分散型侵入検知システムにおける免疫型診断モデルを用いた自己監視[†]

渡 邊 裕 司*・石 田 好 輝*

Self-monitoring Using Immunity-based Diagnosis in Distributed Intrusion Detection Systems[†]

Yuji WATANABE* and Yoshiteru ISHIDA*

In distributed intrusion detection systems, self-monitoring is a subject of current research. One possibility is that each intrusion detection system is checked periodically by others, and then corrupted intrusion detection systems can be identified. In this paper, we employ both an immunity-based diagnostic model and mobile agents for the self-monitoring. The immunity-based diagnosis, which was proposed by one of the authors in 1990, is performed by the mutual tests among units and the dynamic propagation of active states. In simulated distributed intrusion detection systems, we compare three methods: (1) a simple majority vote on host-to-host communication, (2) the immunity-based diagnosis on host-to-host communication, and (3) the immunity-based diagnosis with mobile agents. Some simulation results show that the immunity-based diagnosis can work better than the majority vote. We also clarify advantages and disadvantages of using mobile agents in comparison with the host-to-host communication.

Key Words: immunity-based distributed diagnosis, distributed intrusion detection systems, self-monitoring, mobile agent

1. はじめに

侵入検知システム (Intrusion Detection System, 以下 IDS と略す) とは, コンピュータおよびネットワークに対するセキュリティ侵害の検出, 通知, 検出情報の管理を行なうシステムである. そのタイプとして, ホスト型とネットワーク型, 不正検出と異常検出, 集中型と分散型などに分類される¹⁾. 最近では, 単独の IDS だけではなく, ネットワークに分散して配備された複数の IDS を連携させることの必要性が唱えられている²⁾.

IDS 構築の一つのアプローチとして, 生物の防御機構である免疫系から着想を得ることは有益であり, すでにいくつかの研究が存在する. Forrest らは, 免疫系の自己・非自己の識別方法をもとに, Unix プロセスに対して自己 (正常なふるまい) を定義し, 実験により異常なふるまいを検出できることを示した³⁾. さらに, 胸腺における T 細胞の選択 (negative selection) を参考にした検出器の生成法⁴⁾や, そのほかの免疫特性 (多様性, 分散検出, 記憶など) も取り入れた ARTIS フレームを適用したネットワーク型 IDS (LISYS)⁵⁾を提案した. 免疫細胞をエージェントとみなす

のならば, Crosbie, Spafford らによる自律エージェントを用いた IDS も挙げられ, 免疫系との類似性が指摘されている^{6), 7)}. また, Nishiyama らは免疫細胞エージェント間の協調に基づいて不正侵入の検知・排除を行なった⁸⁾. さらに, 生体の免疫細胞が体内を移動 (循環) していることから, エージェントを移動させるアプローチ, すなわちモバイルエージェントを用いたシステムも存在する^{9)~11)}.

これら複数のエージェント (あるいは IDS) をネットワーク上に配置したものは並列分散型の範疇に入る. 分散型 IDS は, 一枚岩な (monolithic) IDS と比較して, 拡張性, 耐障害性, 動的な再構成などで優るといわれている⁷⁾. しかし, 耐障害性に関して, 侵入者によって IDS 自身が改ざんされたり, IDS の設定に誤りがあった場合を想定した研究はまだ少ない. これまでに自己監視 (self-monitoring) という用語を用いて, 各 IDS が相互にチェックするという考えは示されている^{5), 7)}ものの, 詳細な検証は行なわれていない. 誤りのある IDS は, 正常なふるまいを異常とみなす誤報 (false positive) や異常なふるまいを正常とみなす欠報 (取りこぼし, false negative) を引き起こすため, IDS 自身の誤り検出も重要である.

この問題に対処するために, 免疫系のほかの応用を眺め直してみると, B 細胞間の相互作用を唱えた「免疫ネットワーク説¹²⁾」から発想を得た「免疫型診断モデル」が提案されている^{13), 14)}. この診断モデルは, 相互にテストしうるユニットからなるシステムに対して, そのテスト結果ならびにユニットの活性・非活性をもとに, ダイナミカルモデルに

[†] 第 15 回自律分散システム・シンポジウムで発表 (2003・1)

* 豊橋技術科学大学知識情報工学系

* Department of Knowledge-based Information Engineering,
Toyohashi University of Technology
(Received July 7, 2003)
(Revised April 7, 2004)

よって各ユニットの正常・異常を判定する。このモデルは、主にプラントシステムなどの診断に応用されてきた^{14), 15)}が、最近、筆者らによってモバイルエージェントシステムにおける誤動作検出に応用された¹⁶⁾。

本論文では、この免疫型診断モデルを用いて分散型IDSの自己監視を行ない、ルールが改ざんされたIDSを分散的に検出する。分散型IDSを抽象化したシミュレーションシステム上で、各種パラメータを変更し「多数決モデル」と比較することで、「免疫型診断モデル」の有効性を確認する。さらに、前述のモバイルエージェントを用いたアプローチがあるように、「相互作用」だけでなく「移動（循環）」も免疫細胞が持つ重要な特徴である。そこで、自己監視の方法として、「ホスト間通信を用いる方法」と「モバイルエージェントを用いる方法」の二つを提案し、両者の比較によってモバイルエージェントを用いる利点と欠点を明らかにする。

2. 分散型IDSのシミュレーションシステム

診断モデルの基本的なふるまいを調べるために、分散型IDSを抽象化したシミュレーションシステムを用いる。まず、Fig. 1に示すように N 台の各ホスト上でIDSが起動しており、各ホストは N_c ($2 \leq N_c < N$) 台のほかのホストと接続し、ルールやエージェントなどを送受信する。各ホストは、最低限の接続性を保障するためにリング状に接続した後、ランダムに残った残り $N_c - 2$ 台のホストと接続し合うことでネットワークを形成する。

つぎに、各IDSが備えるルールを簡略化する。現実のIDSの多くは、大量のルールを持ち、そのルールとホストの監査証跡（ログ）を比較することにより、ホストの異常の有無を判定する。たとえば、フリーウェアのsnort¹⁷⁾には、約1600個のルールがデフォルトでは/etc/snort/にサービスごとにファイルとしてまとめられている（Fig. 2）：snortの各ルールの記述は、複雑に見えるが、基本的にはあるログが観測されたときにいかなる行動をとるか（警告する（alert）や通過させる（pass）など）を定めているに過ぎない。そこで、ルールをラベルとデータ（条件部と行動部）の対で表わし、ラベルはA, B, ... を、データは0と1のみをとるとする。また、利用可能なルールの総数を L とすると、一般に各ホストは異なるサービスを提供していることから、各IDSが利用するルールにも相違があると仮定し、各IDSが持つルール数を L_h ($\leq L$) とする。

最後に、全ルールに対して正しいラベル・データ対があらかじめ決定されているもとの、改ざんルールをデータ部の01反転で表わす（たとえばFig. 1におけるIDS_NのラベルCやWのルール）。改ざんされたIDSは、所持ルールが割合 e ($0 < e \leq 1$) で改ざんされているとする。また、すでに何らかの侵入が行なわれ、 N_f ($\leq N$) 個のIDSが改ざんされた状態からシミュレーションを開始する。

Fig. 1 Simulated distributed intrusion detection systems

```
[root@wata root]# ls /etc/snort/
local.rules      snmp.rules
attack-responses.rules  snort.conf
backdoor.rules    multimedia.rules  snort_rutkie.conf
bad-traffic.rules  mysql.rules      sql.rules
chat.rules        netbios.rules    telnet.rules
classification.config  nntp.rules      virus.rules
ddos.rules        oracle.rules     web-attacks.rules
deleted.rules      other-ids.rules  web-cgi.rules
dns.rules         p2p.rules       web-client.rules
dos.rules         policy.rules     web-coldfusion.rules
experimental.rules  pop3.rules      web-frontpage.rules
exploit.rules      pcrn.rules      web-iis.rules
finger.rules       reference.config web-misc.rules
ftp.rules          rservices.rules web-php.rules
icmp-info.rules    scan.rules      x11.rules
imap.rules         shellcode.rules
info.rules         smtp.rules

[root@wata root]# head /etc/snort/web-cgi.rules
# (C) Copyright 2001,2002, Martin Roesch, Brian Caswell, et al.
# All rights reserved.
# $Id: web-cgi.rules,v 1.56 2002/08/18 20:28:43 cazz Exp $
#-----
# WEB-CGI RULES
#
alert tcp $EXTERNAL_NET any -> $HTTP_SERVERS $HTTP_PORTS (msg:"WEB-CGI
HyperSeek hxx.cgi directory traversal attempt"; uri:content:"/hxx.cgi";
content:"../../../../"; content:"%00"; flow:to_server,established;
reference:bugtraq,2314; reference:cve,CAN-2001-0253;
alert tcp $EXTERNAL_NET any -> $HTTP_SERVERS $HTTP_PORTS (msg:"WEB-CGI
HyperSeek hxx.cgi access"; uri:content:"/hxx.cgi"; flow:to_server,
established; reference:bugtraq,2314; reference:cve,CAN-2001-0253;
classtype:web-application-activity; sid:1607; rev:3;)
[root@wata root]#
```

Fig. 2 Example of snort rules¹⁷⁾

3. 免疫型診断モデルによる自己監視

本研究では、免疫型診断モデルを用いて、相互にテストし各IDSの信用度を更新することで、ルールが改ざんされたIDSを検出する。その方法として、「ホスト間通信を用いる方法」と「モバイルエージェントを用いる方法」の二つを提案する。以下では、免疫型診断モデルと二つの方法について詳説する。

3.1 免疫型診断モデル¹⁴⁾

簡単な例として、Fig. 3の左に示す5個のユニットからなる相互評価のネットワークにおいて、信用できないユニット（抗原）を探すことを考える。ここで、ユニット i から j への矢印が +1 ならば信用できると評価し、逆に -1 ならば信用できないと考えているとする。信用できるユニットは正しい評価をするが、信用できないユニットは正しい評価をするとは限らない。この例では、信用できないユニットは正反対の評価つまり信用できるユニットに -1 を、信用できないユニットに +1 を与えるとする。これにより同図に示すように矢印の双方向で同じ評価結果となる。ここではユニット4と5が信用できないとしている。

ユニット i に信用度 $R_i^* \in \{0, 1\}$ ^(注1)を与え、1であれば信用できる、0ならば信用できないと診断する。各ユニット

(注1) 後述の信用度 R_i と異なり時間・状態ともに離散的である点を区別するためにアスタリスクをつける。

Fig. 3 Example of mutual evaluation network

で相手からの評価結果を単純に合計して、その合計が正のとき $R_i^* = 1$ 、負のとき $R_i^* = 0$ とすると、ユニット 1 以外は合計が負となり信用できないことになる。つまりユニット 2 と 3 に対して誤報が起こる。そこで、信用できないユニットの評価結果は信用できないことから、評価結果に自分を評価した「相手の信用度」を重みづけして集票し、その集票結果から信用度更新を繰り返すと、 R_i^* は Fig. 3 の右に示すように変化する。各ステップの変化は以下のとおりである。

- (1) 初めはすべてのユニットが信用できるつまり $R_i^* = 1$ とする。
- (2) 各ユニットで評価結果に相手の信用度（この時点ではすべて 1）を重みづけした合計が正のとき R_i^* を 1 に、負のとき 0 に更新すると^(注2)、ユニット 1 以外は $R_i^* = 0$ になる。たとえばユニット 2 では $(+1) \times 1 + (-1) \times 1 + (-1) \times 1 = -1$ より $R_2^* = 0$ になる。
- (3) 更新された信用度を用いると、ユニット 1 から出ている矢印以外は 0 が重みづけられる。つまり現時点で信用できるユニット 1 からの評価結果のみを用いることになる。たとえばユニット 2 では $(+1) \times 1 + (-1) \times 0 + (-1) \times 0 = 1$ となり、 $R_2^* = 1$ に戻る。ユニット 4 と 5 は $R_i^* = 0$ のままである。
- (4) 同様に更新しても R_i^* は変化せず、正しい診断に収束する。

Fig. 3 は、単純に評価結果の多数決を行なうと誤報となる一方で、重みづけによって正しい診断に収束する例である。しかし、たとえば下記の場合は単純な多数決でもユニット 2 と 3 に対して誤報は起こらない。

- 信用できないユニット数を減らしてユニット 4 のみとする場合、ユニット 5 からの評価結果の正負が反転し、ユニット 2 と 3 への評価結果の単純合計は正となる。
- ユニット 2 と 3 の間でも相互評価する場合、その評価結果は +1 となり、ユニット 2 と 3 への結果の合計は 0 となる（0 ならば信用できるとする）。
- 信用できないユニット 5 が信用できるユニット 2 と 3 に対して正しい評価 (+1) を行なう場合、評価結果の合計は正となる。

つまり正しい診断に収束するかどうかは、信用できないユニット数や評価し合うユニット数などに依存する。そこで、

(注2) 合計が 0 ならば更新しない。

シミュレーションでは、対応するパラメータを変化させて診断性能を評価する。

免疫型診断モデルの第 1 の特徴は、この信用度の重みづけつまり信用度できるユニットのみが相互作用に関与することにある。免疫系と対応させるならば、信用度は抗原や免疫細胞の濃度、正負の評価結果は細胞間の刺激抑制とみなすことができ、活性化され濃度（信用度）の高い細胞のみが相互作用し、最終的には抗原（信用できないユニット）の濃度を減らすことになる。

さらに、上記の離散的なモデルを連続系の動的モデルで捉え直すと、ユニット i の信用度 R_i は、以下のダイナミカルモデルによって変化する：

$$\frac{dr_i(t)}{dt} = \sum_j T_{ji} R_j + \sum_j T_{ij} R_j - \frac{1}{2} \sum_{j \in \{k: T_{ik} \neq 0\}} (T_{ij} + 1) \quad (1)$$

$$R_i(t) = \frac{1}{1 + \exp(-r_i(t))} \quad (2)$$

ここで、 $r_i \in (-\infty, \infty)$ は媒介変数であり、(2) 式により信用度 $R_i \in [0, 1]$ へ変換される。また、 T_{ji} は、ユニット j がユニット i をテストした結果であり、テストが正しければ 1、まちがっていれば -1、存在しなければ 0 とする。具体的なテスト方法については 3.2 節で述べる。

(1) 式の右辺の第 1 項は、ユニット i がほかからの評価を総合している項であり、テスト結果 T_{ji} に相手の信用度 R_j を重みづけする。右辺第 2, 3 項はユニット i がほかを評価することにより、ほかから反射的に評価されているものを総合している項である。これは、「反射効果」と呼ばれ、免疫型診断モデルの第 2 の特徴である。この効果は、ある人が信用の高い人に対してうそつきだということ、逆にその人自身の信用が低く評価されてしまうことに相当する。なお、このモデルはホップフィールドネットワークと同様にある種のエネルギー関数を用いることで、信用度 R_i の 0 または 1 への収束性が保証されている。

免疫型診断モデルは多数決モデルの一種であり、(1) 式を以下の式に変更することで、重みづけや反射効果のない単純な「多数決モデル」となる。

$$\frac{dr_i(t)}{dt} = \sum_j T_{ji} \quad (3)$$

免疫型診断モデルの有効性を明確にするために、この「多数決モデル」と比較する。

3.2 ホスト間通信を用いた方法

診断を行なうためには、(1) 式の T_{ji} を決定する必要がある。本節ではホストつまり IDS 間で直接求める方法を示す。Fig. 4 に示す IDS_i と IDS_j の間のやり取りを通じて相互テストと信用度の更新を説明する。

- (1) IDS_i は、テスト用にランダムに選んだ L_t ($\leq L_h$) の個のルールを IDS_j に送信する (Fig. 4 では $L_t = 2$)。

Fig. 4 Self-monitoring using host-to-host communication

(2) IDS_j は、受信したルールと所持するルールを比較し、対応するラベルがあるデータに対して、値の一致・不一致を調べ、テスト結果 T_{ji} を以下のように決定する。

$$T_{ji} = \begin{cases} 1 & \text{比較できる全ルールが一致} \\ -1 & \text{不一致のルールがある} \\ 0 & \text{比較できるルールがない} \end{cases} \quad (4)$$

Fig. 4 では、ラベル A のデータは所持していないため、ラベル C のデータだけをみて $T_{ji} = 1$ となる。

(3) IDS_j は、テスト結果 T_{ji} と自分の信用度 R_j を IDS_i に返信する。

(4) 逆に、 IDS_i による IDS_j のテストにより T_{ij} が求まる。

(5) IDS_i は、接続している残り $N_c - 1$ 台のホスト上の IDS とともに上記の処理を行なった後、(1), (2) 式により信用度 R_i を更新する。信用度の更新はほかの IDS と同期して行なう。

(6) 以上の処理を 1 ステップとし、相互テストと信用度の更新を繰り返す。

3.3 モバイルエージェントを用いた方法

前節の接続ホスト間で送受信するテスト用ルールを、ホストを介して別のホストに転送することも可能である。これは、テスト用ルールがホスト間を移動しているとみなせば、テスト用ルールに「移動」と「診断」の機能を与えた「モバイルエージェント」を用いる方法とも考えられる。Fig. 5 に示すモバイルエージェントを用いた方法では、 M 個のモバイルエージェントが各 IDS と相互にテストする。同じホスト上にいるエージェント同士（同図の Ma_1 と Ma_2 ）でも相互にテストできるとする。ただし、異なるホスト間のエージェント同士（ Ma_1 と Ma_M ）はテストできないとする（できてしまうとエージェントが移動する意義がなくなるためである）。相互テストと信用度の更新は、各 IDS のテスト相手がエージェントであることとそのエージェントが診断後にランダムに選んだ接続先ホストに移動することを除き、ホスト間通信と同じ処理 (1) ~ (6) を行なう。

この方法において、エージェントが持つルール数、エージェントの生成方法とその数を決めなければならないが、ホスト間通信の方法と正確に比較するためにできるだけ同じ条件にする必要がある。そこで、まずエージェントが持つルール数は前節のテスト用ルール数 L_t と同じにする。つぎに、送受信するテスト用ルールは IDS のルールからランダムに選ばれることから、エージェントはシミュレーション開始時

Fig. 5 Self-monitoring using mobile agents

に IDS のルールからランダムに選んだ L_t 個のルールを複製して作られるとする。ここで、ホスト間通信の場合で改ざん IDS が改ざんルールを送信しうることと同様に、改ざん IDS が改ざんエージェント（Fig. 5 の Ma_M ）を生成することに注意する。改ざんエージェントは正常な IDS を誤診する（信用度を下げる）可能性がある。最後に、ホスト間通信の場合には各 IDS は N_c 台の隣接 IDS とテストを行なうことから、各 IDS あたりのモバイルエージェント数 M/N を N_c と同じにする。つまり、1 ステップで各 IDS がテストに用いるルール数を $L_t N_c = L_t M/N$ とする等しい条件のもとで両手法を比較する。

4. シミュレーション

シミュレーションでは、以下の三つの方法（カッコ内は略記）の検出性能を比較する。

- ホスト間通信で多数決モデルを用いる方法 (MV_H2H)
- ホスト間通信で免疫型診断モデルを用いる方法 (ID_H2H)
- モバイルエージェントで免疫型診断モデルを用いる方法 (ID_MA)

以下では、評価指標とパラメータについて説明し、結果と考察を述べる。

4.1 評価指標とパラメータ

シミュレーションの各ステップでは、IDS（またはエージェント）の信用度 R_i を記録するとともに、「誤報 (false positive)」と「欠報 (false negative)」も求める。誤報とは「正常」にもかかわらず「異常」と判定されることで、逆に欠報とは「異常」にもかかわらず「正常」と判定されることであり、両者をできるだけ小さくすることが望まれる。

本シミュレーションでは、誤報率 α と欠報率 β を以下のように定義する：

$$\alpha = \frac{N_t^{low}}{N_t} \quad (5)$$

$$\beta = \frac{N_f^{high}}{N_f} \quad (6)$$

ここで、 $N_t (= N - N_f)$ は正常な IDS の個数、 N_t^{low} は正常な IDS のうち信用度 R_i が θ -未満のものの個数を表わ

Table 1 List of parameters

	Explanation	Default
$r_i(0)$	Initial value of intermediate variable	5.0
$R_i(0)$	Initial value of credibility	1.0
θ	Diagnostic threshold	0.5
N	Total number of IDSs (hosts)	50
N_f	Number of corrupted IDSs	15
N_c	Number of connected hosts per host	7
M/N	Number of mobile agents per host	
L	Total number of available rules	5000
L_h	Number of IDS's rules	5000
L_t	Number of transferred rules	150
	Number of agent's rules	
e	Corruption rate	0.5

す。また、 N_f は改ざん IDS の個数であり、 N_f^{high} は改ざん IDS のうち信用度 R_i が θ 以上のものの個数を表す。1 試行 30 ステップまで計算し、初期配置などを変えて 50 試行行ない、平均誤報率 $\bar{\alpha}$ と平均欠報率 $\bar{\beta}$ を求める。

つぎに、前述の全パラメータを列挙すると Table 1 となる。このうち最初の三つは診断モデルが有する「診断パラメータ」であり、適切な値に設定する必要がある。一方、残りの七つは診断されるシミュレーションシステムが持つ「環境パラメータ」である。環境パラメータは、ユニット (IDS とエージェント) 数に関する最初の三つとルール数に関する残りの四つに分けることができる^(注3)。環境パラメータに対しては、デフォルト値以外でも各手法が適切な診断を行なうか調べる必要がある。以下では、パラメータを変化させながら検出性能の変化も観測する。

4.2 信用度の推移による比較

まず全パラメータをデフォルト値に固定して、信用度の推移を観察した。Fig. 6 に三つの方法それぞれにおける正常 IDS と改ざん IDS の信用度の変化を示す。ここでの信用度は、ある一つの IDS に着目して、試行を 50 回繰り返したときの平均である。各方法ともに、正常 IDS は高い信用度に、改ざん IDS は低い信用度となり、改ざん IDS を検出できている。しかし、収束値と収束速度に違いがある。

多数決モデルで正常 IDS の信用度が 1 に達しない理由は、50 試行のうち 1 割程度において正常 IDS の信用度が減少し誤報が発生しているためである (次節でより明らかにする)。また、モバイルエージェントを用いる方法で信用度の収束に約 2 倍の時間を要している理由として、まずエージェントが診断され、その後で IDS が診断されるために遅れが生じていると考えられる。信用度の推移では、「ホスト間通信で免疫型診断モデルを用いる方法」が最も良いといえる。

4.3 環境パラメータに対する検証

4.3.1 ユニット数に関するパラメータ

改ざん IDS 数 N_f と一ホストあたりの接続ホスト数 N_c (モバイルエージェントの場合はエージェント数 M/N) は、

(注3) 改ざん率 e は、改ざん IDS の改ざんルール数を決めるパラメータであるから、後者に分類する。

(a) Majority vote on host-to-host communication

(b) Immunity-based diagnosis on host-to-host communication

(c) Immunity-based diagnosis with mobile agents

Fig. 6 Transition of the average credibility R_i over 50 trials for normal and corrupted IDSs

3.1 節でも触れたように診断性能に影響を与えると予想される。そこで、これらのパラメータを変化させて各手法の性能を比較する。Fig. 7 の (a) に N_f を変化させた (ほかのパラメータはデフォルト値で固定した) 場合、(b) に N_c ($= M/N$) を変化させた場合の 30 ステップ後の各手法の平均誤報率 $\bar{\alpha}$ と平均欠報率 $\bar{\beta}$ を示す。

同図 (a) の改ざん IDS 数 N_f に関しては、より多くの改ざん IDS を検出すべきであり、異常を正常とみなす欠報率 $\bar{\beta}$ は非常に少なく (多数決モデル MV_H2H では 0)、異常は異常と判定している。しかし、誤報率 $\bar{\alpha}$ を見ると、モバイルエージェント ID_MA の場合でも N_f が 30 を超えると急激に増えており、これは単にすべての IDS を異常とみなしているだけで正しい診断とはいえない。これは、診断モデルが基本的に多数決であり、総 IDS 数 N に対して正常 IDS が半数未満 (N_f が半数以上) になると正しく診断されなくなるためである。以下では正しく診断しなければならない $N_f < 25$ に絞り、欠報率に比べて値の大きい誤報率について

Fig. 7 Average false positive and negative rates ($\bar{\alpha}$ and $\bar{\beta}$) after 30 steps over 50 trials changing N_f or $N_c (= M/N)$ for three methods

て各手法を比較考察する。

まず、多数決モデル MV_H2H では、免疫型診断モデル ID_H2H と比べて少ない N_f でも誤報が発生している（デフォルト値の $N_f = 15$ でも一割ほど誤報がある）。これは、Fig. 3 の例で示したように多数決では誤報となるが、免疫型診断モデルでは信用度の重みづけにより正しい診断になるためと考えられる。

つぎに、ホスト間通信の免疫型診断モデル ID_H2H でも N_f が半数に達する前に誤報が発生している。Fig. 8 に示す具体例を用いてこの理由を考察する。実際に用いた信用度 R_i は連続値であり、テスト結果 T_{ji} は比較するルールに依存するが、説明を簡単にするために、Fig. 3 と同様に離散信用度 R_i^* を使い、双方向でテスト結果が同じであるとする。この例では、 $N = 8, N_f = 3, N_c = 3$ であり、 N_f はまだ半数に達していない。Fig. 3 と同様に信用度を推移させると、図に示すようにまず IDS₇ の信用度のみが 1 となり、その後は重みづけにより IDS₇ からのテスト結果のみが働き、正常 IDS の 6 と 8 の信用度は 1 に戻る。しかし、正常 IDS の 1 と 2 はともに信用度が 0 であるため、お互いの +1 のテスト結果に 0 が重みづけられて機能せず、 $R_i^* = 0$ のままで誤報となる。つまり、信用度が高い正常 IDS が隣接していないために、一度下がった信用度が戻らない。一方、モバイルエージェントでは、テスト相手は隣接の IDS ではなくエージェントであるため、正常なエージェントが移動してきてそのような IDS も正しく診断していると考えられる。

Fig. 8 Failure case where normal IDS 1 and 2 come to low credibility

この例で接続数 N_c を増やして、たとえば正常 IDS₁ が正常 IDS₇ と相互テストするようにすれば、信用度は上がり誤報は発生しない。また、Fig. 3 の例でもユニット 2 と 3 の相互評価を追加すれば、多数決モデルでも誤報は起こらない。 N_c が増えると誤報率と欠報率ともに 0 になることが Fig. 7 (b) の結果からも確認できる。しかし、通信量を減らすためにもできるだけ少ない接続数やエージェント数を望むのであれば、誤報が発生しない「モバイルエージェントで免疫型診断モデルを用いる方法」が最も良いといえる。

4.3.2 ルール数に関するパラメータ

ここでは、各 IDS が持つルール数 L_h 、送受信するルール数（エージェントが持つルール数） L_t 、改ざん率 e を変更

して、各手法の診断性能の変化を調べる。Fig. 9の(a)に L_h の場合、(b)に L_t の場合、(c)に e の場合の30ステップ後の各手法の平均誤報率 $\bar{\alpha}$ と平均欠報率 $\bar{\beta}$ を示す^(注4)。

すべての結果で値が小さくなるにつれて、多数決モデルMV_H2Hでは誤報が減り、全手法で欠報が急激に増加する。これは誤報と欠報のトレードオフによると予想される。そこで、誤報と欠報が起こる各パラメータの境界値を大まかな計算で求める。ここでは、ホスト間通信を行なう正常IDSと改ざんIDSのみに限り、その間で1ステップだけやり取りする簡単な状況において、誤報と欠報が起こる理由を考える。

誤報は、正常IDSが送った正常ルールのどれか一つが改ざんIDSの改ざんルールに対応し、(4)式により改ざんIDSがテスト結果として-1を正常IDSに返すために起こる。改ざんIDSの各ルールが正常、改ざん、未所持である確率は、それぞれ $(1-e)L_h/L$, eL_h/L , $(L-L_h)/L$ であるから、改ざんIDSが受け取った L_t 個の正常ルールに対応する改ざんルール数の期待値は eL_tL_h/L となる。対応する改ざんルールが一つでもある、つまり $eL_tL_h/L \geq 1$ ならば誤報は起こりうる。

一方の欠報は、改ざんIDSが送ったルール中の改ざんルールのすべてが正常IDSの所持ルールに含まれず、正常IDSがテスト結果として-1を改ざんIDSに返せないために起こる。改ざんIDSは所持ルールが割合 e で改ざんされているため、改ざんIDSが送った L_t 個に含まれる改ざんルール数の期待値は eL_t となる。正常IDSの各ルールの所持確率は L_h/L であるから、 eL_t 個の改ざんルールに対応する所持ルール数の期待値は eL_tL_h/L となる。対応する所持ルールが一つもない、つまり $eL_tL_h/L < 1$ ならば欠報は起こりうる。

誤報と欠報で不等号のみが異なる式となり、 $eL_tL_h/L = 1$ が境界値を決めるパラメータ関係式と予想される。そこで、デフォルト値に対する L_h , L_t , e の境界値を求め、Fig. 9の各グラフと比較する。図(a)の L_h の場合、デフォルト値の $e = 0.1$, $L_t = 150$, $L = 5000$ を代入すると $L_h = 333$ となり、欠報はこの値以下で急激に増加している。図(b)の L_t と図(c)の e の場合も、それぞれ $e = 0.1$, $L_h = 5000$, $L = 5000$ を代入すると $L_t = 10$ に、 $L_t = 150$, $L_h = 5000$, $L = 5000$ を代入すると $e = 0.0066$ になり、同じことがいえる。しかし、誤報に関しては、多数決モデルMV_H2Hでは境界値以上で増加するもののほぼ一定となり、免疫型診断モデルID_H2HやID_MAでは増加後に0に戻る。また、各手法で増加し始める値も異なる。これは、上記の誤報の計算には、正常IDSが改ざんIDSからテスト結果の-1を返されたとしてもほかの正常IDSから+1を受け取れば誤報とならないことが含まれていないためである。今後、ほかの

IDSとのやり取り、接続数、改ざんIDS数、診断ステップ数なども考慮してより正確な境界値の算出が必要である。

4.4 診断パラメータに対する検証

本節では、診断パラメータに対する性能の変化を調べる。Fig. 10の(a)(b)(c)に、媒介変数の初期値 $r_i(0)$ 、信用度の初期値 $R_i(0)$ 、診断のしきい値 θ それぞれに対する30ステップ後の各手法の平均誤報率 $\bar{\alpha}$ と平均欠報率 $\bar{\beta}$ を示す。

同図(a)より $r_i(0)$ が0に近づくと急激に誤報が増加することがわかる。 $r_i(0) = 0$ の場合のID_H2Hの信用度 R_i の推移を見ると、Fig. 11に示すように最初に正常IDSの信用度の落ち込みがある。これは、(1)式の r_i の変化量が小さいとき、(2)式には初期値 $r_i(0)$ がそのまま代入され、 $r_i(0) = 0$ ならば $R_i = 0.5$ になってしまうためである。(2)式に $r_i(0) = 5$ を代入すると $R_i = 0.9933$ となること、ならびにFig. 10(a)から $r_i(0) \geq 7$ でわずかながら欠報が観測されたことから $r_i(0)$ のデフォルト値は5とした。なお、Fig. 11で正常IDSの信用度が1に達しない理由は、Fig. 10(a)からわかるように誤報が発生しているためである。

つぎに、信用度の初期値 $R_i(0)$ を0から1まで変化させた場合、図(b)より診断性能は変化しないことがわかる。ここでFig. 3の離散信用度 R_i^* の場合、初期値0で推移させるとすべて $R_i^* = 0$ のままで正しく診断されない。しかし、連続値 R_i では初期値0でも診断できている。これは、連続値では変数 r_i を介するために、たとえ $R_i(0)$ が小さくても $r_i(0)$ がある程度大きければ、(2)式によって R_i が1近くに近づくためである。実際 $r_i(0) < 5$ の場合、小さい $R_i(0)$ では正しく診断できないことを確認している。 $R_i(0)$ のデフォルト値は大きな値すなわち1にするのが良いといえる。

最後に、 θ のデフォルト値は単純に信用度 R_i の取りうる範囲 $[0, 1]$ の中間の0.5としたが、図(c)の結果より θ を変更しても診断性能は変わらないことがわかる。これは、3.1節で述べたように免疫型診断モデルの信用度は0または1への収束が保障されており、Fig. 6の信用度推移が示すように数ステップではほぼ収束した状態になるためである。しかし、30ステップではなく数ステップで判定する場合などには、 θ が影響すると予想される。

5. 議 論

本章では現時点での課題と今後の展開について議論する。

●信用度や診断モデルの改ざんと偽装

現時点では、信用度や診断モデル自体を改ざんすることや改ざんIDSがテスト結果や信用度を偽って返答することまでは考慮していない。改ざんに対しては、暗号化や署名も組み入れる必要があるだろう。一方、偽装に関しては、(4)式を変更して改ざんIDSがでたらめなテスト結果を返すことまでは行ない、適切に検出できることは確認した。しかし、信用度まで偽装する場合には、適切に検出できないことがあり、今後検討する必要がある。

●ルールの同一性の欠如

(注4) 各パラメータは実際には0から $L = 5000$ (e の場合は1)まで変化させたが、値が大きいたまはほぼ一定であったため、変化が観測された部分のみを示す。

Fig. 9 Average false positive and negative rates ($\bar{\alpha}$ and $\bar{\beta}$) after 30 steps over 50 trials changing L_h , L_t or e for three methods

Fig. 11 Transition of the average credibility R_i over 50 trials when $r_i(0) = 0$ using ID_H2H

本診断はルールの同一性に基づいているため、たとえば Fig. 1 のラベル A に対して正しいデータが 1 だけでなく 2, 3 などもある場合や途中で正しいデータが変化した場合には、新たな仕組みが必要である。モバイルエージェントの場合では、エージェントの自律的生成と消滅により、適切なテスト用ルールを持ったエージェントだけが生き残

り検査を続ける方法が考えられる。

- ルールの修正と改ざん

IDS やエージェントが持つルールはシミュレーション開始時に固定され、途中で正常 IDS やエージェントによる修正または改ざん IDS やエージェントによる改ざんは行なわれない。ルールの修正と改ざんも取り入れ、どのような状況で適切に「自己修復」されるか調べる価値はある。

- ほかの分散的診断手法との比較

IDS の分野では分散的診断手法の議論は少ないものの、分散処理システムの故障を扱った基本問題として「ビザンチン合意問題」があり、いくつかの解決アルゴリズムが存在する¹⁸⁾。IDS の自己監視にこれらの手法も適用し、免疫型診断モデルと比較する必要がある。その結果、分散的診断手法として免疫型診断モデルの有効性がより明らかになれば、IDS の自己監視に限らず、広く分散システムの診断にも適用できるであろう。

- 実装レベルでの検証

Fig. 10 Average false positive and negative rates ($\bar{\alpha}$ and $\bar{\beta}$) after 30 steps over 50 trials changing the value of diagnostic parameters for three methods

実装レベルでの検証を考えた場合、3.2節の処理(5)で述べた「同期した診断」や2節で定めた「ネットワークの形状」などを再考すべきである。同期を取るための通信コストを考えれば、非同期のほうが好ましい。この点に関しては、すでに非同期の免疫型診断モデルも提案されており¹³⁾、このモデルで実装する必要がある。一方、ここでのネットワークモデルは、各ホストがランダムに接続され、接続数がほぼ等しいランダムグラフ(random graph)である。しかし、WWWから社会関係に至るまで多くのネットワークをよりの確に表わしているといわれるスケールフリー(scale free)¹⁹⁾があり、より現実的なこのネットワークモデルを用いなければならない。

●侵入相関分析への展開

IDSの分野では、自己監視よりもIDSの誤検知を減らすことが現状では重要な課題である。免疫型診断モデルをプラントのセンサネットに適用した例¹⁴⁾では、各センサの「出力」に対してセンサ間の「依存関係」(たと

えば圧力センサ値 P と温度センサ値 T に対して $P=cT$ など)を用いて故障箇所を絞り込む。そこで、各IDSの「ルール」を相互監視する代わりに、各IDSの「出力結果」がどれほど信用できるかをIDS相互の「関連性」によって判定することが考えられる。たとえばIDS $_i$ が警告を出力した場合、IDS $_j$ でも関連する警告があれば侵入を検知した信用できる結果として $T_{ji}=1$ とする。一方、IDS $_j$ で警告がなければ、IDS $_i$ の誤報またはIDS $_j$ の欠報の可能性を疑い $T_{ji}=-1$ とし、ほかのIDSも含めた相互評価により i と j のどちらが信用できるかを判定する。これは、複数のIDSの出力結果の相関を調べて検出精度を上げる「侵入相関分析²⁾」ともいえる。しかし、このためには、プラントの例でも十分な依存関係を見つけておかなければならない。各IDSの出力結果の関連性を見つける別の仕組みが必要となる。

●移動のメリット

免疫細胞の体内循環(移動)を参考にしてモバイルエー

ジェントを用いる方法も提案したが、現時点では診断速度において必ずしも優れていると結論できない。また、今回のエージェントの移動はランダムとしたが、さまざまな移動戦略も考えられる。引き続きモバイルエージェントを使う場合と使わない場合を比較し、明確な「移動のメリット」を見つける必要がある。

6. おわりに

本論文では、免疫型診断モデルを用いて分散型IDSの自己監視を行ない、ルールが改ざんされたIDSを分散的に検出した。分散型IDSを抽象化したシミュレーションシステム上で、三つの方法「ホスト間通信で多数決モデルを用いる方法」「ホスト間通信で免疫型診断モデルを用いる方法」「モバイルエージェントで免疫型診断モデルを用いる方法」の検出性能を比較した。多数決モデルとの比較から免疫型診断モデルの有効性を確認した。また、ホスト間通信とモバイルエージェントを比較すると、診断速度ではホスト間通信のほうが良く、診断精度ではモバイルエージェントのほうが良いことがわかった。

今後は、5章で列挙した課題に取り組む予定である。最終的には、本研究を通して、以下の点を見極めたい。

- なぜ生物の免疫系は細胞の「局所的な相互作用」と「移動」でうまく生体を防御できるのか？
- 果たしてそれらはコンピュータネットワークのセキュリティに対しても有効であるのか？

謝辞 本研究は、文部科学省の科研費若手研究(B)(15700050)、財団法人堀情報科学振興財団および21世紀COEプログラム「インテリジェントヒューマンセンシング」の支援を一部受けて行なわれた。ここに謝意を表します。

参考文献

- 1) B. Mukherjee, L. T. Heberlein and K. N. Levitt: Network intrusion detection, IEEE Network, **8**-3, 26/41 (1994)
- 2) 武田, 磯崎: ネットワーク侵入検知, ソフトバンクパブリッシング (2000)
- 3) S. Forrest, S. Hofmeyr, A. Somayaji and T. Longstaff: A sense of self for unix process, In Proc. of 1996 IEEE Symposium on Security and Privacy, 120/128 (1996)
- 4) S. Forrest, S. Hofmeyr and A. Somayaji: Computer immunology, Communications of the ACM, **40**-10, 88/96 (1997)
- 5) S. Hofmeyr and S. Forrest: Architecture for an artificial immune system, Evolutionary Computation Journal, **7**-1, 45/68 (2000)
- 6) M. Crosbie and E. Spafford: Defending a computer system using autonomous agents, In Proc. of the 18th National Information Systems Security Conference (1995)
- 7) E. Spafford and D. Zamboni: Intrusion detection using autonomous agents, Computer Networks, **34**, 547/570 (2000)
- 8) H. Nishiyama and F. Mizoguchi: Design and implementation of security system based on immune system, Software Security — Theories and Systems, LNCS, **2609**, 234/248 (2003)
- 9) 浅香緑: モバイルエージェントによる侵入検出システムのための情報収集方式, 信学論 (DI), **J81-D-I-5**, 532/539 (1998)
- 10) D. Dasgupta: Immunity-based intrusion detection systems: a general framework, In Proc. of the 22nd National Information Systems Security Conference, 18/21 (1999)
- 11) W. A. Jansen: Intrusion detection with mobile agents, Computer Communications, **25**-15, 1392/1401 (2002)
- 12) N. K. Jerne: The immune system, Scientific American, **229**-1, 52/60 (1973)
- 13) Y. Ishida: Fully distributed diagnosis by PDP learning algorithm: towards immune network PDP model, In Proc. International Joint Conference on Neural Networks, 777/782 (1990)
- 14) 石田好輝: 免疫型システムとその応用—免疫系に学んだ知能システム—, 72/100, コロナ社 (1998)
- 15) 片岡, 中田, 古田: プロセス制御におけるマルチエージェントシステムの合意形成手法, 計測自動制御学会論文集, **38**-3, 309/315 (2002)
- 16) 渡邊, 石田: 情報収集モバイルエージェントシステムにおける分散的信用度評価, 信学論 (DI), **J85-D-I-8**, 758/766 (2002)
- 17) Snort.org, <http://www.snort.org/>
- 18) 山下雅史: ビザンティン合意問題—信頼性の低い分散ネットワーク上での合意問題—, 情報処理, **32**-6, 682/693 (1991)
- 19) A. L. Barabasi, R. Albert and H. Jeong: Mean-field theory for scale-free random networks, Physica A, **272**, 173/187 (1999)

[著者紹介]

渡邊 裕 司 (正会員)

1994年名古屋大学工学部電子機械工学科卒業。96年同大学大学院電子機械工学専攻前期課程修了。98年同大学大学院電子情報学専攻後期課程修了。現在、豊橋技術科学大学助手。免疫型システム、エージェントシステムなどの研究に従事。計測自動制御学会学術奨励賞受賞。電子情報通信学会、情報処理学会、日本ロボット学会の会員。

石田 好 輝 (正会員)

1979年京都大学工学部数理工学科卒業。81年同大学大学院工学研究科修士課程修了。京都大学工学部助手、奈良先端科学技術大学院大学助教授を経て、豊橋技術科学大学教授となり現在に至る。その間、カーネギー・メロン大学、サンタフェ研究所などの客員研究員。システム制御情報学会、電子情報通信学会、情報処理学会、IEEEの会員。