社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

メール特徴を用いたウィルスメール検知に関する一考察

宮本　大輔† 　飯村　卓司† 　門林　雄基†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0101 奈良県生駒市高山町 8916–5
E-mail: †{daisu-mi,takuji-i,youki-k}@is.naist.jp

あらまし　本論文では，新しいウィルスメール検知システムについて議論する．近年のウィルスは多様化・複雑化しており，ウィルス検知システムがウィルスを検知することが困難になってきている．先行研究では，新井氏はウィルスそのものではなく主たるウィルスの感染媒体であるメールに着目し，メールから特徴を抽出することによって添付されているファイルがウィルスか非ウィルスかを分類できるとしている [1]．本研究では，独自に採取した 4130 通のウィルスメールと 2504 通の非ウィルスメールを用いて，メール特徴を用いたウィルスの検知が行えるかを否かを考察する．まず，採取したメールからシグネチャを作成し，シグネチャとのパターンマッチに基づくウィルス検知システムについて調査を行い，次に，確率機械学習のアルゴリズムである AdaBoost を用いたウィルス検知システムについて考察した．最後に，この二つのシステムを組み合わせたハイブリッドなシステムを提案した．そして，この 3 つのシステムについて評価を行い，ハイブリッドなシステムを用いることにより，3 つの検知システムの中で最も検知率が高く，ウィルスメールを非ウィルスメールとして誤検知する確率も最小であることがわかり，有効性が示せた．
キーワード　ウィルス検知，メール，確率機械学習，フィルタリング

A Consideration for Detecting Virus Email by Using Mail Form Information

Daisuke MIYAMOTO†, Takuji IIMURA†, and Youki KADOBAYASHI†

† Graduate School of Information Science, Nara Institute of Science and Technology, Takayama 8916–5, Ikoma, Nara, 630–0101 Japan
E-mail: †{daisu-mi,takuji-i,youki-k}@is.naist.jp

Abstract　In this paper, we argue that a novel virus mail detection system. Recently, detecting virus programs is difficult because they have been complicated and been diversified in order to hinder exiting anti-virus solutions. According to [1], e-mails can be classfied into virus mails or non virus mails by using Mail Form Information, instead of analyzing attached virus programs. We collect 4130 virus mails and 2504 non virus mails, and consider a signature-based virus detection method with Mail Form Information. We also explore a way of decreasing the false negative errors, and consider an AdaBoost-based virus detection method. Finally, we present a hybrid detection method which employs both the signature-based detection method and the AdaBoost-based detection method. Our preliminary results show that true positive and total accuracy can be improved by comparing with signature-based detection.
Key words　Virus Detection, Email, AdaBoost, Filtering

1. はじめに

近年，コンピュータウィルス (以下単にウィルス）が猛威を震っている．コンピュータがウィルスに感染すると，このコンピュータのシステムが破壊されたり，コンピュータを悪意の第三者に犯罪目的で利用されたり，あるいはコンピュータを使うユーザの機密情報が漏洩するなど，様々な被害が生じる可能性がある．IPA の報告によると，2007 年 4 月には 3199 件のウィルス感染事例が届け出られており，そのうち 97.0% がメール経由での感染であった [2]．このため，メールに添付されているウィルスの存在を検知し，ウィルスメールと普通のメールを判別する技術の確立は急務である．

これまで，ウィルスメールへの対策として，ウィルス対策ソフトやメールフィルタが用いられていた．代表的なウィルス対策技術としては，ウィルスを特定可能な情報をシグネチャとして保持し，メールに添付されているウィルスをパターンマッチにより発見するという方式が挙げられる．しかし，このようなウィルス対策技術に対し，ウィルス作成者はウィルスメール検知を難しくする対抗措置として，ウィルスの多様化が行われるようになった．ウィルスの多様化においては，既存の対策技術のシグネチャには一致しないウィルスの亜種を作成することにより，パターンマッチによる検出を難しくしている．また，ウィルスを複雑化する技術も開発されている．例えば，ポリモーフィック型と呼ばれるウィルスは，ウィルスの本体である攻撃を行う機械コードを暗号化してある状態のプログラムである．ウィルス対策技術では暗号化してある機械コードを読み解いて実行することは困難であるが，プログラムとして動作する際には自動的に復号化が行われ，攻撃を行う機械コードが実行される．また，メタモーフィック型と呼ばれるウィルスの場合，ウィルス本体のコードをプログラム実行時にコンパイルし，ウィルスプログラムを自動的に書き換える．このようなウィルス作成者の対抗措置に対し，既存のメールに添付されているファイルをパターンマッチによりウィルスメール検知を行う方法が困難になりつつある．

新井氏の提案するメール特徴を用いたウィルスメール分類法 [1] は，ウィルスの主たる感染媒体であるメールに着目し，メールに添付されているウィルスそのものではなく，メールのヘッダなどから特徴を抽出し，添付されているファイルがウィルスか否かの分類を行う手法である．他のウィルス対策研究と比較すると，メールに添付されているウィルスがいかに多様化・難読化されていたとしても，このウィルスの検査を行わずにウィルスメールか非ウィルスメールを識別可能であるという長所を持つ．

新井氏は，ウィルスメール及び非ウィルスメールメールの検体から，メールの特徴を抽出していた．この特徴は，メールにおいてウィルスの多様性の影響を受けにくい下記 5 項目の値を特徴として抽出する．

（1） 添付ファイルの EW 値の乱れの有無

プログラムなどのバイナリデータをメールに添付する際に，MIME [3] では添付ファイルを MIME Base64 の形式にエンコードすることが定められている．MIME Base64 の形式にエンコードされたデータは文字列形式となり，この文字列を Encoded-Words (EW) と呼ぶ．MIME では，EW の文字長が 76 文字より大きい場合に，1 行の長さが 76 文字以内となるように EW を複数の行に改行する．ここで，一般的なメールにおいて改行が必要な場合，最終行を除いて 76 文字で改行が行われるのが普通である．しかし，ウィルスメールが送信される際は，この EW に「乱れ」があった，すなわち，最終行以外にも 76 文字以外の文字数で改行された行があったと新井氏は述べている．この観測結果から，EW の乱れに基づいたウィルスメールの発見が可能であるとしている．

（2） EW の文字長を 1000 で整数除算した値

添付ファイルの EW 行に含まれる文字長の合計を 1000 で整数除算した値．

（3） MIME マルチパートの構成

メールヘッダに含まれる Content-type ヘッダ部位を出現順にカンマ区切りで並べた値．

（4） 拡張ヘッダ情報

メールヘッダに含まれる X-Priority, X-MS-Mail-priority, X-Mailer ヘッダの部位を出現順にカンマ区切りで並べた値．

（5） MIME boundary の内容

MIME パートの境界を表す MIME boundary [4] の文字列を，英数文字は全て ”X” に，連続する文字は 1 文字にまとめるという方法によりコード化した値．

このように，各メール毎に抽出した特徴を順番に並べることにより，特徴量と呼ばれる文字列が得られる．新井氏は，この特徴量を用いて分類実験を行い，ウィルスメールと非ウィルスメールが同じ特徴量となることはなく，分類が可能であったと述べている．

しかし，新井氏の論文では，ウィルスメールと非ウィルスメールの分類が可能ということは示されているが，例えば全く新しいメールが 1 通来たときに，これがウィルスメールに分類されるのか，非ウィルスメールに分類されるのかという実験は行われていない．また，ウィルスメールにおける特徴量の分布と，非ウィルスメールにおける特徴量の分布についての違いも考察されていない．このため，この未判別のメールの特徴量が，ウィルスの特徴量に近いのか，非ウィルスのメールの特徴量に近いのかを判断できるか否かは明かではない．

そこで，本論文では，ウィルスメール 4130 通，添付ファイル付きの非ウィルスメール 2508 通に対して特徴量抽出を試み，ウィルスメールの特徴量の分布と非ウィルスメールの特徴量の違いについて独立性の検定を行った．検定に用いた検定統計量はクラメールの連関計数であり，その値は 0.607 となったことから，ウィルスメールの分布と非ウィルスメールの分布は異なると判断できる．また，ウィルスの多様性の影響を受けやすいと思われる EW の文字長を除いて特徴量を抽出したところ，連関計数は 0.761 となり，よりウィルスメールと非ウィルスメールの分布の違いが浮き彫りとなり，EW の文字長を 1000 で整数除算した値をウィルスメールの特徴として使用しない方が，ウィルスメールの検知に良い結果が得られるのではないかと推測した．

次に，この 4 つの特徴量を用いたウィルスメールの検知方式を考える．まず，標本としたウィルスメールの半分である 2065 通と，非ウィルスメール 1254 通を非復元的に無作為抽出した 3319 通のメールを学習用データセットとして用い，特徴量のシグネチャを作成した．そして，残りの 3319 通のメールを検証用データセットとして用い，学習用データセットのシグネチャによりメールがウィルスか非ウィルスかの分類を行った．このうち，2859 通のメールが，シグネチャを基にウィルスメールもしくは非ウィルスメールに分類可能であった．また，ウィルスメールを非ウィルスメールとして分類したり，非ウィルスメールをウィルスメールとして分類することはなかった．

一般的なシグネチャ型の検知では，ウィルスの特徴をシグネチャとして保持し，この特徴に一致しないメールについてはウィルスとして検知されなかったメール，すなわち通常のメールであるとして取り扱われる．仮に残りの 460 通のメールについて全て非ウィルスメールであるとした場合，ウィルスメールをウィルスメールであるとして識別した True Positive(TP) は 98.21%，非ウィルスメールをウィルスメールとして識別した False Positive(FP) は 0%，全体的な正解率は 98.89% となった．ただしこの方式では，判定できなかったメールを全て非ウィルスメールとして扱っており，判定できなかったメールについてさらなる検知を行い，より多くのウィルスメールを検知する方法が必要であると考えた．

このため，本論文では，確率機械学習のアルゴリズムである AdaBoost [5] をウィルスメール検知に適応することを試みた．AdaBoost は，正答率が 5 割以上の仮説を複数個組み合わせ，その複数の仮説において重み付け多数決を行うことにより，正答率の高い仮説を作成するアルゴリズムである．このアルゴリズムを用いるためには，各仮説において判定を行った結果，真であった場合は 1 を，偽であった場合は-1 を返す必要がある．EW の乱れの有無は真偽値であるため適応可能であり，また EW の文字長は上に述べたように仮説として含まない．残りの 3 つに対しては，1 つメールから抽出した特徴に対して同じ特徴を持つメールが何通あるかをまず調べ，その値が定義された閾値と比較し，非ウィルスだと思われるときに 1 を，ウィルスだと思われるときに-1 を返すような仮説として設計した．このように学習用データセットを用いた AdaBoost を用いた学習を行い，その成果であるウィルスメール判定アルゴリズムを検証用データセットに適応した．この結果，TP は 87.46%，FP は 11.16% であることが分かった．この方式はシグネチャ型の方式と比べて精度が低いように思われるが，シグネチャ型の方式で判定不能となったメールに対してのみ AdaBoost を用いたウィルスメール判定を行うことにより，TP は 99.76%，FP は 1.59%，全体的な正解率は 99.25% となることが分かった．また，シグネチャ型では検知できなかった 37 通のウィルスメールのうち 32 通をさらに検知できていることも確認できた．シグネチャ型の方式より FP は増加するが，TP や全体的な正解率は増加していることから，シグネチャ型と AdaBoost を用いた方式を組み合わせることで，ウィルスメールの判定の精度は上昇していると考えることができる．

以下， 2. 節において，ウィルスメールと非ウィルスメールの特徴量の分布に関する考察を行い， 3. 節では特徴量を用いたウィルスメール検知と AdaBoost を用いたウィルスメール検知，二つを組み合わせたハイブリッドなウィルスメール検知について述べる． 4. 節では，ウィルスメールらしさを決めるための閾値について考察を行い， 5. 節に結論と今後の課題について述べる．

2. ウィルスメールと非ウィルスメールの特徴量の分布

本節では，ウィルスメールと非ウィルスメールの特徴量の分布がどのように異なるのかを説明する．まず，新井氏の論文を参考として実装した特徴量抽出を行うプログラムの説明を行う．次に，このプログラムを基にメールの分類を行い，ウィルスメールと非ウィルスメールの特徴量の分布について解析し，最後に，特徴とすべき項目について考察を行う．

2.1 特徴量抽出プログラムの実装

本論文の実装は，基本的に新井氏の論文を参考に実装している．しかし，新井氏の論文では述べられていなかった点もあるため，本論文の実装について補足的な説明を行う．

- EW の乱れの有無について

新井氏は EW を示す部位に 76 文字以上の行がある場合に，このメールには乱れがあるとしている．非ウィルスメールに関しても EW の値は 76 文字となることが多く，76 より大きい値であった場合に乱れがあるとする．

- 拡張ヘッダ情報について

新井氏は拡張ヘッダである `X-Priority`, `X-MS-Mail-Priority`, `X-Mailer` ヘッダを出現順ごとにカンマ区切りで記録している．検体からは，は `X-MSMail-Priority` という文字列も見受けられたため，本論文ではこの値も拡張ヘッダ情報として取得対象とする．

また，今回の検体ではメールそのものにウィルスが添付されているわけではなく，メールにウィルスメールが添付されている場合があった．このため，メールに添付されているファイルがメールである場合，その添付されている方のメールの特徴量を抽出することにした．

なお，分類プログラムは未公開であったため，今回は Perl で新たに分類プログラムの実装を行った．その行数は 426 行程度であった．

2.2 ウィルスメールと非ウィルスメールの特徴量の分布

次に，実装したプログラムを用い，ウィルスメールと非ウィルスメールの検体から特徴量を抽出した．検体として用いたデータセットは，著者らが所属している奈良先端科学技術大学院大学で 2006 年 9 月から 2007 年 1 月までの期間に収集したウィルスメール 4130 通と，著者らが仕事用，個人用途で使っているメールアドレスで 2006 年 4 月から 2007 年 3 月までの期間に収集した非ウィルスメール 22256 通のうち，添付ファイルが含まれていた 2508 通の，合計 6638 通のメールからなる．なお，奈良先端大では，メール配送時にウィルスが含まれているか否かをアンチウィルスソフトにより検査する．今回検体としたウィルスメールは，この際にウィルスが検出されたメールである．また，非ウィルスメールも同様に，アンチウィルスソフトによりウィルスが検出されなかったことを確認している．

分類は，一つの特徴量に対し，どの程度のメールが分類されるかを調べることによって行う．新井氏の論文では，ウィルスが多様性を持っていたとしても，特徴量によりウィルスの分類が可能であるとしている．そこで，一つの特徴量に対して何通のメールが分類されるのか，またウィルスメールと非ウィルスメールではこれらの分類はどのように異なってくるかを検証する．

まず，単純にウィルスメール及び非ウィルスメール全てに分

類を行った所，ウィルスメールは 299 種類に，非ウィルスメールは 1179 種類の特徴量に分類された．なお，ウィルスメールの特徴量と非ウィルスメールの特徴量が合致するケースは見られず，非ウィルスメールを誤判定によりウィルスメールとして判定する事象は観測されなかった．また，ウィルスメールを分類した結果，98.37 % において同じ特徴量の中には同じウィルスしか分類されず，添付されているウィルスが異なるにもかかわらずメールの特徴量が同じという事例は 1.63% でしかなかった．なお，新井氏の論文においても 98.91 % という正分類率が示されているが，この結果はそれと同じ傾向であると考えられる．

次に，ウィルスメールの分布と非ウィルスメールの分布に果たして差があるのか否かを調査する．まず，ウィルスメールを分類し，1 通のメールしか分類されなかった特徴量，2 通のメールが分類された特徴量，というように 299 種の特徴量毎に分類されたメールの数を調べた結果を図 1 に示す．分布座標は，x 通のメールが分類された特徴量において，その場合の種類の数 y を示している．例えば，ウィルスメールの分類結果において，1 通のメールが分類された特徴量は 660 種類，2 通のメールが分類された特徴量は 233 種類であるため，座標 (x, y) が (1,660), (2,233) の点にプロットしている．なお，グラフの可読性を高めるため，x 軸，y 軸ともに対数スケールを用いた．

同様に，非ウィルスメールについても同様の調査を行い，その分布を図 2 に示す．これらのグラフは，どちらも x が増えると y が下がるという右肩下がりの傾向を示しており，二つの分布に差があるとはこれだけでは判断できない．

そこで本論文では，パラメータの相関を調べるため，度数分布表を用いることにした．前述のグラフの x 軸の変数とした「同じ特徴量を持つメールの数」を「メールの特異性」であると近似する．例えば，これまで行ってきた特徴量による分類を行った結果，ある特徴量には 1 件のメールしか分類されなかったとする．この場合，このメールは他のメールの持たない特徴量を持つと判断し,「メールの特異性が高い」とみなす．反対に，1000 通のメールと同様の特徴量を持つメールに関しては,「メールの特異性が低い」とみなす．ここで，前述の x において，$log(x) < 1$ となる場合に「特異性が極めて高い」，$1 \leqq log(x) < 2$ となる場合に「特異性が高い」，$2 \leqq log(x) < 3$ となる場合に「特異性がやや高い」，$3 \leqq log(x) < 4$ となる場合に「特異性は高いとも低いとも言えない」，$4 \leqq log(x) < 5$ となる場合に「特異性はやや低い」，$5 \leqq log(x) < 6$ となる場合に「特異性は低い」，$6 \leqq log(x)$ となる場合に「特異性は極めて低い」という階級を定義した時の度数分布は表 1 のようになった．度数分布から分かるとおり，ウィルスメールは特異性が低い階級に多く分布し，非ウィルスメールは特異性が高い階級に多く分布しているように見える．そこで，二つの分布が異なる分布であるか否かを調べるため，クラメールの連関計数を用いて独立性の検定を行った．クラメールの連関係数は 0 から 1 までの値を取る検定統計量であり，0 に近ければ近いほど分布は似通っており，1 に近ければ近いほど分布は異なると判断できる．ウィルスメールと非ウィルスメールの度数分布に対して連関計数を測定した所，値は 0.607 であり，ウィルスメールと非ウィルスメールの分布は異なると推測するのが妥当だということが分かった．

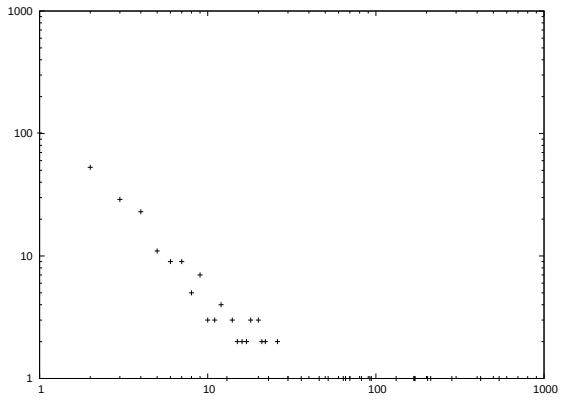

図 1　ウィルスメールの分布

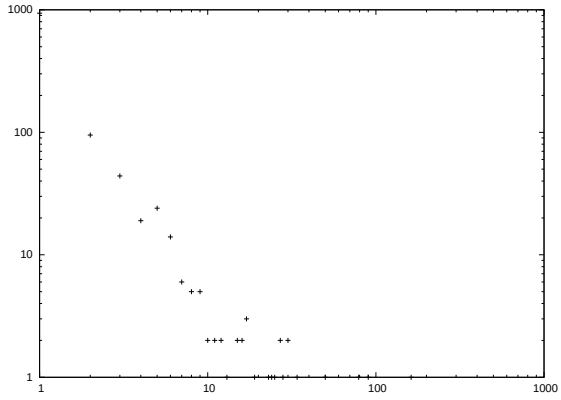

図 2　非ウィルスメールの分布

表 1　度 数 分 布

検体	ウィルスメール	非ウィルスメール
$log(x) < 1$	207	1129
$1 \leqq log(x) < 2$	351	454
$2 \leqq log(x) < 3$	479	296
$3 \leqq log(x) < 4$	325	298
$4 \leqq log(x) < 5$	600	169
$5 \leqq log(x) < 6$	1207	162
$6 \leqq log(x)$	961	0
合計	4130	2508

2.3　メール特徴の妥当性に関する考察

新井氏の論文では，抽出すべきウィルスメールの特徴として，ウィルスの多様性による影響を受けないことを前提とした個所を選択しているが，その特徴量に EW の文字長を 1000 で整数除算した値をメールの特徴として含めるよう設計している．EW の文字長とは，添付されたウィルスを MIME Base64 形式でエンコードした時の文字列の長さの事であり，多様性を持ったウィルスであればこの長さは可変長である．近年のウィルスの複雑化を鑑みるに，添付されたウィルスの大きさが常に 1000 バイト以内の変化しかしないとは考えにくい．

このため，EW の文字長という特徴量を使用しないメールの

表 2　EW の文字長を考慮しない場合の度数分布

検体	ウィルスメール	非ウィルスメール
$log(x) < 1$	50	886
$1 \leqq log(x) < 2$	106	453
$2 \leqq log(x) < 3$	116	406
$3 \leqq log(x) < 4$	130	226
$4 \leqq log(x) < 5$	462	375
$5 \leqq log(x) < 6$	1220	162
$6 \leqq log(x)$	2046	0
合計	4130	2508

分類を試みた所，ウィルスメールが非ウィルスメールと同一の特徴量を持つことはなかった．また EW の文字長を用いた場合，用いなかった場合の両方において，異なるウィルスであるにもかかわらず同じウィルスであるとして分類されるメールの数は変化しなかった．また，この分類結果の分布を，表 1 と同様な度数分布を用いて表 2 に示す．この分布においてクラメールの連関係数は 0.761 であり，EW の文字長を考慮した場合の度数分布と比較すると，よりウィルスメールと非ウィルスメールの違いが浮き彫りになる事が分かった．

これらの結果により，ウィルスを分類する上で，EW の文字長を考慮せず，残りの 4 項目の特徴から特徴量を形成し，ウィルスメール検知を行う手法について考察する．

3.　メールの特徴量を用いたウィルスメール検知

本節では，メールの特徴量を用いた単純なウィルスメール検知方法にのべ，問題点を解決するため機械学習の手法を用いたウィルスメール検知手法について述べる．

3.1　シグネチャによるウィルスメール検知手法

新井氏の論文及び本論文の 2. 節において，メールの特徴量を用いたウィルスメール及び非ウィルスメールの分類をすると，ウィルスメールと非ウィルスメールは同じ特徴量となることはなく，またその分布も異なることがわかった．このため，例えば添付ファイルを持つ新しいメールが 1 通来た時に，このメールの特徴量を抽出し，これがウィルスメールの特徴量に一致するか，非ウィルスメールの特徴量に一致するかを判定することによって，ウィルスメール検知ができると考えられる．

これを確認するため，検体の半分をウィルスメール検知アルゴリズムを形成するための学習用データセットとして用いてシグネチャを作成し，残りの半分を検証用データセットとして用いて，作成したシグネチャの精度を測るために用いることにした．なお，本論文では学習用データセットを，ウィルスメール 4130 通の半分となる 2065 通を，非ウィルスメール 2508 通についても同様に 1254 通を非復元的な無作為抽出によって取り出し，計 3319 通のメールによって構成した．また，検証用データセットについては，残りの無作為対象の抽出とならなかった残りの 3319 通のメールから構成した．この学習用データセットを用いて，ウィルスメール検知アルゴリズムを作成した．学習用データセットでは教師付き学習，すなわちウィルスメールと非ウィルスメールの分類がすでになされている状態を前提とする．この前提のもとに，学習用データセットに対して 2. 節で述べた実装によって特徴量を抽出し，分類を行う．検証用データセットからは，1 通毎に特徴量を抽出し，これが学習用データセットで抽出したウィルスメールの特徴量群に一致するか，非ウィルスメールの特徴量群に一致するかを調べる．検査用データセットの特徴量が，ウィルスメール，非ウィルスメールの特徴量群のどれにも一致するものがなかった時は，判定不能であるとする．このようにして，本論文ではウィルスメールの判定を行った．

表 3　メールの判定結果

検体	ウィルスメール	非ウィルスメール	合計
ウィルスと判定	2028	0	2028
非ウィルスと判定	0	831	831
判定不能	37	423	460
合計	2065	1254	3319

まず学習用データセットを分類した結果，ウィルスメールは 70 種類の特徴量に，非ウィルスメールは 549 種類の特徴量にそれぞれ分類された．次に検査用データセットにも同様に特徴量を抽出し，その特徴量が学習用データセットから得られたウィルスメール，非ウィルスメールの特徴量群のどちらに類するかを調べ，ウィルスメールの特徴量群に一致すれば，このメールをウィルスメールとして検知する．同様に，非ウィルスメールの特徴量群に一致すれば，このメールは非ウィルスメールとして検知する．なお，検査用データセットから抽出した特徴量が，両方の特徴量群のいずれにも一致しない場合，このメールは判定不能であるとして，別に集計することとした．

表 3 は，検証用データセットに対し検知を行った結果である．判定可能であった 2859 通のメールに関しては，ウィルスメールを非ウィルスメールであると誤検知したり，非ウィルスメールをウィルスメールであると誤検知することは観測されなかった．ただし，検知不能となったメールは全体で 13.86% であり，特に非ウィルスメールに関しては 33.73% のメールが判別不能であった．これは，非ウィルスメールの方が，個々のメールの特徴量における特異性が高い事ことに起因していると考えられる．

一般的なシグネチャ型の検知では，ウィルスの特徴をシグネチャとして保持し，この特徴に一致しないメールについてはウィルスとして検知されなかった，すなわち非ウィルスであるとして取り扱われる．これは，例えば非ウィルスのメール，非ウィルスの添付ファイルは特異性が高く，これらのシグネチャを保持することは，特異性の低いウィルスのメール，ウィルスの添付ファイルのシグネチャを保持することよりも大変だからだと考えられる．

そこで，ウィルスメールとして検知した 2028 通以外の残り 1281 通のメールを，ウィルスが含まれていると検知できなかったため，非ウィルスメールとして取り扱った場合を考える．判定結果は表 4 の通りとなり，37 通のウィルスメールに対する False Negative(FN) が発生した．この方式の TP は 98.21%，FP は 0.00 %，全体的な正解率は 98.89% であった．

表 4　シグネチャ型の検知による判定結果

検体	ウィルスメール	非ウィルスメール	合計
ウィルスと判定	2028	0	2028
非ウィルスと判定	37	1254	1291
合計	2065	1254	3319

3.2　AdaBoost を用いたウィルスメール検知

シグネチャ型のウィルスメール検知では，37 通のウィルスメールが分類できなかったため，非ウィルスであると判定されてしまった．本節からは，この 37 通のウィルスメールを，なるべく FP を増やさないようにして 0 通に近づける方法について考える．

このため，本論文では確率機械学習のアルゴリズムである AdaBoost を用いたウィルスメール検知について考察する．AdaBoost は，正答率が 5 割強の仮説を複数個組み合わせ，その複数の仮説において重み付け多数決を行うことにより，正答率の高い仮説を作成するアルゴリズムである．

そこで，本論文においてメールの特徴として考えている EW の乱れ，MIME マルチパートの構成，拡張ヘッダ情報，MIME boundary の内容の 4 項目から，メールがウィルスメールであるか非ウィルスメールであるか否かの判定を行う仮説を作成する．なお，AdaBoost の仮説は，仮説による判定結果を真偽値で返す必要がある．ここでは，メールがウィルスだと判定したら-1(偽) を，非ウィルスだと判定したら 1(真) を返すこととし，それぞれ特徴量について以下のような真偽値を設定した．

- H_1: EW 値に乱れは無いか

EW 値に乱れが無ければ 1 を，乱れがあれば -1 を返す

- H_2: MIME マルチパートの構成は特異性が高いか

MIME マルチパートの構成の特異性がある閾値より高ければ 1 を，低ければ -1 を返す．

- H_3: 拡張ヘッダ情報は特異性が高いか

拡張ヘッダ情報の特異性がある閾値より高ければ 1 を，低ければ -1 を返す．

- H_4: MIME boundary の内容は特異性が高いか

MIME boundary の特異性がある閾値より高ければ 1 を，低ければ -1 を返す．

仮説 H_2, H_3, H_4 は，非ウィルスメールであれば特徴量の特異性が高く，ウィルスメールであれば特異性が低くなるというこれまでの観測結果に基づいている．

特異性が高いか低いかを判断する指標としては， 2 節に述べた手法を用いる．表 5, 6, 7 に，各メールの特徴量の分布を示す．例えば，同じ MIME マルチパートの構成を持つメールが他にない場合，log(1)(= 0) であり，このメールの特異性は極めて高いと考える．また，1193 通のメールが 1 つの同じ特徴量を共有していたが，この場合は，log(1193)(≒ 7.08)，特異性が極めて低いと考える．ここでは，同じ特徴を共有するメールの数 x が $\log(x) \geqq 4$ から特異性が低くなるとする．なお，この閾値の決定手法は暫定的なものであり， 4. 節においてさらに考察する．

次に，このようにして作成した仮説を用いて， 3.1 節において作成した学習用データセット対し AdaBoost による機械学習を行った．その結果，得られたウィルスメールの検知アルゴリズムを式 1 に示す．計算によって導き出される H_{ADA} が正の値であれば非ウィルスメール，そうでなければウィルスメールであると判断することができる．

表 5　MIME マルチパートの構成の度数分布

検体	ウィルスメール	非ウィルスメール	全体
$log(x) < 1$	23	235	258
$1 \leqq log(x) < 2$	33	174	207
$2 \leqq log(x) < 3$	45	336	381
$3 \leqq log(x) < 4$	181	231	411
$4 \leqq log(x) < 5$	291	279	570
$5 \leqq log(x) < 6$	299	0	299
$6 \leqq log(x)$	1193	0	1193

表 6　拡張ヘッダの情報の度数分布

検体	ウィルスメール	非ウィルスメール	全体
$log(x) < 1$	21	103	124
$1 \leqq log(x) < 2$	31	100	131
$2 \leqq log(x) < 3$	0	175	175
$3 \leqq log(x) < 4$	5	281	286
$4 \leqq log(x) < 5$	95	249	344
$5 \leqq log(x) < 6$	0	0	0
$6 \leqq log(x)$	1913	346	2259

表 7　MIME boundary の度数分布

検体	ウィルスメール	非ウィルスメール	全体
$log(x) < 1$	5	37	42
$1 \leqq log(x) < 2$	0	27	27
$2 \leqq log(x) < 3$	8	82	90
$3 \leqq log(x) < 4$	48	29	77
$4 \leqq log(x) < 5$	52	255	307
$5 \leqq log(x) < 6$	185	288	473
$6 \leqq log(x)$	1767	536	2303

$$H_{ADA} = 0.20 \cdot H_1 + 0.80 \cdot H_2 + 0.41 \cdot H_3 + 0 \cdot H_4 \quad (1)$$

AdaBoost では正解率の高い仮説から順番に学習させていき，仮説の重みを算出していく．この例では，仮説 H_2 の次に H_3, H_4, H_1 の順番で仮説を学習用データセットに適応することとなった．ここで，1 番最初に適応した 仮説 H_2 が正しく判定できなかったメールを，2 番目に適応した 仮説 H_3 が正しく判定できた場合，この 仮説 H_3 の重みは高くなる．反対に，仮説 H_2 が正しく判定できたメールを 仮説 H_3 が正しく判定できたとしても，仮説 H_3 の重みはそれほど高くならない．仮説 H_4 の重みが 0 となっているのはこのためである．

次に，検証用データセットの各メールにおいても各項目毎に特徴を抽出する．この抽出した特徴について，学習用データセットにおいて何通のメールが同じ特徴を保持しているかを調べることにより，特異性を算出し，各仮説毎に真偽値を割り出す．この割り出された真偽値と， 1 のウィルスメールの検知アルゴリズムを用いてウィルスメールの検知を行った結果を表 8 に示す．TP は 86.25%, FP は 5.98%, 全体的な正答率は 89.18%

表 8　AdaBoost を用いた場合のメールの判定結果

検体	ウィルスメール	非ウィルスメール	合計
ウィルスと判定	1781	75	1956
非ウィルスと判定	284	1179	1363
合計	2065	1254	3319

表 9　ハイブリッド方式によるメールの判定結果

検体	ウィルスメール	非ウィルスメール	合計
ウィルスと判定	2060	75	2135
非ウィルスと判定	5	1179	1184
合計	2065	1254	3319

であった． 3.1 節で述べたシグネチャ型のウィルス検知を行う方式では TP が 98.21 %，FP が 0.00% であったことを考えると，AdaBoost を用いた検知では TP が減少し，FP が増加しており，単純に比較すると性能が劣化していると言える．

3.3　ハイブリッド型のウィルスメール検知

これまで， 3.1 節において，シグネチャ型のウィルスメール検知は，判定不能となるウィルスメールがあり，ウィルスメールの見逃しが発生するという問題があることを述べた．また，3.2 節において， AdaBoost を用いたウィルスメールの検知は実現可能であるが，シグネチャ型のウィルスメール検知と比較して，ウィルスメール，非ウィルスメールの検知共に精度が下がっていることが分かった．

そこで，両方式を組み合わせたハイブリッド型のウィルスメール検知について考える．すなわち，まず精度の高いシグネチャ型のウィルスメール検知手法を適応し，ウィルスメールであると判定できなかった残りのメールに対して，AdaBoost を用いた検知手法を適応する．これにより，判定不能となっていたメールについても判定が行われるため，FN が減ぜられると考えられる．

表 9 に，検証用データセットに対し，上記の手法によりウィルスメール検知を行った際の検知結果を示す．3319 通のメール中，2028 通はウィルスメールであると検知可能であり，残りの 1291 通のメールに対して AdaBoost を用いた検知を行った所，さらに 32 通のウィルスを検知できた．これより，ウィルスメールを非ウィルスメールだと誤検知した件数がシグネチャ方式では 37 件あったが，ハイブリッド方式では 5 件に減少している．この方式の TP は 99.76%，FP は 5.98 %，正解率は 97.59% であった．

ここで，FP をさらに減らすためには，非ウィルスメールのシグネチャも併用すれば良い．すなわち，シグネチャ型でウィルスメールであると判定した 2028 通，非ウィルスメールであると判定した 831 通を除いた，残りの 460 通の判定不能となったメールに対してのみ AdaBoost による検知を行い，その結果を表 10 に示す．この方式の TP は 99.76%，FP は 1.59%，正解率は 99.25 % であった．

この結果とシグネチャ型のウィルスメール検知方式を比較すれば，FP こそ増加しているものの，TP や全体的な正解率は増加しており，検知精度は高められる．ウィルスの検知において，TP，FP，全体的な正解率のどれを重視するかはシステム

表 10　非ウィルスメールのシグネチャも用いたハイブリッド方式によるメールの判定結果

検体	ウィルスメール	非ウィルスメール	合計
ウィルスと判定	2060	20	2080
非ウィルスと判定	5	1234	1239
合計	2065	1254	3319

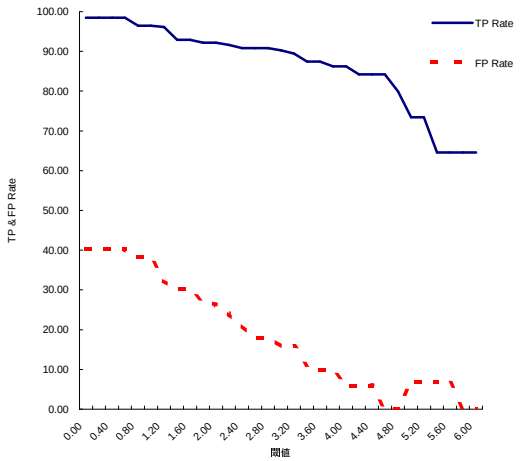

図 3　閾値が TP 及び FP に与える影響

運営のポリシーによって異なると考えられるが，まず TP が高く，次に FP が少ないことが要求されるような場合では，シグネチャ型のウィルス検知よりも有効である．このため，シグネチャ型のウィルス検知に，AdaBoost による検知方式を組み合わせたハイブリッド型検知方式は無価値ではなく，有効性があるということが分かった．

4.　考　　察

本論文において，メールから抽出した特徴量の特異性が高ければ非ウィルスメール，低ければウィルスメールであると考えている．この閾値をどのように決定するかを前もって予測することが，今回の実験では困難であった．図 3 は，$\log(x) > \alpha$ となる閾値 α を 0 から 6 まで 0.2 ずつ値を変えて変化させた時に， 3.2 節で述べた方式の TP と FP について調べ，グラフとしてまとめたものである．X 軸に閾値 α，Y 軸は TP と FP の発生する確率 (単位は%) である．傾向としては，閾値を高くすることによりメールがウィルスメールであると判定される確率が下がり，その結果 TP と FP の両方が下がる．反対に，閾値を低くすることにより，メールがウィルスメールであると判定される確率が上がり，その結果 TP と FP の両方が上がる．ここで，閾値が 4.6 のとき，TP は 84.21%，FP は 0.00% となる．仮にこの閾値を算出できるのであれば， 3.3 節に述べるハイブリッド方式を用いることにより，TP は 99.71%，FP は 0.00%，全体的な正解率は 99.82% となるウィルスメール検知が可能となる．しかし，メールがウィルスメールか非ウィルスメールかの検知を行う前に TP や FP を知ることはできないため，検知前に閾値の 4.6 という数字を決定することは困難である．

最適な閾値を探索する手法としては，ウィルスメールの検知が終わった後に，他のウィルス検知ソフトなどと連携してこのウィルスメール検知結果が正しかったのかどうかを調べ，最適な閾値を自動的に探索する方法が挙げられる．このような事後分析は，学習用データセットを充実する上でも役立てられると考えられ，今後の課題とする．

5. 結　　論

ウィルスの主な感染源はメールである．本研究は，メールに添付されているウィルスそのものを解析するのではなく，メールに含まれている特徴量に基づいてウィルスメールと非ウィルスメールを区分する方式に着目し，これを基づいてシグネチャ型の検査方式について考察した．また，この方式では判定不能となるメールに対して，ウィルスメールを非ウィルスメールであると誤検知する確率をさらに減らすべく，確率機械学習のアルゴリズムである AdaBoost を用いたウィルスメール検知手法を提案した．その上で，シグネチャ型のウィルス検知方式では検知では判定不能となるメールを，AdaBoost を用いたウィルス検知方式に検知させることにより，非ウィルスメールをウィルスメールとして判定する事例が増加するものの，ウィルスメールを非ウィルスメールとして判定する事例が減少し，全体的な正解率も上昇するという成果が得られることが分かった．

今後の課題は，事後学習を行うシステムの開発である．現在のウィルス検知における正解率が高いのか低いのかを調査し，それによって「ウィルスメールである」と疑う特異性の閾値を，最適な TP と FP が得られるように自動的に探索するシステムが求められる．また，このシステムの応用により学習用データセットの更新も可能となり，ひいてはより高い精度のウィルスメール検知が可能となると考えられる．

文　　献

[1] 新井貴之：“メール特徴を用いたウィルスメール分類法”, 2007 年暗号と情報セキュリティシンポジウム (SCIS2007) (2007 年).

[2] IPA: “コンピュータウイルスの届出状況 [2007 年 4 月分] について”, http://www.ipa.go.jp/security/txt/2007/documents/virus-full0705.pdf.

[3] K. Moore: “MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII text”, RFC 2047, Internet Engineering Task Force (1996).

[4] N. Freed and N. Borenstein: “Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types”, RFC 2046, Internet Engineering Task Force (1996).

[5] Y. Freund and R. E. Schapire: “A Short Introduction to Boosting”, Journal of Japanese Society for Artificial Intelligence 14(5), pp. 771–780 (1999).