，あるクラスが与えられたときの特徴量を持つ確率は他のどの特徴量とも条件付き独立との仮定を導入し，計算量を減らす．Naive Bayes には Weka [6] のスタンダード版を使用する．

Bayesian Networks もベイズの定理に基づくが，Naive Bayes のようにあるクラスが与えられたときの特徴量を持つ確率は他のどの特徴量とも条件付き独立との仮定は設けない．因果関係を条件付き確率でラベル付けした非循環有向グラフ（確率ネットワーク）で表す．Bayesian Networks についても Weka [6] のスタンダード版を使用する．

4. 実験方法

ネットワークから隔離された環境を構築し，その上でウイルスを実際に動かしてその動作を観測する．図 1 のように，ウイルスを動作させるクライアントマシンと DNS サーバと SMTP サーバが動作するサーバマシンを用意する．これらのマシンの OS は Linux（CentOS 5）とする．クライアントマシンは，その上でウイルスを動作させるので，簡単に元の状態に戻せるよう仮想マシンとする．仮想環境は Windows XP SP2 とする．

ウイルスの動作は [2] と同様 Windows Performance Counter(WPC) [10] を使用して観測する．WPC は複数のカウンタから構成され，各カウンタは一つの特徴量を保持する．特徴量には例えば TCP: Connection_Passive（TCP 接続

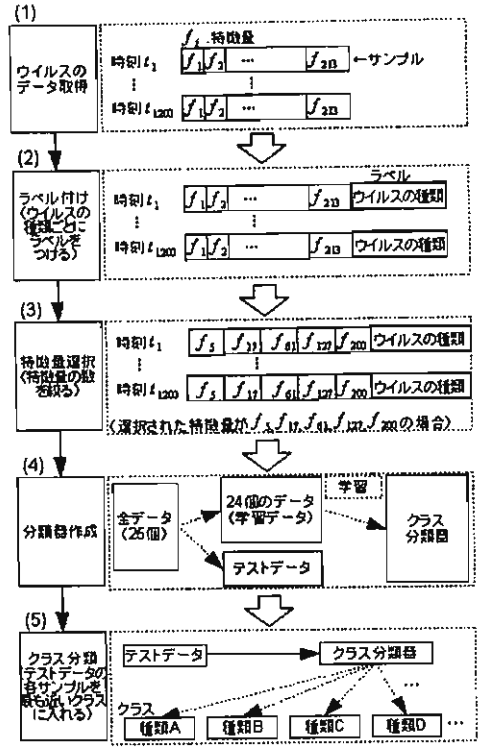

図 3 実験の手順

Fig. 3 Experimental steps

が LISTEN 状態から SYN-RCVD 状態に直接移行した回数）などがある．図 2 に例を示す．これは BagzC を動作させた際に Memory: Cache_Fault/sec カウンタ（ページフォールトの回数を表し I/O 操作量の指標となる）の値を 20 分間観測したものである．実験では 213 個のカウンタ，すなわち 213 個の特徴量を用いる．カウンタ i は特徴量 f_i，$i = 1, \ldots, 213$，を保持する．

実験の手順を図 3 に示す．(1) まず，ウイルスのデータを取得する．8 種類のウイルス (Bagz, Bagle, Doombot, Fanbot, Kipis, Klez, Mimail, Mydoom) 各 3 個計 24 個を動作させ，20 分間 1 秒ごとに WPC をサンプルし，各時刻 $t_1, \ldots, t_{1200}$ で $f = (f_1, f_2, \ldots, f_{213})$ を求める．(2) 次に，t_i におけるサンプル値に種類ラベルをつけ，$f(t_i)$ とする．$f(t_1), \ldots, f(t_{1200})$ を 1 つのデータとする．ウイルスが動作していない状態でも同様にサンプルし，計 25 個のデータを取得する．(3) さらに，特徴量を選択する．213 個の特徴量のうちウイルス分類に貢献する度合が高いと考えられる特徴量ほど，より高いランクを付ける．特徴量のランク付けに前節で述べた Chi Square(CS), Gain Ratio(GR), ReliefF(RF) を用いる．各サンプル値 $f = (f_1, f_2, \ldots, f_{213})$ の成分のうち，上位にランクされるもの以外は削除する．(4) そして，25 個のデータのうち，検出を試みるデータ（テストデータ）一つを取り除き，残り 24 個のデータを学習データとする．学習データを使ってクラス分類器を学習させる．ク

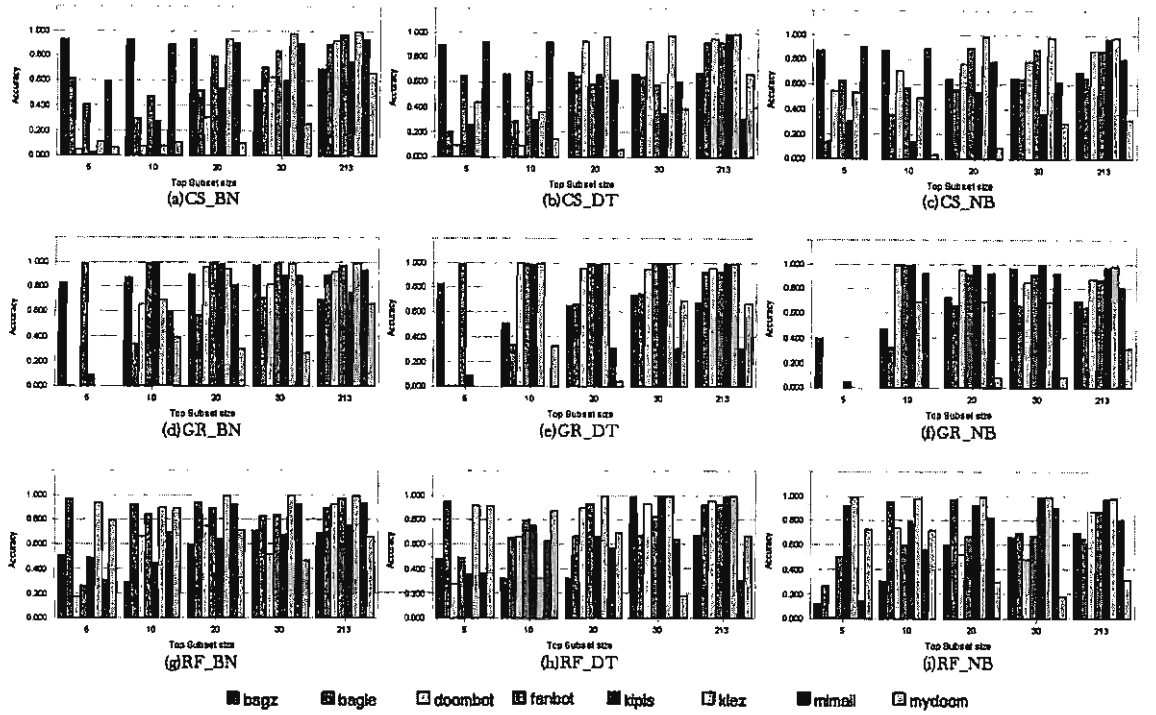

図 4 ウイルスごとの識別率の平均値：(a), (b), (c) は特徴量ランク付けに CS，クラス分類手法にそれぞれ BN，DT，NB を，(d), (e), (f) は特徴量ランク付けに GR，クラス分類手法にそれぞれ BN，DT，NB を，(g), (h), (i) は特徴量ランク付けに RF，クラス分類手法にそれぞれ BN，DT，NB を適用した場合を表す．横軸は高いランクのものから選択した特徴量の数を表す．

Fig. 4 Average accuracy of identifying variants for each kind of viruses. (a), (b), (c): Applying CS for attribute selection and BN, DT, NB for classification, respectively. (d), (e), (f): Applying GR for attribute selection and BN, DT, NB for classification, respectively. (g), (h), (i): Applying RF for attribute selection and BN, DT, NB for classification, respectively. The horizontal axis represents numbers of attributes selected from the highest.

ラス分類手法に Bayesian Networks(BN), Decision Tree(DT), Naive Bayes(NB) を用いる．(5) 最後に，クラス分類器にテストデータを入力し，正常種類を含む 9 種類に対応する 9 つのクラスのどれに分類されるかを調べる．これを，テストデータ 24 個それぞれに対して行い，正しく識別された割合を測定する．

5. 実験結果とその考察

実験結果を図 4 に示す．(a), (b), (c) は特徴量ランク付けに CS，クラス分類手法に BN，DT，NB を，(d), (e), (f) は特徴量ランク付けに GR，クラス分類手法に BN，DT，NB を，(g), (h), (i) は特徴量ランク付けに RF，クラス分類手法に BN，DT，NB を適用した場合を表す．図の縦軸は，8 種類のウイルスの各種類ごとに 3 個の亜種が正しく識別された割合（識別率）を求め，その種類ごとに平均をとった（3 で割った）ものを表す．横軸は高いランクのものから選択した特徴量の数を表す．

はじめに，ウイルス全体に対しウイルスの識別率が特徴量ランク付け手法とクラス分類手法を変えるとどう変わるかを見るため，図 4 から図 5，6，7，8，9，10 を作成した．特徴量ランク付け手法に着目した場合が図 5，6，7，クラス分類手法に着目した場合が図 8，9，10 である．図の縦軸はウイルスの種類ごとの識別率の平均値を表す．図 5，6，7，8，9，10 から次のことが分かる．

(1)8 割強の精度で識別が可能である．

(2) ほとんどの場合 30 個の特徴量で十分である．

(3) 特徴量ランク付け手法を固定すると，クラス分類手法を変えてもあまり変わらない．

(4) クラス分類手法を固定すると，ランク付け手法により大きな違いがある．

(3) と (4) について考察する．特徴量はクラスとの関連度合によってランク付けされるが，この場合多くのクラスと関係が

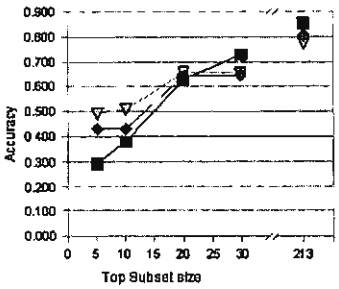

図 5　識別率の平均値:CS を用いた場合
Fig. 5　Average accuracy of identification in case of using CS.

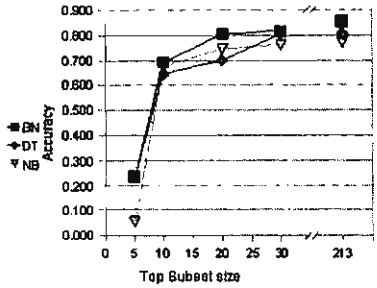

図 6　識別率の平均値:GR を用いた場合
Fig. 6　Average accuracy of identification in case of using GR.

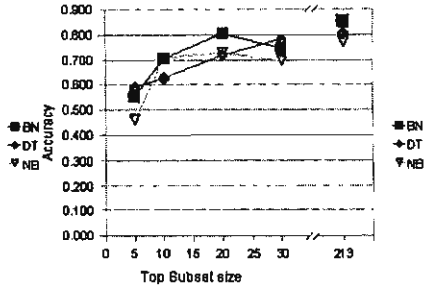

図 7　識別率の平均値:RF を用いた場合
Fig. 7　Average accuracy of identification in case of using RF.

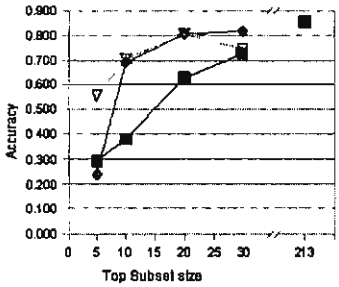

図 8　識別率の平均値:BN を用いた場合
Fig. 8　Average accuracy of identification in case of using BN.

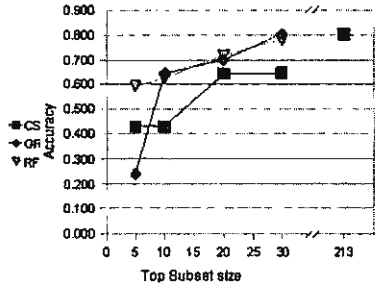

図 9　識別率の平均値:DT を用いた場合
Fig. 9　Average accuracy of identification in case of using DT.

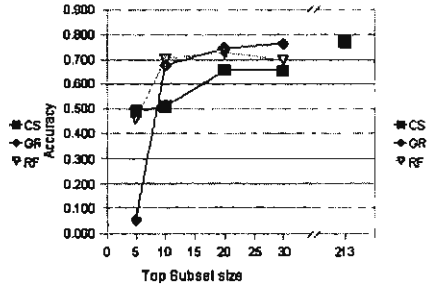

図 10　識別率の平均値:NB を用いた場合
Fig. 10　Average accuracy of identification in case of using NB.

ある特徴量が上位になると推定される．よって，あるクラスを見分けるのには有効な特徴量でも，他のクラスとは関連が低い特徴量は特徴量選択に残らないと考えられる．また逆に，あるクラスに対しては有効な特徴量も他のクラスでは有効でないということも考えられる．したがって，特徴量ランク付け手法が適切であるかないかは精度に大きく影響を与える．これに対しクラス分類は与えられた特徴量ランク付けを基に分類を行うので，特徴量ランク付け手法が適切であれば，分類手法の優劣はさほど精度に影響を与えない．

クラス分類にかかる時間は，クラス分類に使う特徴量の数にほぼ比例している（図 11）．その時間は DT が最も大きく，NB が最も小さい．また全部の特徴量を使った場合の精度は NB が最も低く，BN が最も高い．BN は精度がよく，かつ分類にかかる時間も中程度ということが分かる．

次にウイルスの種類によって識別の結果がどのように異なるかをみるため図 12，13，14 を作成した．図は，精度に大きく影響を与える特徴量ランク付け手法ごとにまとめた．図 12，13，14 はそれぞれ CS，GR，RF に着目した場合を表す．

図 12 から次のことが分かる．多くのものは特徴量を多く用いると精度があがる．しかし Bagz は逆に少ない特徴量で精度が高い．また Mimail は特徴量の数を増やすほど精度が下がる傾向がある．Mydoom はクラス分類手法によらず精度がよくない．

図 13 から次のことが分かる．CS に比べ，少ない特徴量（10 個以上）で良い精度が得られる．特に Doombot, Fanbot, Kipis, Klez はほとんど 100%に近い精度である．Mydoom はクラス分類手法によらず精度がよくない．

図 14 から次のことが分かる．CS に比べ，少ない特徴量（10 個以上）で良い精度が得られる．特に Klez はほとんど 100%に近い精度である．Mydoom は他の特徴量選択手法に比べ精度がよい．

Bagz は CS と GR では RF を用いる場合より精度がよい．Bagle は逆に RF を用いる方が CS と GR を用いるより精度よい．Doomobot と Fanbot は GR を用いる場合が最も精度がよい．Kipis は GR と RF を用いるとよい精度が得られる傾向がある．Klez は全体に精度がよい．Mimail は BN または NB をクラス分けに用いるとよい精度が得られる傾向がある．Mydoom は RF のときを除き精度が悪い．

以上のようにウイルスの種類が異なると識別率もまた変化する．またその変化の様相に特に規則性は見られない．このことについて考察する．第 2. 章で述べたように用いたウイルスの亜種サンプルは確認された時期がそれぞれ異なる．たとえば Klez の亜種サンプル KlezA，KlezB，KlezC は確認された時期が近接している．また Mydoom の亜種サンプル MydoomA，MydoomM，MydoomQ は確認された時期が大きく離れている．表 1 はウイルスの各種類の亜種サンプルの末尾に付けたアルファベットの違いをまとめて示したものである．

クラス分類の精度がよいウイルスは，用いたサンプルが確認

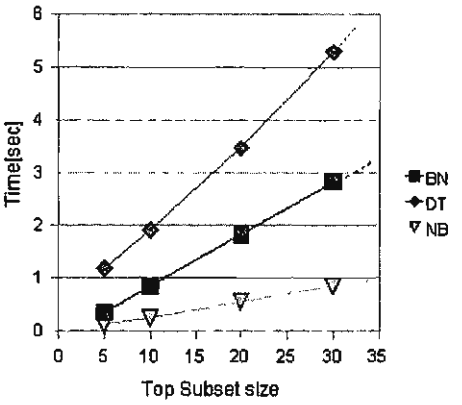

図 11　クラス分類にかかる時間

Fig. 11　Processing time required for classification

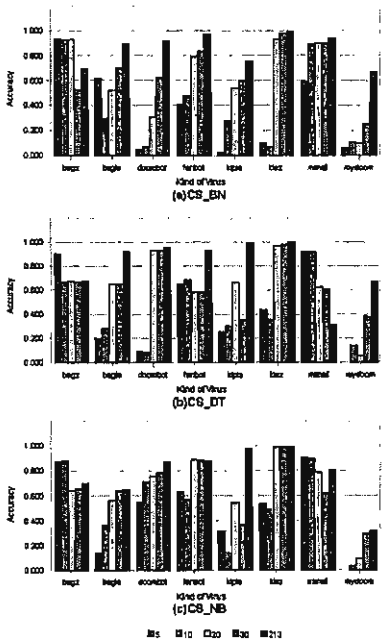

図 12　CS に着目したときのウイルスの種類ごとの識別結果

Fig. 12　Accuracy of identification of virus variants when using CS.

された時期が比較的近接している傾向がある．確認時期が近い亜種はふるまいも近く，ふるまいの近い亜種は一つのクラスに分類しやすいため識別率も上がると考えられる．一方クラス分類の精度が悪いウイルスは，用いたサンプルが確認された時期が比較的離れている傾向がある．確認時期が離れている亜種はふるまいも異なり，ふるまいの異なる亜種は一つのクラスに分類しにくいため識別率は下がると考えられる．

同じ種類の亜種としてラベル付けされたウイルスでも，ふるまいが大きく異なればその対策も異なってくると考えられる．本実験で用いたサンプルについては [4] に基づくラベル付けに問題があると考えられる．このラベル付けは，本実験のふるまいによる分類とは合わない可能性がある．今後は，同じ亜種でも確認された時期に基づき異なるラベル付けをするか，教師なし学習により亜種を分類することが考えられる．

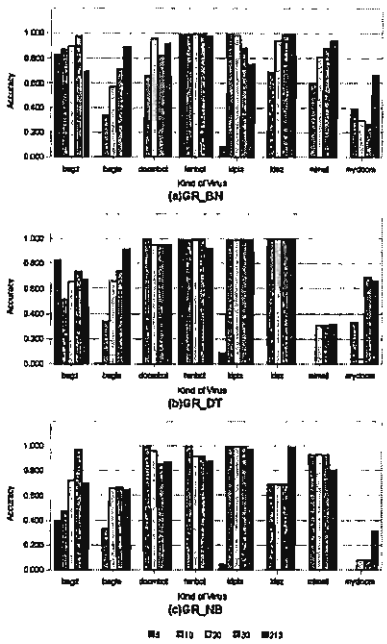

図 13　GR に着目したときのウイルスの種類ごとの識別結果

Fig. 13　Accuracy of identification of virus variants when using GR.

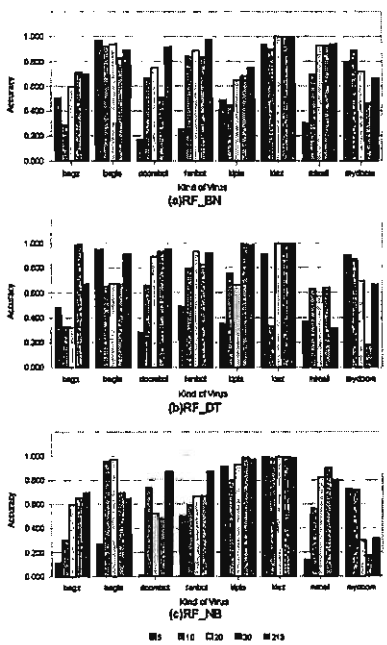

図 14　RF に着目したときのウイルスの種類ごとの識別結果

Fig. 14　Accuracy of identification of virus variants when using RF.

表 1　ウイルスの各種類の亜種サンプルの末尾に付けたアルファベット

Table 1　Distribution of alphabets attached to samples used for virus variants.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Bagz			○	○		○											
Bagle	○		○		○												
Fanbot		○				○	○										
Doombot	○		○	○													
Kipis	○	○					○										
Klez	○	○	○														
Mimail	○	○	○														
Mydoom	○												○				○

6. おわりに

データマイニングを適用したウイルス検出手法により，未知の亜種ウイルスが 8 割強の割合で正しく識別されることが分かった．ウイルスの種類が異なると識別率も異なるが，確認時期が近い亜種の識別率は高い傾向にあることが分かった．ウイルスのラベル付けに関しては，教師あり学習の場合のラベルのつけ方の変更や，教師なし学習でラベルを決定するなどの方法が考えられる．また，未知種類のウイルスの処理や，大量メール送信ウイルス以外のウイルスへの拡張等は今後の課題である．

謝辞　本研究は，一部，日本学術振興会科学研究費補助金 (基盤研究 (C)18500048) による．

文　献

[1] 独立行政法人情報処理機構,"未知ウイルス検出技術に関する調査," 2004.
[2] R. Moskovitch, I. Gus, S. Pluderman, D. Stopel, C. Glezer, Y. Shahar, and Y. Elovici,"Detection of Unknown Computer Worms Activity Based on Computer Behavior using Data Mining,"IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp.169-177, 2007.
[3] VX heavens:http://vx.netlux.org/
[4] Kaspersky:http://www.kaspersky.co.jp/
[5] Symantec:http://www.symantec.com/ja/jp/index.jsp
[6] Weka:http://www.cs.waikato.ac.nz/ml/weka/
[7] M. Bramer,"Principles of Data Mining," Springer-Verlag New York Inc, 2007.
[8] K. Kira and L. Rendell, "The feature selection problem: traditional methods and new algorithm," Proc. the 10th National Conference on Artificial Intelligence, pp.129-134, 1992.
[9] J. R. Quinlan, "C4.5: programs for machine learning," Morgan Kaufmann Publishers Inc., San Francisco, CA. USA, 1993.
[10] Windows Performance Counters: http://msdn.microsoft.com/library/default.asp?url=/library/en-us/counter/counters2_lbfc.asp/