情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-IOT-7 No.12
2009/10/9

実トラフィックを用いたネットワーク不正侵入検知システムのための学習データ生成支援アプリケーションの開発

小林恭平[†1]　新居学[†1]
湯本高行[†1]　高橋豊[†1]

不正アクセスへの対策として知的不正侵入検知システム（IIDS）の開発を続けている．実用的な侵入検知のためには検知対象ネットワークのトラフィックから学習用データを作成し，それを用いて IIDS の学習を行うことが望ましい．しかし，学習用データの作成のための実トラフィックへのラベリングは人の手で行わねばならず，莫大な労力を要する．そこで，本研究では実トラフィックから学習用データを効率的に作成するためのシステムを開発する．また，作成したデータを用いて IIDS-SVM を学習し，数値実験により提案手法の有効性を示す．

Development of study data generation helper application for intelligent intrusion detection system using real network traffic

Kyohei Kobayashi,[†1] Manabu Nii,[†1]
Takayuki Yumoto[†1] and Yutaka Takahashi[†1]

We have already developed an intelligent intrusion detection system (IIDS) to fight illegal network access. To detect an illegal intrusion practically, we need real world network traffic data because each IDS is used at various real environments. Therefore, we have to develop a system that aids us to generate benchmark data from real world network traffics. We show that proposed method's availability by a numeric experiment.

†1 兵庫県立大学大学院
Graduate School of Engineering, University of Hyogo

1. はじめに

近年，インターネットは接続コストの低下などにより一般的ユーザへも爆発的に普及している．また，インターネット接続に対応した端末の種類も増え，その利用目的も WWW（World Wide Web）や電子メールのみならず，オンラインショッピングやネットバンキング，オンライントレードなど多様化している．

インターネットは，高価な専用回線を使用せずに遠隔地との通信を容易に実現できる利点をもつ一方で，接続したホストはネットワーク上のあらゆるホストからのアクセスが可能となるため，悪意のある第三者やコンピュータウィルスによる攻撃や不正侵入の標的となるという問題が存在する．現在では，個人レベルでも不正アクセスを試みることが可能であり，このような不正アクセスは増える一方である．特に，インターネットを用いたビジネス分野ではこのような被害を受けることは信用問題であるため，不正アクセスへの対策は必要不可欠である．

このような不正アクセスへの対策の１つとして，ソフトコンピューティング的手法を用いてネットワークの傾向を学習し検知指標を動的に設定することにより，ネットワークの傾向に応じた柔軟で精度の高い検知を行うことのできる知的侵入検知システム（IIDS:Intelligent Intrusion Detection System）が開発されている[1)]．IIDS が検知を行うためには，事前にネットワークの傾向を学習しておく必要があり，ネットワークトラフィックに不正アクセス情報を付与した学習データを用いて学習を行う．人工的に正常な通信と不正アクセスを発生させることで学習データを作成することはできるが，そのようなデータを用いて学習を行っても IIDS が実際に運用されるネットワークの傾向を学習しているとはいえない．侵入検知を可能な限り正確に行うためには実際に IIDS が運用されるネットワークのトラフィックから学習データを作成し，それを用いて学習を行うことが望ましい．

そこで本研究では，実トラフィックから学習データを作成するシステムの開発を目的とする．作成したデータは学習データとしてだけでなく，ベンチマークとしても利用できる．また，作成したデータを用いて，サポートベクターマシンを利用する IIDS-SVM を学習し，数値実験によって提案手法の有効性を示す．

2. 学習データの作成

2.1 実トラフィックへのラベリング

IIDS が学習を行うためには，ネットワークトラフィックに不正アクセス情報を付与した

© 2009 Information Processing Society of Japan

学習データが必要となる．DARPA データ集合[⋆1]のような人工的に発生させたトラフィックには不正アクセス情報を容易に付与できるが，実トラフィックに不正アクセス情報を付与する場合，やり取りされたパケットを１つ１つ観察し，それぞれの通信が不正アクセスによるものであるかどうかを人の目で判断する必要がある．しかし，やり取りされるパケットの情報を見るだけでそのパケットが不正アクセスによって発生したものであるかどうかを判断することが困難な場合には，トラフィックログから TCP による通信を取り上げ，TCP コネクション単位に分類して観察することで，付与するラベルの判断を行う．

本研究で作成する学習データは，実ネットワークトラフィックを TCP コネクション単位に分類し，TCP コネクションごとに不正アクセスであるか否かを示すラベルを付与したものである．本稿では，不正アクセスの疑いがない TCP コネクションを “正常” なコネクション，不正アクセスの疑いがある TCP コネクションを “不正” なコネクションと呼ぶ．また，分類した TCP コネクションをリスト化したものをコネクションデータと呼ぶ．

2.2 ラベリングの流れ

実トラフィックから学習データを作成するために，トラフィックの生データをトラフィックログとして保存し，それを TCP コネクション単位に分類する．実トラフィックにはラベルは付与されていない．ラベリングは各 TCP コネクションの状態やヘッダ情報を観察し，それらの情報をもとにそのコネクションが正常であるかどうかを判断してラベルを付与する処理を繰り返し，コネクションデータに含まれる全ての TCP コネクションにラベルを付与した時点で完了する．

しかし，ネットワークにおける通信量は非常に膨大であるため，ラベリング作業は作業者にとって大きな負担となる．そのため，ラベリング作業を効率化するためのツールが必要となる．

3. 学習データ作成システム

学習データの作成は以下の手順で行われる．

手順１　トラフィックログから TCP による通信のみを抜き出し，ヘッダ情報と取得時間を読込む．

手順２　読込んだデータを TCP コネクション単位に分類し，コネクションデータを作成する．

手順３　コネクションデータに “正常” または “不正” のラベルを付与する．

このうち手順１および手順２は自動化が可能であるが，手順３のラベル付けに関しては各 TCP コネクションのコネクション状態を観察して手動でラベル付けを行わなくてはならない．そこで，実トラフィックから学習データを生成するには，トラフィックログから TCP コネクションを抽出するプログラム，および，ラベリング支援ツールが必要となる．

3.1 TCP コネクションの抽出

TCP による通信では，ある時刻において複数の TCP による通信が行われている場合でも，これらの通信の両端のホストで同一の IP アドレスとポート番号の組合せが同時に用いられることは無い[2]．そこで，トラフィックログに記録されたパケットのヘッダ情報を参照し，通信の両端の IP アドレスとポート番号の組合せが同じものを同一の TCP コネクションとみなすことでトラフィックログから TCP コネクションを抽出することができる．

また，ラベリングの際の判断基準とするため，コネクションデータには各コネクションが TCP におけるスリーウェイハンドシェイクの手続きをどこまで完了したかを示す情報や端末間で交換されたパケットのサイズの情報を付加する．

3.2 ラベリング

ラベリング作業において，TCP コネクションを１つ１つ見るだけではその通信が正常であるかどうかは判断することが困難である．しかし，調査行為や DoS 攻撃では短時間に大量のコネクション確立要求が発生するなどログに特徴が残る[3]．そこで，コネクションを同一ホストに関するものや同じホスト間の通信に関するものなどの基準でグループ分けし，グループごとにまとめて観察を行うことがラベリングに有効である．そこで，図 1 に示す開発したラベリング支援ツールでは，TCP コネクションをグループ分けし，そのグループを見やすく表示する機能を実装している．

さらに，大量の TCP コネクションを観察する必要のあるラベリング作業を効率化するためにラベリング支援ツールをマルチユーザ化する．

3.2.1 ラベリング支援ツールのマルチユーザ化

ラベリング作業では大量の TCP コネクションを人が観察する必要があるため，効率化のためには複数人によるラベリングしかない．ラベリングを行うコネクションデータの時間範囲を分担して複数人でラベリングを行えば，作業者の人数に比例して効率は向上する．しかしながら，ラベリングの目的は正確な学習データを作成することであり，この方法は作業者の力量により作成される学習データの信頼性にばらつきが発生すると考えられる．

⋆1 DARPA Intrusion Detection Data Sets
http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html

© 2009 Information Processing Society of Japan

Vol.2009-IOT-7 No.12
2009/10/9

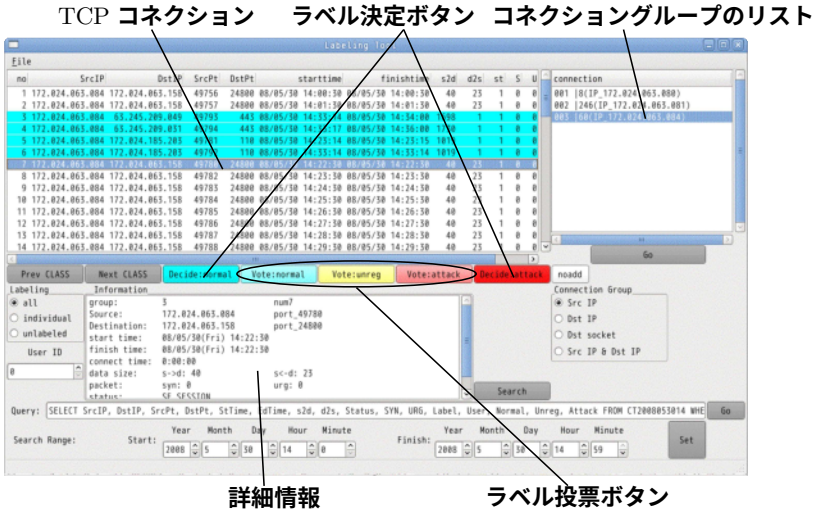


図 1　ツールの表示画面
Fig. 1　Screenshot of Labeling tool

作業者の力量の差を考慮しつつ学習データの信頼性を上げるには，付与するラベルに複数人の意見を反映させることが有効である．そこで，ラベリングシステムをマルチユーザ化するにあたり，ラベル決定に複数人の意見を反映させる仕組みを導入し，その上でラベリングの効率を向上させるシステムを考える．

3.2.2　ラベル投票システム

付与するラベルの信頼性を上げるため，複数人の意見を反映させて付与するラベルを決定する仕組みを導入する．

付与するラベルの投票を全ての TCP コネクションに対して行う場合，ラベルの信頼性は向上するものの，ラベルの投票に参加する者全員が全てのコネクションを観察し投票を完了するまでラベルは決定されないため，ラベリングの完了までに非常に時間がかかる．ラベリング作業では，TCP コネクションを関連性のあるグループに分け，そのグループを観察することでそれらの TCP コネクションが不正アクセスによって発生したものであるかどうかを判断し，付与するラベルを決定する．その際，観察対象のグループに含まれる TCP コネクションの数が十分な量であり，かつ不正アクセスに見られる特徴的な傾向が顕著に現れている場合などは，他人の意見を仰ぐまでもなく容易にラベルを決定することができる．そのような場合に際しては，投票システムの有無はラベルの信頼性にほとんど影響を及ぼさず，逆に投票システムを用いることで作業効率の悪化を招いてしまう．

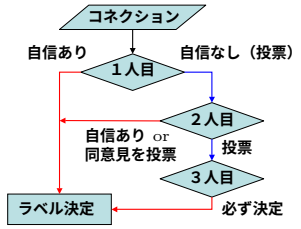


図 2　投票システムのフロー
Fig. 2　Vote System's Flow

そこで，全ての TCP コネクションへのラベル決定に投票システムを利用するのではなく，作業者が自信を持ってラベルを決定することができないものに対してのみ投票システムを利用することで，ラベリング作業の効率を確保しつつ信頼性の向上を図る．

また，投票を行う際，ラベルの決定に必要な得票数が多いとその分だけラベル決定までに時間がかかる．そこで図 2 に示すようなフローを提案し，最大３人の投票で確実にラベルを決定する事ができるように設計を行う．

4.　システムの評価実験

本研究で構築したシステムの評価を行った．

4.1　評 価 実 験

本研究で構築したシステムの評価を行うため，投票システムを使用せずに１人でラベリングを行った場合と投票システムを使用して３人でラベリングを行った場合で実際に学習データを作成し，ラベリング結果の比較を行った．また，それらを用いて学習した IIDS-SVM でコネクションデータの識別を行った．

© 2009 Information Processing Society of Japan

表 1　ラベリング作業の所要時間
Table 1　The time required for Labeling

投票システム不使用	投票システム使用	
4時間 45 分	6時間	（作業者 A）
	3時間 30 分	（作業者 B）
	2時間 30 分	（作業者 C）

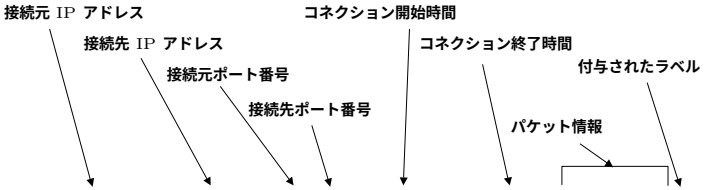
実験に使用したデータは，2008 年 11 月 12 日 0 時から１時間の間に兵庫県立大学姫路書写キャンパスのネットワークの出入り口を流れたトラフィックのログである．このトラフィックログよりコネクションデータを作成し，ラベリングを行った．また，作成したデータを用いて IIDS-SVM を学習し，先に用意しておいた別の時間（2008 年 10 月 23 日 0 時〜3 時）のトラフィックログより作成した識別用のコネクションデータを１時間分ずつ識別させた．

4.2　実験結果および考察

表 1 にラベリングの完了までにかかった所要時間を示す．投票システム不使用は１人でラベリングを行った場合，投票システム使用は３人でラベリングを行った場合であり，時間は各人の所要時間である．図 3 に投票システムを使用せずにラベリングを行ったコネクションデータの一部，図 4 に投票システムを使用してラベリングを行ったコネクションデータの一部をそれぞれ示す．

表 1 について，３人でラベリングを行った場合，３人の作業時間に大きな偏りができた．この理由は，作業を早く始めた者に作業が集中してしまい，他の作業者は主にラベルが確定されなかったコネクションへの投票のみを行うだけとなっていたためであった．このことから，３人でラベリングを行った場合，一番先に作業を始めたものにかかる負担は実質的にはほぼ１人でラベリングを行った場合と同じであり，全体的に見た場合，コネクションの投票待ちの分だけ余分な時間がかかるため，１人でラベリングを行った場合よりも効率は低下してしまった．

図 3，図 4 は投票システムを使用しなかった場合と使用した場合においてラベリング結果が異なった部分を示している．図中の “パケット情報” は左から接続元ホストが送信したパケットのサイズ，接続元ホストが受信したパケットのサイズ，コネクション状態，SYN パケット数，URG パケット数を表す．また，コネクション状態は “1” が正常にコネクションが開始・終了したことを示し，“3” は SYN パケットによるコネクション確立要求のみがなされた状態であることを示す．ラベルは “1” が正常，“-1” が不正を表す．図中では正常ラベルが付与された TCP コネクションを青色，不正ラベルが付与された TCP コネクション




```
123.204.141.67 150.12.190.18 2735 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.19 2736 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.20 2738 25 1226416916 1226416917 220 267 1 1 0 -1
123.204.141.67 150.12.190.21 2737 25 1226416916 1226416956 220 267 1 1 0 -1
123.204.141.67 150.12.190.22 2740 25 1226416916 1226416917 220 267 1 1 0 -1
123.204.141.67 150.12.190.23 2739 25 1226416916 1226416956 220 267 1 1 0 -1
123.204.141.67 150.12.190.24 2741 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.25 2742 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.26 2943 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.27 2944 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.28 2945 25 1226416916 1226416916  48   0 3 1 0 -1
123.204.141.67 150.12.190.29 2946 25 1226416916 1226416916  48   0 3 1 0 -1
                                   ⋮
```

図 3　投票システム不使用の場合のラベリング結果（一部）
Fig. 3　Labeled connetion data by voting system nonuse

を赤色で示している．

このようなコネクション群が見つかった場合，接続元が全て同一のホストであること，接続先ポート番号が全て同一であること，接続先が同一ネットワーク上の複数のホストであること，コネクションの大半が接続要求のみの TCP コネクションであることからこれらの TCP コネクションはアドレススキャンにより発生したコネクション群であると判断し，図 3 のように全てに不正ラベルを付与することができる．しかし，アドレススキャンを行う際に，SYN パケットの送信先にホストが存在する場合，接続の確立が行われることがあり，その場合同じ接続元 IP アドレスを持つコネクション群には正常に開始・終了した TCP コネクションが一部含まれる．複数人でラベル付けを行った場合，先に一部の正常に開始・終了した TCO コネクションのみを見て正常であると判断し，正常ラベルを付与してしまったため，図 4 のような誤ったラベリングがなされたコネクションデータが作成されてしまった．

これらの作成した学習データを用いて IIDS-SVM を学習させ，別に用意した識別用のコ

© 2009 Information Processing Society of Japan

Vol.2009-IOT-7 No.12
2009/10/9

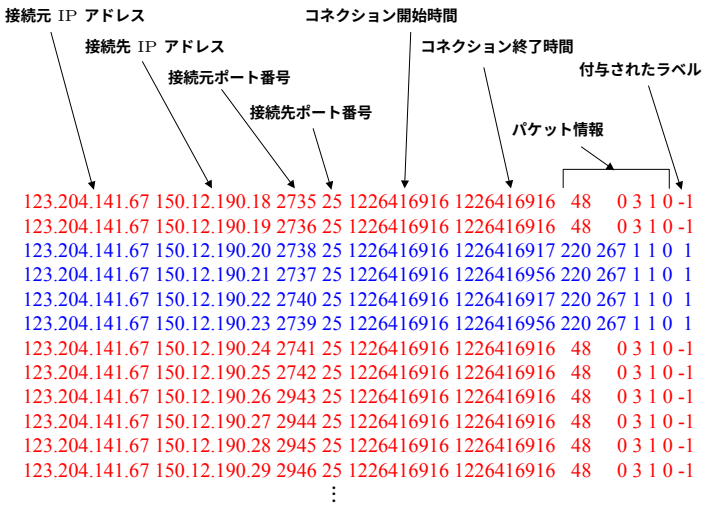


図 4　投票システム使用の場合のラベリング結果（一部）
Fig. 4　Labeled connetion data by voting system nonuse

ネクションデータの識別を行った．投票システムを使用せずに作成したデータで学習させた場合の識別結果を表 2 に，投票システムを使用して作成したデータで学習させた場合の識別結果を表 3 に示す．$R_{\mathrm{Missdetect}}$ は正常な通信を不正と判断した場合，R_{Undetect} は不正な通信を正常と判断した場合を示す．なお，IIDS-SVM のカーネル関数にはガウシアンカーネルを用いて，σ は 0.001 とした．

先に述べたように，作成された学習データは投票システムを使用した方が不正アクセスの判断基準が甘くなり，表 2，表 3 に見られるように，投票システムを使用した場合は投票システムを使用しなかった場合よりも不正な通信を正常と判断した割合が増加した．

このように，複数人で同一のコネクションを観察した場合，不正アクセスによるコネクション群の一部に正常な通信に見えるコネクションが含まれているときに，先に正常な通信に見える部分を見た作業者がその部分に正常ラベルを付与してしまい，ラベルの信頼性が低下していることがわかった．

表 2　投票システム不使用の場合の識別結果
Table 2　Identification result by voting system nonuse

時間	$R_{\mathrm{Missdetect}}$ [%]	R_{Undetect} [%]
0	0.07 (7/ 9549)	0.65 (3/ 460)
1	0.06 (4/ 7198)	0.74 (3/ 405)
2	0.02 (1/ 6210)	6.13 (1465/23905)
3	0.00 (0/ 1746)	0.00 (0/ 75)

表 3　投票システム使用の場合の識別結果
Table 3　Identification result by voting system use

時間	$R_{\mathrm{Missdetect}}$ [%]	R_{Undetect} [%]
0	0.01 (1/ 9549)	3.48 (16/ 460)
1	0.01 (1/ 7198)	0.74 (3/ 405)
2	0.00 (0/ 6210)	6.26 (1497/23905)
3	0.00 (0/ 1746)	8.00 (6/ 75)

5. 問題点への対応策

5.1 複数ユーザによるラベリングの問題点

システムの評価実験より，本研究で作成したシステムでは複数人でラベリングを行った場合，各人の作業開始時刻の違いにより１人に作業が集中してしまう問題があり，そのことが作業効率の悪化を招いてしまった．そのため，複数人でラベルを行うにあたっては，作業の負荷を各人に均等に分散させる仕組みが必要となる．

また，複数人が同一のコネクションを観察する際，大きなグループの一部分だけしか観察せずにラベリングを行ったユーザがいた場合，それによって学習データの信頼性が低下する可能性がある．この問題は，十分な教育によってラベリングに参加する者の力量を上げることにより回避できるが，多人数を一定のレベルまで教育することは難しく，それだけの人材をラベリングのたびに確保するのは困難である．しかし，ユーザに提供する情報をある程度絞り，その場で見るべき情報のみを提供することができれば，ユーザがラベリングの判断を行いやすくなり，ある程度の知識があるユーザであれば正しくラベリングが行えると考えられる．

5.2 ロードバランサ

前節で述べたように，本研究で作成したシステムの問題を改善するためには，ラベリング作業の負荷を各作業者に均等になるように分散する仕組みと作業者に与える情報を絞り込む機能が必要となる．そこで，２つの解決策を実現する仕組みとして，ラベリング対象の TCP コネクションをラベリング作業者に振り分けるロードバランサを実装することを考える．ラベリング対象の TCP コネクションとは，具体的には以下の条件を満たすものとする．

- ラベルが未確定（投票待ちを含む）
- 自分がラベルの投票を行っていない
- 他ユーザが観察中でない

© 2009 Information Processing Society of Japan

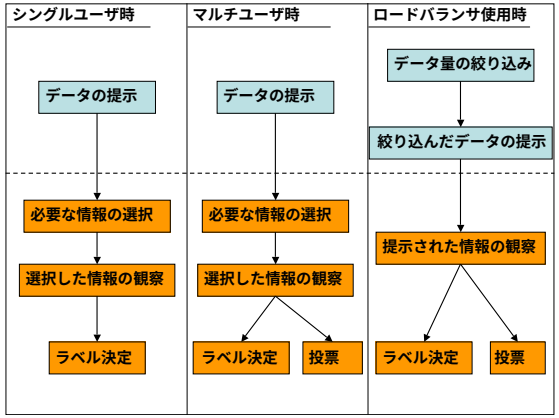


図 5　ラベリングのフロー
Fig. 5　Labeling Flow

また，ラベリングの判断を行いやすくするため，対象となる TCP コネクションと関連性のある TCP コネクションを検索し，それらの TCP コネクション群をユーザに提示する．このとき提示する TCP コネクション群の量は少数にし，作業者がそれらへのラベリングを完了してから別の TCP コネクション群を提示する．こうすることで作業者に与えられる情報は絞り込まれたものとなり，観察中の TCP コネクションに関連性のない TCP コネクションのような必要のない情報を含まないため，作業者は必要な情報のみを観察でき，付与するラベルの判断がしやすくなる．図 5 にラベリングをシングルユーザで行うとき，マルチユーザで行うとき，ロードバランサを使用してマルチユーザで行うときのそれぞれのフローを示す．図中の破線より上の部分はツール側で行う処理を，下の部分はユーザが行う処理を示す．

このような機能を持つロードバランサを導入することで前章で述べた問題点は解決することができると考えられる．さらに，ラベルの判断の難しいコネクション群を力量のある作業者に振り分けるなど振り分けアルゴリズムの最適化を行うことができれば，より作業効率を向上させることができ，作成される学習データの信頼性向上も実現できると考えられる．

6. おわりに

本研究では，実トラフィックに各通信が正常であるか不正であるかを示すラベルを付与することで IIDS の学習に用いる学習データを作成するためのシステムを開発した．また，開発したシステムを用いて実際に学習データを作成し，作成したデータの解析および IIDS-SVM に学習させての評価を行った．

ラベリングの効率向上と作成される学習データの信頼性向上を考え，ラベリング支援ツールをマルチユーザ化し，投票システムを使用せずに１人でラベリングを行った場合と投票システムを使用して３人でラベリングを行った場合の両方で学習データを作成し，データの評価を行った．その結果，３人でラベリングを行った場合でも各人の作業開始時刻の違いにより１人に作業が集中し，人海戦術によるラベリングの効率向上が実現できなかった．また，ラベルの信頼性についても３人でラベリングを行った場合では複数人が同一の TCP コネクションを観察できる仕組みが逆にラベルの信頼性低下の原因となってしまった．しかしその一方で，ラベルの判断をしづらい TCP コネクションに対しては，３人でラベリングを行った方が複数人の意見を取り入れた判断が行えたため，投票システムが効果的であったといえる．

そこで，評価実験で明らかになった問題点を解決するためにロードバランサをラベリング支援ツールに新たに導入することを提案した．今後の課題としては，提案したロードバランサを実装し，問題点を解決した学習データ作成システムを開発することが挙げられる．

参　考　文　献

1) 石堂裕章：実際のネットワークトラフィックを用いた知的侵入システムの性能検証, システム制御情報学会 研究発表講演会講演論文集, pp.293-294 (2006).
2) Information Sciences Institute University of Southern California: RFC 793 - Transmission Control Protocol, pp.5-6 (1981).
3) Chris McNab，鍋島公章：実践ネットワークセキュリティ監査，オライリー・ジャパン，pp.52-70 (2005).

© 2009 Information Processing Society of Japan