情報処理学会論文誌 Vol.53 No.9 2086–2093 (Sep. 2012)

# 攻撃通信検知のための合成型機械学習手法の一検討

小久保 博崇[1,†1,a)] 金岡 晃[2] 満保 雅浩[3] 岡本 栄司[2]

受付日 2011年12月2日, 採録日 2012年6月1日

**概要**：マルウェアの脅威は日々拡大しており，いまや社会に実害を及ぼす脅威となっている．また未知のマルウェアの侵入や活動を検出し，被害を防ぐことの重要性が高まっている．本論文ではCCC DATAset2011の攻撃通信データを利用し，通信プロトコルヘッダの特性を，性質の異なる複数の機械学習手法を組み合わせて学習することで未知攻撃を含む攻撃通信の持続的な検知を試みた．決定木の定期的な再学習に加え二次元自己組織化マップ（SOM）による逐次学習を取り入れることで安定して高い精度を保てるように工夫することにより，99%前後の確率で攻撃通信の検知を行うことが可能となった．

**キーワード**：機械学習，ネットワーク攻撃検知，CCC DATASet2011

# A Combined Machine Learning Method for the Detection of Attacks

Hirotaka Kokubo[1,†1,a)] Akira Kanaoka[2] Masahiro Mambo[3] Eiji Okamoto[2]

Received: December 2, 2011, Accepted: June 1, 2012

***Abstract:*** Growing threats of malwares has already caused great damage to the world. It is necessary to detect invasions and activities of unknown malwares, and to prevent damage. In this paper, we combine multiple machine learning methods to achieve sustainable detection of attack communication including unknown attacks. We use the attack communication data of the CCCDATAset2011 for the analysis of the proposed method. As a result, it succeeded in stably detecting in high accuracy.

***Keywords:*** Machine Learning, Network attack detection, CCC DATASet2011

## 1. はじめに

コンピュータウイルスやワーム，トロイの木馬といった悪意のあるソフトウェア（マルウェア）の脅威は日々拡大しており，いまや社会に実害を及ぼす脅威となっている [1]．

マルウェアの侵入や攻撃を検出する手法はミスユース型とアノマリ型に大別できる [2]．商用製品で主流となっているミスユース型手法は，それぞれのマルウェアの特徴を示すシグネチャパターンを持ち，通信などをシグネチャとマッチングさせることでそれらを検出する．シグネチャに合致すれば確実に検知できるため，シグネチャが存在する既知のマルウェアやその通信に対して高い検知能力を持つが，近年のマルウェアは膨大な亜種・変異種が存在するケースも多いため，シグネチャの数も膨大となるという問題がある．また，近年大きな問題となっているゼロデイ攻撃など，未知の攻撃に対しての検知は行うことができない．

これに対して，アノマリ型手法は機械学習や統計的手法などを用いて通常の振舞いと異なる行為を検出することで未知の攻撃などを検知する手法である．未知の攻撃に対応できる一方で，ミスユース型と比較して本来は正しい振舞いであるところを攻撃と検知してしまう誤検知の発生が高くなるという問題がある．

アノマリ型の検出は様々なアプローチで研究が行われており，決定木（Decision Tree）やk-近傍法，サポー

1 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan
2 筑波大学システム情報系
Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan
3 金沢大学理工研究域
Institute of Science and Engineering, Kanazawa University, Kanazawa, Ishikawa 920–1192, Japan
†1 現在，株式会社富士通研究所
Presently with FUJITSU LABORATORIES LTD
a) kokubo.hirotaka@jp.fujitsu.com

トベクタマシン，自己組織化マップ（Self Organization Map，以後 SOM）などの機械学習を用いた手法も多い [3], [4], [5], [6], [7], [8], [9]．機械学習を用いたアノマリ型の検出手法では，現在の主流は単一の検出手法を採用しているものであり，複数の手法を組み合わせた場合の効果については，必ずしも十分に検討しきれているとはいえない [8], [9]．単一の検出手法のみを使って検出を行う場合，時間とともに動的に変化するデータへの対応が難しいことや，再学習のタイミングが難しいことなどが課題にあげられる．たとえば決定木を用いて検出を行う場合，マルウェアの流行など時間とともに傾向が強く変化するケースに対しては頻繁に再学習する必要が生じてしまう．

そこで本論文は，決定木と SOM を組み合わせた手法を提案する．SOM は逐次学習（判別と同時に学習を行う方式）が可能であり，刻々と変化する状況を判別結果に反映することができるため，判別が高速ではあるが逐次学習はできない決定木と合わせた．2 つの手法を合わせることで，アノマリ型の検知手法の性能向上を目指す [10]．

提案手法の評価のために，試作システムを開発し，実際の攻撃データを利用してマルウェアの検知率や誤検知率，処理速度を測定した．

評価実験ではサイバークリーンセンター（CCC）で収集された研究用データセットである CCC DATAset 2011 の攻撃通信データと，家庭環境や大学研究室における小規模環境での正常通信データを用いて評価を行った．

その結果，最も精度が良かったもので平均検知率（攻撃通信を攻撃通信だと正しく判別する割合）99.48%，平均誤検知率（正常通信を入力として与えると誤って攻撃通信だと判別する割合）0.04%を実現した．

本論文の構成は以下のとおりである．2 章で関連研究として機械学習と機械学習を応用したアノマリ型検出の先行研究を解説する．続く 3 章で決定木と SOM を合わせたアノマリ型マルウェア検知手法の提案を行う．4 章で評価の手法と前提について延べ，5 章でその評価結果を示す．最後に 6 章でまとめる．

## 2. 関連研究

提案手法で組み合わせる機械学習手法の動作などの基本的事項の説明と本論文での使用法について記述する．また，機械学習を使用して検知を行っている関連研究の紹介を行う．

### 2.1 機械学習

機械学習とは，データを人間が見て理解し処理するのではなく，コンピュータがデータを解析し学習を行う手法である．人間が対応困難な大量のデータを解析し，そのデータから有用なルールや分類法則などの知見を得ることができる．迷惑メールのフィルタリングや POS システムのデータからの購買分析などの大量のデータを背景に行う処理や，文字認識，顔認識や検出など従来では人間でなければ難しかった分野でも使われている．

機械学習にはベイズ分類器，サポートベクタマシン，決定木，自己組織化マップ（Self Organization Map），主成分分析，遺伝的アルゴリズム，k-近傍法など様々な手法が存在する．それぞれの手法は教師データと呼ばれる正答データを使用し学習を行う手法を教師あり学習，教師データなしで学習行う手法を教師なし学習に分けることができる．

#### 2.1.1 決定木

決定木（Decision Tree）は，多数の条件式から構成された木構造の分類器である．枝の分岐点である節に条件式があり，その条件式を評価し枝を進んでいくことで入力データの判別を行う．決定木はデータマイニングにおいて幅広く利用されている手法である．入力データの分類だけでなく，節の条件式に注目することである集合からルールを抽出し意思決定を行うことに利用することや，ルールを利用したマーケティングなどにも利用されている．決定木の利点として，教師あり学習のために高い正答率が期待できることや分類速度が高速なことがあげられる．決定木を構築するためには教師データと呼ばれる答えの付いた入力データが大量に必要となる．教師データが少ない場合，その現象を十分に分類できていない決定木が構築されてしまう可能性がある．また，多数の特徴量から構成された大量の教師データを扱う場合，決定木の構成時間が長くなる傾向があることが知られている．

具体的な決定木の構築アルゴリズムの 1 つに ID3（Iterative Dichotomiser 3）[11] がある．教師データ集合の平均情報量と，その集合をある特徴量を基準に分割した部分集合の平均情報量の期待値の差（情報ゲイン）が最大になるような特徴量を選ぶことで，節の条件式を決定するアルゴリズムである．連続値を取り扱えないという欠点も存在し，ID3 の拡張として連続値を取り扱える C4.5 も存在する．

ID3 ではまず，教師データ全体の平均情報量 $H(T)$ を計算する．次に特徴量を 1 つ選び，その特徴量を基準に教師データ全体を分割する．そして，分割されたデータの平均情報量の期待値 $H(S)$ を算出し，その特徴量の情報ゲイン $H(T) - H(S)$ を得る．情報ゲインが高い特徴量ほど教師データに対する識別力が高いといえる．そこで，最も情報ゲインが高い特徴量を節の分岐に用いる特徴量とする．

#### 2.1.2 自己組織化マップ

自己組織化マップ（Self Organizing Map，SOM）とはコホネンが提唱した教師なしニューラルネットワークアルゴリズムである [12]．多元の入力データを，近い性質のデータどうしが近い座標に配置されるように n 次元平面上に写像を行う．視覚的な効果から 2 次元平面上への写像が用いられることが多い．本論文ではこれ以降 2 次元自己組織化マップを単に SOM と呼ぶこととする．

SOM も決定木と同じくデータマイニングにおいてよく使われる手法であり，マーケティングや気象情報の分析，脈波解析やメタボリックシンドローム解析などの医療分野での応用も存在している．SOM のメリットとして，未知の通信の性質を類推できることや，判別と同時に学習を行う逐次学習ができることなどがあげられる．

### 2.2 機械学習をアノマリ型検知に適用した例

機械学習をネットワークセキュリティにおけるアノマリの検知に用いる研究も多い [3], [4], [5], [6], [7], [8], [9], [13], [14], [15].

柿本ら [6] は SOM をアノマリ検知に使用することを提案している．マルウェア動作時に呼び出されるシステムコールの列を SOM の入力として与えることでアノマリ検知を行おうとしている．SOM を使用する利点として，未知のデータの性質をマップ上で近接する既知のデータから類推することができることなどをあげている．

山田は，アノマリ型の侵入検知の教師情報としてミスユース型の検知結果を用いる方式を提案している [8]．テストデータとして 1999 年に公開された DARPA Intrusion Detection Data Sets [16] を用いており，主に HTTP リクエストを対象にミスユース型が見逃した攻撃通信の検知に成功している．この方式では，ミスユース型とアノマリ型の組合せを考察している．

野上はデータマイニングツール Weka を用いたネットワーク侵入検知法に関する研究を行っている [7]．DoS 攻撃とネットワークマッピングのどちらの検知においても高いものでは 99%以上の検知率を出しており，特に決定木を使ったアンサンブル学習がこれらの検知に有効であることを示している．Weka とはオープンソースのデータマイニングツールであり，内部に多くの機械学習・データマイニングアルゴリズムを含んでいる．教師データおよびテストデータに 1998 年版の DARPA Intrusion Detection Data Sets [16] を用いており，DARPA Intrusion Detection Data Sets に含まれる tcpdump 形式のデータから，特に TCP ポート，UDP ポート，ICMP メッセージ，TCP・UDP のトラフィック量，FTP・telnet・HTTP のトラフィック量に着目して抽出を行っている．そしてこれらのデータを用いて，Weka で DoS 攻撃とネットワークマッピングの検知を行っている．

## 3. 決定木と SOM の併用によるアノマリ型検知

これまでのアノマリ型の研究では，機械学習方法だけでなく幅広い手法により多くの研究がされてきた．一方で，商用化や運用実績が高い手法は現れておらずアノマリ型の研究はいまだ基礎研究段階にあるといえよう．

決定木は分類の高速性が優れた点であるが，特徴量が増加するとその構築時間が長くなるという問題を持つ．また逐次学習ができず，入力データが動的に変化する場合には木を再構築する必要がある．

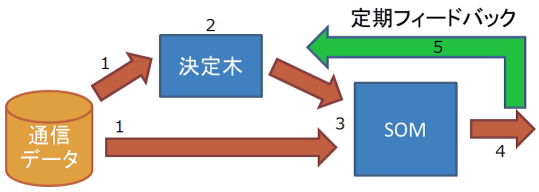


図 1 決定木と自己組織化マップの連結
**Fig. 1** Connection of Decision tree and SOM.

再構築の時間を短くするために特徴量を削減させることも考えられるが，マルウェアによる攻撃通信の検知の場合入力データはマルウェアの流行により大きく特性が変わるため，特徴量は広く利用可能にしておくことが望ましい．また再構築の負担を軽減するために，再構築を行う時間間隔を大きくとることも考えられるが，時間間隔が開くことによる動的な特性変化に決定木が対応できず，精度が落ちるという問題が発生する．

決定木が持つ分類の高速性を生かしつつ，特徴量を広く利用可能かつ動的な入力特性変化に対応するために，本論文では決定木と SOM を併用したアノマリ型検知手法を提案する．

以下 3.1 節に提案手法の概要を述べ，提案方式の詳細については SOM と決定木に着目し，それぞれ 3.2 節と 3.3 節に決定木アルゴリズムと SOM の構成を述べる．

### 3.1 提案手法概要

提案手法の概要を**図 1** に示す．決定木は判別対象を非常に高速に判別するため，SOM と連結させても SOM 単体で判別を行う場合とほぼ変わらない速度での判別が期待できる．一方，決定木を構成するために使用した教師データが古くなった場合，新しい教師データで再構成を行わなければ環境の変化から精度が落ちていくが，SOM 部分は最新のデータを判別しつつ学習することができるため，決定木の再構築までに生じる精度低下の穴を埋めることができると期待される．このように決定木の結果を参考に SOM で最終的な判別を行うことで，高い判定率を目指している．さらに，システムの判別結果を定期的にチェックし，決定木の教師データとしてフィードバックすることで持続的に安定した検知が行えることを期待している．以後，決定木のフィードバックと SOM の逐次更新をあわせて更新機能と呼ぶ．

### 3.2 決定木アルゴリズム

今回は決定木構築アルゴリズムとして，利用実績が高い ID3 を選択した．しかし，特徴量のうち値の割当てが定められているプロトコル番号などとは異なりパケットサイズ

やヘッダ長などは連続値のため，通常の ID3 では取り扱えないため連続値の取扱いが可能となるように拡張を行った．

### 3.3 SOM の構成

まず $i \in [0, L]$，$j \in [0, M]$（$L$，$M$ は任意の整数）の整数インデックスを持つ二次元平面上にユニット $u_{i,j}$ を並べる．ユニットはそれぞれ参照ベクトル $\vec{R}_{i,j} = [e_{i,j,1}, e_{i,j,2}, \ldots, e_{i,j,n}]$ を持っている．ただし，$e_{i,j,1}$ から $e_{i,j,n-1}$ にはそのユニットが持つ特徴量が格納されており，$n-1$ は特徴量の個数である．$e_{i,j,n}$ はそのユニットが正常通信と判別された回数と攻撃通信と判別された回数の両方を保持する特殊な要素とした．

SOM は学習と判別を行う．

**学習**：SOM に入力ベクトル $\vec{I} = [e'_1, e'_2, \ldots, e'_n]$ を与えると SOM は各ユニットの持つ参照ベクトル $\vec{R}_{i,j}$ ごとに，類似度 $s_{i,j}$ を次のように計算する．

$$s_{i,j} = \sum_{k=1}^{n} c_k \cdot Sim(e'_k, e_{i,j,k}) \quad (1)$$

ただし，$c_k$ は $k$ 番目の特徴量に設定された重み，$Sim(\cdot,\cdot)$ は特徴量ごとの類似度を算出する関数とする．そして，最も $s_{i,j}$ の高いユニットを，勝者ユニットとする．勝者ユニットとその周囲にあるユニットの参照ベクトル $\vec{R}_{i,j}$ を，

$$\vec{R}_{i,j} + (\vec{I} - \vec{R}_{i,j}) * f(distance, time)$$

で更新する．ただし，$f$ は勝者ユニットとの距離 $distance$ と時間 $time$ によって変化する係数列．

**判別**：入力ベクトル $\vec{I} = [e_1, e_1, \ldots, e_n]$ を与えると，入力ベクトル $\vec{I}$ とのノルムが最も小さい参照ベクトル $\vec{R}_{i,j}$ を持ったユニット $u_{i,j}$ を選び，勝者ユニットとする．勝者ユニットの参照ベクトル中の判別結果が書かれている要素 $e_{i,j,n}$ を参照する．$e_{i,j,n}$ を見て，正常通信と判別された回数と攻撃通信と判別された回数の大きい方を判別結果として返し，該当する通信の回数を 1 つ増やす．

## 4. システムの試作と評価手法

提案手法の処理速度性能と検知率，誤検知率を評価するために提案システムを試作した．本章では評価の手法と評価に用いたシステムと評価環境について解説する．

### 4.1 評価方法

評価は実際に流れたパケットを取得したデータ（キャプチャデータ）を事前に用意して行う．データは正規通信データと攻撃通信データの 2 種類を用い，まずそれぞれのデータを学習用とテストデータに分ける．

評価は学習用の正規通信データと攻撃通信データを用いて決定木の作成と SOM の学習を行い，その後テストデータを用いて攻撃/正規の判断をシステムが行う．

表 1 試作システムの環境
**Table 1** Environment of prototype system.

| | |
|---|---|
| OS | Windows7 64 bit |
| 言語 | Visual C#.NET |
| プロセッサ | Intel Core i7 2.2 GHz |
| RAM | 8 GByte |
| pcap 操作ライブラリ | 自作物 |

評価は以下の点から行う．

- 処理性能
- 検知精度
  - 平均検知率
  - 平均誤検知率

処理性能は 1 パケットあたりの処理時間を求め，試作システムを評価する．また CAIDA の調査による平均 IP パケット長 420 バイトを用いて処理速度を試算する．

検知精度について，山田の論文では，特定の方式や特定のデータに依存しない一般の目標値として検知率 99%，誤検知率 0.01%をあげている．本論文でもこれらの目標値を採用し，平均検知率 99%，平均誤検知率 0.01%とした．

また，決定木自身の性能がシステム全体の検知率にどう影響するかを調べるため，決定木部分を疑似判定器に置き換えた場合のシステム全体の検知率の計測を行った．疑似判定器は，設定した任意の精度で決定木の判定を行うプログラムである．

さらに，更新機能の有無による精度の差異を評価するために，更新機能を使ったケースと使わなかったケース双方でテストを行った．

### 4.2 試作システムと評価環境

システムは Visual C# .NET で開発を行った．パケットデータを扱うための pcap 操作ライブラリは自作のものを利用した．開発したシステムを 64 bit 版 Windows 7 が稼働するホストに搭載し，検知処理を行った．ホストの CPU は Intel Core i7（2.2 GHz），RAM は 8 GB である（**表 1**）．

試作システムは次のような手順で動作する．

(1) 通信データから特徴量を抽出し入力ベクトルを生成．
(2) 入力ベクトルを決定木に入れ，結果を得る．
(3) (2) で得た結果を入力ベクトルの末尾に加え，SOM に入力する．
(4) SOM から結果を得て，システムの判定結果とする．
(5) システムの判定結果を定期的に決定木の教師データとしてフィードバックする．

決定木と SOM に入力される特徴量は，攻撃通信データから得られるプロトコルヘッダ情報からチェックサムなどの明らかに攻撃との関連性が低い項目を除いたものを特徴量として抽出し用いる．使用した特徴量を**表 2** に示す．これらの特徴量はパケット単位で学習や判別に使用する．

表 2 特徴量一覧
Table 2 Feature value list.

| 種別 | 特徴量名 |
|---|---|
| キャプチャ時獲得 | パケット全体のサイズ |
| データリンク層 | 上位層プロトコル番号 |
| ネットワーク層（IP） | ヘッダ長 |
| | データグラム長 |
| | Flag |
| | TTL |
| | 上位層プロトコル番号 |
| ネットワーク層（ICMP） | メッセージタイプ |
| | コード |
| ネットワーク層（IGMP） | タイプ |
| トランスポート層（TCP） | 送信元ポート |
| | 宛先ポート |
| | ヘッダ長 |
| | URG，ACK，PSH，RST，SYN，FIN |
| トランスポート層（UDP） | 送信元ポート |
| | 宛先ポート |
| | データ長 |

本評価では宛先または送信元の IP アドレスおよび MAC アドレスを特徴量として採用していない．これらは実運用上において重要な手がかりとなる可能性の高い特徴量であるが，後述する今回用意したデータは正常通信データと攻撃通信データが異なるネットワーク環境下で取得されており，正常通信データと攻撃通信データの取得環境の差が決定木の学習と SOM の学習に強く反映され本来のシステムの精度よりも高く精度が出てしまうことが予想され，不当な結果が出てしまう恐れがあるためである．

SOM では精度向上のため次の方法でまず SOM の初期化を行った．

初期化：正常通信データと攻撃通信データから 1 日分を抜き出し教師データとする．インデックス $i$，$j$ を 2 つの領域に分割し，片側を正常通信データから抽出した入力ベクトルのみ，もう片側を攻撃通信から抽出した入力ベクトルのみで学習させる（ただし，1 度も勝者ユニットとなっていないユニットを発見した場合，そのユニットの参照ベクトルを入力ベクトルで置き換える）．

SOM の逐次更新はシステムが自動で行うが，フィードバックは実際に運用する際には人手による設定が必要であり，決定木の再構築を含むため手間と時間がかかる作業である．実際の運用のされ方を考え，今回の実験では更新機能を使用する場合，逐次更新はパケットごとに行いフィードバックはテストデータが流れている間に 1 回だけ行うものとする．フィードバックのタイミングは，テストデータの半分が処理された時点とした．

また SOM で用いる特徴量の重み $c_k$ はポート番号など関係の深そうなものの影響が強くなるように発見的に設定したほかに，ID3 で決定木を構成する際にノードの決定に用いられている平均情報量による情報ゲインによる特徴量抽出の方法を応用した．たとえば，パケット長は 100 byte 以下か，プロトコルは IP を使用しているかなどを基準とし，基準により分割する前のデータの集合の平均情報量と，分割後の平均情報量の期待値から求められる情報ゲインの値を基に $c_k$ を設定した．基準に基づき分割して得られたデータ集合が $n$ 個の要素を持つ場合の平均情報量は次の式で求められる．

$$H(P = \{p_i | 1 \leq i \leq n, i \in \mathbf{N}\}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

今回の実験では，検出プログラム部を SOM 部分の重みと更新機能の有無を変えることにより 20 種作成し，利用した．重みの設定方法は，5 種は情報ゲインを使った手法，残り 5 種は発見的な手法で設定した．これらそれぞれに対し更新機能ありの構成となしの構成を作成し 20 種のケースを作成した．

### 4.3 評価に用いるデータセット

評価に用いたデータセットは正常通信データと攻撃通信データの 2 種類である．

攻撃通信データはマルウェア対策研究人材育成ワークショップ（MWS2011）で提供された CCC DATAset2011 を利用した．CCC DATAset2011 は 3 種類のデータセットから構成されており，そのうち観測装置で取得した通信のキャプチャデータである「攻撃通信データ」を利用した．これはホスト OS 上の 2 台のゲスト OS で動作しているハニーポットの通信を tcpdump でパケットキャプチャしたデータである．ゲスト OS はそれぞれインターネット接続されており，2 台ともに Windows XP SP1 で定期的にクリーンな状態にリセットされる．今回実験に使ったデータの総パケット数は 45,367,110 件，96.08%が TCP，2.09%が UDP，1.82%が ICMP，0.006%が IGMP での通信である．データに関する詳細は畑田らの文献を参照されたい [17]．

正常通信データは，家庭環境における通信と大学研究室における通信を取得したデータを利用した．それぞれの通信は Web サイトの閲覧やファイル転送（SMB），電子メールの送受信など日常的な通信を行っているもので，外部との接続ポイントであるルータにおける通信を tcpdump で取得した．それぞれの通信は Norton Internet Security 2011 および Symantec Endpoint Protection において攻撃やマルウェアなどが検出されていないデータである．データ中の 90.77%が TCP，9.23%が UDP であった．

## 5. 評価結果と考察

### 5.1 評価結果

試作システムで評価した 20 種類のケース結果のうち，

表 3 判別結果
**Table 3** Classification result.

| ケース | 平均 TP | 平均 FP | 備考 |
|---|---|---|---|
| case A | 98.35% | 2.72% | 更新機能不使用&TP 最大 |
| case B | 95.84% | 0.39% | 更新機能不使用&FP 最小 |
| case B' | 99.48% | 0.04% | 更新機能使用&TP 最大&FP 最小 |

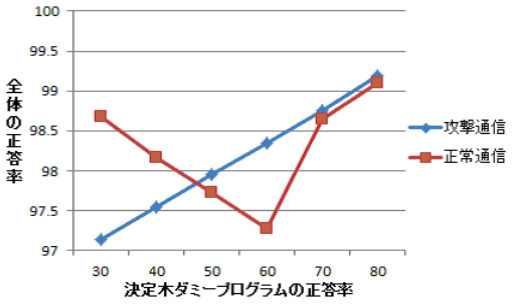


図 2 決定木の精度がシステム全体に与える影響
**Fig. 2** The effect of a decision tree.

更新機能を使用せず検知率（表中は True Positive の略で TP と表記）が最も高かった case A，更新機能を使用せず誤検知率（表中は False Positive の略で FP と表記）が最も低かった case B，更新機能を使用し，最も高い検知率と最も低い誤検知率を実現した case B' を**表 3** に示す．case B' としたのは，case B と caseB' の違いは更新機能の有無だけであり特徴量の重みは同じ値だったためである．

決定木の精度に対するシステム全体精度の評価結果を**図 2** に示す．決定木の精度を疑似判定器で任意に変更することで，システム全体の精度がどのように変化するかを見たものである．攻撃通信の正答率はつねに決定木の正答率の上昇に応じて大きな値となる．正常通信の正答率は決定木の精度が 60%付近のときに最も悪くなった後に決定木の正答率の上昇に応じて大きな値となっている．攻撃通信と正常通信のいずれにおいても，正答率がすべて 97%から 99%程度の範囲に収まっており，ある程度，高い値で推移していることも分かる．

判別にかかる速度は最も速くなった構成で 1 パケットあたり平均 3.2 msec であった．そのうち決定木部における処理時間は 0.00074 msec であった．最も遅い構成では 1 パケットあたり 22.2 msec であった．構成したものすべての平均値は 9.5 msec であった．

平均パケットサイズを 420 バイトとしたとき，平均 3.2 msec/packet のスループットは 1.001 Mbps，同じく 22.2 msec/packet では 0.144 Mbps，9.5 msec/packet では 0.3373 Mbps である．

## 5.2 考察

### 5.2.1 評価目標に対する考察

検知精度は，表 3 を見ると，更新機能が働くことで平均検知率と平均誤検知率の改善が行われていることが分かる．また case A はシステムが入力データを攻撃通信と判別する傾向に，case B は正常通信と判別する傾向にあり，検知率と誤検知率がトレードオフの関係になっていることが分かる．また更新機能を使用することで目標検知率 99%を達成することができている．更新機能により最新の状態を取り入れ続けることで高い検知率を持続させることができた結果だといえる．一方，誤検知率に関しては目標値 0.01%以下を達成したケースは 1 つもなかった．機械学習の判別傾向が異常側にやや傾いていると考えられるため，バランスをとる必要がある．

疑似判定器を用いて決定木の正答率差によるシステム全体の精度の違いを示した図 2 によると，決定木の正答率が 60%以上のときに決定木の正答率が上がると検知率，誤検知率ともに良くなる傾向があるため，決定木部分のさらなる改良によって目標検知率を達成しつつ誤検知率を下げることができる可能性がある．また，決定木部は高速に動作しているため処理時間をあまり増やさずに性能を向上させる効果が期待できそうである．この結果と case B' から，複数手法を組み合わせることによって実行時処理時間をあまり増やさずに精度維持，性能向上を実現していることが分かる．

### 5.2.2 その他の考察

処理性能が最速の構成で 1 パケットあたり平均 3.2 msec であり，平均パケットサイズを 420 バイトとしたときのスループットとして 1.001 Mbps にとどまった理由は以下のことが考えられる

(1) 手法自体の効率が高いとはいえない．
(2) 試作システムの機器や言語の性能が低い．
(3) 試作システムのソースコードが効率化されていない．

まず (2) については表 1 を見る限り CPU，RAM ともに最高性能とはいえなくても現状で妥当な性能を持っているということができ，また C#自身も他の C/C++や Java 言語などと比較して圧倒的に低いという評価がされてはいないため原因はこの点ではないということがいえよう．

(1) についての評価は，(3) についてのより深い実装評価がされない限り正確な評価は難しいが，試作システムであるため動作させることに重点を置いたことに加え，パケットの処理部分を自作したことや SOM 部分の実装最適化を行うことで高速化を図ることは十分可能であると考えられる．

実際の運用上では正常通信をどう集めるかにも気を配らなければならない．本方式は正常通信も教師データとして使用しているため，正常通信の質がシステムの検知率に直接影響を及ぼす．攻撃通信が混在していたり，ふだんあまり出現しない通信が大量に出現するなどの現象が起こったりした場合，システム全体の精度は落ちると考えられる．企業は個人よりも攻撃を受けやすい傾向にあるため，正常通信だけを厳選するという作業はややコストがかかると

いえるだろう．企業によってネットワークを流れる通信の傾向も異なると考えられるため，共通の正常通信データを作ることも非常に難しい．精度はやや落ちる可能性はあるが，明らかに攻撃のあった時間帯以外の通信データをもって正常通信とすることが現実的だと考える．

また，フィードバックをどう行うかといった問題も存在する．実験環境では入力データの真の分類を知ることができるため，100%の精度でフィードバックを行っている．しかし実環境ではそういったことはできないので，明らかに間違っている分類結果をフィードバックして精度を保つ必要があるだろう．

## 6. まとめ

マルウェアの侵入や攻撃を検出するアノマリ型手法において，判別が高速であるが逐次学習が困難であり再構築に時間がかかる決定木と逐次学習が可能であり動的に変化する入力特性に対応できる 2 次元自己組織化マップという 2 つの機械学習手法を組み合わせることで，効果的に攻撃通信を検知する手法について提案した．

また試作システムを開発し提案手法を実装し，処理速度や検出精度の評価を行った．評価は正常通信データと攻撃通信データの 2 種類を事前に学習させ，その後別の正常通信データと攻撃通信データで検出を行うことで測定した．

その結果，今回使用したデータでは最良の個体は 99.48%の確率で攻撃通信を検知し，誤検知率も 0.04%に抑えることができた．処理速度が遅く，誤検知率は目標値に達していなかったものの，検知率は目標値である 99%を実現したことで提案システムの有効性を示した．

今回の評価では，正常通信データから一部の特徴のあるデータのみを抜き出して比較するといったことを行っていない．このため，正常通信データによっては，データの特殊な性質のために，良い評価結果となってしまうことも考えられる．今後，この点の改善を含めて研究を発展させていく予定である．

**謝辞** 非常に有益な助言をいただいた査読者の皆様に感謝いたします．また，CCC DATAset 2011 を提供してくださった Cyber Clean Center の皆様，マルウェア対策研究人材育成ワークショップ 2011 実行委員会の皆様に感謝いたします．本論文の一部は日本学術振興会日中韓フォーサイト事業の助成により行われました（Part of this paper is supported by JSPS A3 Foresight Program）．

### 参考文献

[1] McAfee Labs：2011 年第 1 四半期脅威レポート，McAfee Labs（オンライン），入手先 ⟨http://www.mcafee.com/japan/media/mcafeeb2b/international/japan/pdf/threatreport/threatreport11q1.pdf⟩（参照 2011-08-01）．

[2] 情報処理推進機構：情報処理推進機構：情報セキュリティ：ネットワークセキュリティ関連用語集，入手先 ⟨http://www.ipa.go.jp/security/ciadr/crword.html⟩（参照 2011-12-01）．

[3] Chandola, V. Banerjee, A. and Kumar, V.: Anomaly detection: A survey, *ACM Computing Surveys*, Vol.41, No.3 (July 2009).

[4] Garcia-Teodoro, P., Diaz-Verdejo, J., Macia-Fernandez, G. and Vazquez, E.: Anomaly-based network intrusion detection: Techniques, systems and challenges, *Computers & Security*, Vol.28, No.1-2, pp.18–28 (Feb. 2009).

[5] Patcha, A. and Park, J-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends, *Computer Networks*, Vol.51, No.12, pp.3448–3470 (Aug. 2007).

[6] 柿本圭介，田中英彦：自己組織化マップを用いた異常検知についての一検討，情報科学技術フォーラム一般講演論文集 6(4)，pp.79–80 (2007).

[7] 野上晋平：データマイニングツール Weka を用いたネットワーク侵入検知法，次世代コンピューティングシステムに関する合同ワークショップ 5-01 (2009).

[8] 山田 明：ネットワーク侵入検知システムの高度化に関する研究，博士学位論文，東北大学大学院情報科学研究科 (2009).

[9] Shon, T. and Moon, J.: A hybrid machine learning approach to network anomaly detection, *Information Sciences*, Vol.177, No.18, pp.3799–3821 (Sep. 2007).

[10] 小久保博崇，満保雅浩，岡本栄司：攻撃通信を持続的に検知する合成型機械学習手法の検討，コンピュータセキュリティシンポジウム 2011 論文集，Vol.2011, No.3, pp.272–276 (2011).

[11] Quinlan, J.R.: Induction of Decision Trees, *Machine Learning 1*, pp.81–106 (1986).

[12] コホネン，T.：自己組織化マップ，シュプリガーフェアラーク東京 (2005).

[13] Labib, K. and Vemuri, R.: NSOM: A real-time network-based intrusion detection using self-organizing maps, *Networks & Security* (2002).

[14] Smith, R. Bivens, A. Embrechts, M. Palagiri, C. and Szymanski, B.: Clustering approaches for anomaly-based intrusion detection, *Proc. Intelligent Engineering Systems through Artificial Neural Networks*, pp.579–584, ASME Press (2002).

[15] Ramadas, M. Ostermann, S. and Tjaden, B.: Detecting anomalous network traffic with self-organizing maps, *Proc. 6th International Symposium on Recent Advances in Intrusion Detection* (*RAID*)-*2003*, Lecture Notes in Computer Science 2820, pp.36–54 (2003).

[16] Lincoln Laboratory, Massachusetts Institute of Technology: DARPA Intrusion Detection Data Sets, Massachusetts Institute of Technology (online), available from ⟨http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/⟩.

[17] 畑田充弘，中津留勇，秋山満昭：マルウェア対策のための研究用データセット—MWS 2011 Datasets，マルウェア対策研究人材育成ワークショップ 2011（MWS2011）(2011).

[18] Packet Length Distributions, The Cooperative Association for Internet Data Analysis (2000) (online), available from ⟨http://www.caida.org/research/traffic-analysis/AIX/plen_hist/⟩.

**小久保 博崇**

2010 年筑波大学第三学群情報学類卒業．2012 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻修士課程修了．同年富士通研究所入社．

**金岡 晃** （正会員）

2004 年筑波大学大学院博士課程システム情報工学研究科修了．同年セコム株式会社入社．筑波大学大学院システム情報工学研究科研究員を経て，2008 年より筑波大学大学院システム情報工学研究科助教．2010 年より情報通信研究機構招聘専門員兼務．ネットワークシステムの安全設計方式，暗号応用，電子認証に関する研究に従事．博士（工学）．電子情報通信学会，IEEE，ACM 各会員．

**満保 雅浩** （正会員）

1988 年金沢大学工学部電気・情報工学科卒業．1993 年東京工業大学大学院理工学研究科博士後期課程修了．博士（工学）．同年北陸先端科学技術大学院大学助手．その後，東北大学助教授，筑波大学助教授・准教授を経て 2011 年より金沢大学教授．情報セキュリティの教育・研究に従事．

**岡本 栄司** （フェロー）

1973 年東京工業大学工学部電子工学科卒業．1978 年東京工業大学大学院博士課程修了．工学博士．同年日本電気入社．その後，北陸先端科学技術大学院大学，東邦大学を経て，2002 年より筑波大学教授，現在に至る．情報セキュリティを中心とする情報数理工学の教育・研究に従事．1990 年電子情報通信学会論文賞，1993 年本会ベストオーサ賞受賞，2008 年本学会論文賞，2007 年・2009 年 CHES Best Paper Award．2003 年電子情報通信学会フェロー，2004 年本学会フェロー．著書『暗号理論入門』（共立出版），『電子マネー』（岩波書店）等．IEEE，ACM 会員