KDD CUP 99 Data Set を用いた
異なる学習データによる機械学習アルゴリズムの評価

高原 尚志† 櫻井 幸一†

†九州先端科学技術研究所
814-0001 福岡県福岡市早道区 百道浜２丁目 1-22 福岡 SRP センタービル 7F
tkhara@unii.ac.jp, sakurai@csce.kyushu-u.ac.jp

あらまし 近年，インターネット上のコンピュータがマルウエアによる攻撃を受け問題となっている．通常のウイルス対策ソフトはシグネチャ型と呼ばれる個別のウイルスの特徴に注目して検知する方式を採用しているため，新種のマルウエアに対応できない場合がある．そこで本研究では，KDD CUP 99 Datasetを用いて機械学習の各手法によりマルウエアの侵入検知を試みる．その際，再現率（Recall）を評価基準として，手法だけでなく，学習データ及び検知されるマルウエアといった3者の様々な組み合わせを試行する．これにより，過去のマルウエアのデータを学習データとして，新種のマルウエアの侵入を検知することができる汎用的な手法を考察する．

Evaluation of Machine Learning Algorithms
with Different Learning Data Using KDD CUP 99 Data Set

Hisashi Takahara† Kouichi Sakurai†

†Institute of Systems, Information Technologies and Nanotechnologies (ISIT)
Fukuoka SRP Center Building 7F
2-1-22, Momochihama, Sawara-ku, Fukuoka, 814-0001, JAPAN
tkhara@unii.ac.jp, sakurai@csce.kyushu-u.ac.jp

Abstract Today, computers in the Internet are attacked by malware. As software for protection from malware adopts a signature system for each individual attack, it cannot offer protection from a new type of malware. In this paper, we try detecting malwares by machine learning methods with KDD CUP 99 dataset. Then we attempt to discover the best combination of methods, learning data and detected malware, focusing on recall rates. Finally, we predict exploitation of new intrusion detection system for an attack from a new type of malware, using past malware behavior data.

1 はじめに

1.1 背景と動機

近年，インターネットの普及により，各種マルウエアの攻撃への対策が大きな課題となっている．マルウエアとは悪意のあるソフトウエア（Malicious Software）の略称[6]である．ネットワーク上のコンピュータがマルウエアに感染すると，個人情報などの機密情報が知らないうちに流出したり，コンピュータ自体が使えなくなったりするなど大きなダメージを受ける．現在，ネットワーク上のコンピュータをマルウエアの攻撃から守るための多くのシステムが開発されているが，常に新たなマルウエアが開発されるため，すべてのマルウエアに有効な対策を見出すのは困難である．

1.2 既存研究と課題

マルウエアの攻撃に対応する方法としては，マルウエアの攻撃を検知してユーザに教える攻撃検知型（IDS=Intrusion Detection System）と検知した場合に，その脅威を排除する攻撃防御型（IPS=Intrusion Prevention System）がある．また，マルウエアの攻撃パターンを把握して，その攻撃パターンに合致したデータを得た場合に，マルウエアと判断するシグネチャ型（signature）と正常通信のパターンを把握して（定義して），これ以外の通信を攻撃と判断するアノマリ型(anomaly)がある[11]．更に，ネットワーク上の通信データを監視して，通常通信と比較することによって攻撃を検知するネットワーク型システム（NIDS=Network Intrusion System）とホスト上の異常データを監視して異常を検知するホスト型システム（HIDS=Host Intrusion System）がある[11]．

IDS 及び IPS の研究で扱われる代表的なデータセットとして，KDD CUP 99 Data Set[14][20]がある．KDD CUP は，ACM の研究会である Knowledge Discovery and Data Mining[22]が毎年主催するデータマイニングの大会である．その中で，1999 年大会（KDD CUP 99）で侵入検知用に提供されたデータセットが，KDD CUP 99 Data Set であり，米カリフォルニア大学により提供されている．KDD CUP 99 Data Set は，約 500 万件(4,898,930)のフルセットとその中から 10%を抽出した約 50 万件(494,021)の 10%データセットからなり，学習用データセットと評価用データセットがある．評価データには各インスタンスに攻撃または正常通信のラベル付けがされたデータセット(corrected.gz)とラベル付けがされていないデータセット(kddcup.testdata.unlabeled.gz ，kddcup.testdata.unlabeled_10_percent.gz ，kddcup.newtestdata_10_percent_unlabeled.gz)があり，ラベル付けされたデータセットには，各インスタンスに攻撃の種類がラベル付けされている．学習用データには，22 種類の攻撃と正常通信が含まれており，評価データには，38 種類の攻撃と正常通信が含まれている．この内，17 種類は，学習データにはない攻撃であり，学習データにはあるが，評価データにはない攻撃も 2 種類ある．

Safaa ら[12]は，KDD CUP 99 Data Set に対し，機械学習の様々な手法を適用して，検知率と学習時間を検証した．Safaa らの論文には，各攻撃に対する検知率と学習時間の一覧が示されており，検知に対して有効な手法を見出すことができるが，学習データとしてラベル付けされた 10%データの中のすべてのインスタンスを用いているが，個別の攻撃を用いた場合や正常通信と攻撃通信の割合を変えた場合など学習データのバリエーションを意識した検知率については触れられていない．そのため，学習データとして，どの攻撃をどのような割合で用いると検知率にどのような影響が出るのかを検証する必要があると考えられる．

1.3 検証と貢献

本稿では，マルウエア対策の中でも既存の攻撃パターンをもとに攻撃を検知するシグネチ

ャ型検知に注目し，シグネチャ型検知の課題である新種の攻撃への対応について，学習データと機械学習の手法の適切な組み合わせによる解決を目指す．学習データに，学習する攻撃や正常通信の割合などを変えた様々なものを用い，これと機械学習の手法との関係を検知率に基づいて検証することによって，新たな攻撃にも対応できる組み合わせを見出す．これにより，従来のシグネチャ型検知の課題である「新種の攻撃への対応」と「定義ファイルのダウンロード」に関する課題を解決するための一助となると考える．

1.4 関連研究（との比較）

本稿では，学習データとして，正常通信と攻撃通信の様々な割合を扱うが，このような場合，インスタンス数が多い方に結果が偏るという不均衡問題[7][8][15][18]が発生する可能性がある．このような場合には，割合の多い方の結果になる傾向にあるといういわゆる不均衡データ問題が知られている．このような場合，データに重み付けをする方法とそれぞれのデータ数を調整する方法がある．重み付けをする方法は，判断を誤った場合に，そのデータに加重（重み）を掛けて調整する．これに対して，データ数を調整する方法では，少ない方のデータを増やすオーバーサンプリングと多い方のデータを減らすアンダーサンプリングがある．このようにして，データ数の不均衡をなくすのである．

本稿では，学習データに対して，後者（データ数を調整する）の方式，特に normal（正常通信）のデータ数を減らして個別の攻撃のデータ数にそろえるアンダーサンプリングを行ったが，学習データを１対１（attack:normal）にするだけではなく，１対2，１対3のデータセットも用意して，結果の違いを比較検討した．

2 評価実験

2.1 概要

シグネチャ型の検知システムは，学習した攻撃の検知には威力を発揮するが，学習した攻撃以外の攻撃，特に新種の攻撃の検知に対しては課題があるとされている．本実験では，学習する攻撃や攻撃及び正常通信の割合などを様々に変えて，各攻撃に対する検知率を測定する．測定結果を比較検討することにより，シグネチャを作成した攻撃以外の攻撃，特に新種の攻撃についても検知することができるシグネチャ型の攻撃検知システムのための手法と学習データの有効な組み合わせを明らかにする．

2.2 識別手法

機械学習[12][15]の識別手法には，様々なものがある．本稿では，その中でも比較的評価が高い手法である SVM[1][4][13]と Random Forest[2][4][13]を用いて識別を行った．SVM は，サンプルをグルーピングして各グループからのマージンが最大となる識別境界線（超平面）を描き，新たなデータの識別を行う手法で，Random Forest は，特徴値をランダムに選択した複数の決定木（弱識別器）の結果をもとにデータを識別する代表的なアンサンブル手法である．今回の実験では，SVM のカーネルとして，exponent=1 の Poly kernel を用い，Random Forest では，選択される特徴値の数を 5，決定木の数を 3 とした．

2.3 評価基準

機械学習の評価基準としてよく用いられる指標に，正例／負例のデータを正しく識別した割合である正解率（Accuracy），正例と判断したものの内，真に正例であったものの割合である精度（Precision），正例全体の内，正例と判定された割合である再現率（Recall）がある[15][17]．また，それぞれの指標には一長一短があるため，これを総合的に評価する指標として，F-measure や AUR などもある[15][17]．

本稿では，攻撃の検知率を最優先に考え，すべての攻撃の内で攻撃と識別されたインスタンスの割合である Recall を評価基準として考える．Recall の値が高いほど，マルウエアの侵入

を検知することができているということになる.

ただし, Recall が高い場合, 攻撃ではないにも関わらず攻撃と判断してしまう誤検知の割合が高くなる可能性もある. しかし, 本稿では, 誤って攻撃と判定する誤検知率を低く抑えるために, 真の攻撃を検知できず感染被害が拡大するということを避けるため, 誤検知が高くても, 攻撃を攻撃と判断できるものの評価(Recall)をよりよくすることとした.

2.4 実験の流れ

評価実験は次の手順で行う.

(手順1)対象のデータセットに対して評価のための前処理を施す

(手順2)機械学習の手法により識別実験を行う

(手順3)(手順2)の結果をもとに, 評価基準(Recall)を求めて, 比較検討する

2.5 実験環境

・データセット

データセットには, KDD CUP 99 Data Set の他にもマルウェア対策研究人材育成ワークショップ(MWS)が提供するもの[10][19]や現在マサチューセッツ工科大学のリンカーン研究所から提供されている"DARPA Intrusion Detection Data Sets"[21], NEW BRUNSWICK 大学にから提供されている NSL-KDD Data Set[5][9]があるが, 本実験では, KDD CUP 99 Data Set を用いる. 学習データとして, kddcup.data_10_percent.gz を用い, 評価データセットとして kddcup.data_10_percent.gz と corrected.gz の 2 つを用いてその結果を比較検討した.

・前処理

データの前処理としては, KDD CUP 99 Data Set では, 1 インスタンスあたり, 41 の特徴値が用意されているが, この内, 値がテキストとなる3つの特徴値を除外した 38 の特徴値を識別のために用いた. この際, 学習データ, 評価データともに正規化を行い, 0 から 1 の範囲にデータを圧縮した.

・学習データ

また, 学習データセットの normal と attack の比率が不均衡であったため, インスタンスの数が多い normal について, アンダーサンプリングを施し, インスタンス数を削減した. この際, attack:normal 比率が1対1のもの, 1対2のもの, 1対3のものを作成した.

学習の対象となる攻撃には, 学習データ kddcup.data_10_percent.gz の 22 種類の攻撃[3][12][16]の内, 1,000 前後のインスタンス数の攻撃 6 種類を用いた. 具体的には, back(2,203 インスタンス), ipsweep(1,247 インスタンス), portsweep(1,040 インスタンス), satan(1,589 インスタンス), teardrop(979 インスタンス), warezclient(1,020 インスタンス)を用いた. back と teardrop は, DoS(Denial of Service)攻撃に属し, ipsweep, portsweep, satan は, IP アドレスやポートのスキャンなどを行う probe 型の攻撃, リモートホストからログインを見てパスワードの推測を行うなどの R2L 型の攻撃である.

本稿では, 各攻撃に対する検知率の差を測定するため, 攻撃を混合したり, 同じカテゴリに属する複数の攻撃を用いたりすることは行わず, 単独の攻撃を用いて検証を行った.

・評価データ

評価データとしては, 学習データで用いたデータセットである kddcup.data_10_percent.gz とこれとは異なるデータセットである corrected.gz を用いた. これにより, 共通の通信の場合と新たな通信の場合の検知率を比較検討した.

評価の対象とした攻撃には, kddcup.data_10_percent.gz, corrected.gz の 2 つのデータセットに存在して双方のデータセットで 1,000 以上のインスタンスを有する攻撃 4 種類を用いた. 具体的には, back, neptune, satan, smurf について評価を行った. neptune と smurf は, DoS 型の攻撃である.

評価データにおいても, 学習データと同様の理由で, 複数の攻撃を混合させず, 単独の攻撃についての検証を行った.

・評価に使用したソフトウエア

実験のソフトウエアには，Weka3.7[17][23]を用いた．Weka はニュージーランドの WAIKATO 大学で開発されたデータマイニング用ソフトウエアでオープンソースソフトウエアとして公開されている．

2.6 実験結果

全体として，評価データに学習データと同じ kddcup.data_10_percent.gz を用いた場合と異なる corrected.gz を用いた場合の結果はほぼ同じとなったが，SVM を用いた場合と Random Forest を用いた場合では結果が異なる部分があった(表 1)．以下，検知する攻撃別にその結果について述べる．

・back の検知

back の検知では，SVM では検知できず，学習データに portsweep を用い，attack:normal の比率が 1:2 及び 1:3 で Random Forest を用いた場合にのみ高い検知率を示した．更に特徴として，学習データに back を用いた場合でも Random Forest で 1:2 の比率の場合以外は，ほぼ検知することができないということが分かった．

・neptune の検知

neptune の検知では，SVM を用いた場合，protsweep，satan，teardrop を学習データとして用いた場合に高い検知率を示すことが分かった．これに対して，Random Forest を用いた場合には，satan のみが安定して高い検知率を示すことが分かった．

・satan の検知

satan の検知では，SVM，Random Forest ともに satan 自身を学習データとして用いた場合には高い検知率を示した．この他，SVM では，portsweep，teardrop 及び warezclient を用いた場合に，attack と normal の比率に関係なく高い検知率を示した．一方，Random Forest では，ipsweep と portsweep を用いた場合に高い検知率を示した．

・smurf の検知

smurf の検知では，SVM を用いた場合，satan を学習データとして用いた場合の検知率が 100%でその他の攻撃を学習データとして用いた場合が 0%となり，極端な結果となった．これに対して Random Forest を用いた場合には，ipsweep，satan 及び teardrop の検知率が，attack と normal の比率によっては低くなるものがあるものの，ほぼ高い検知率を示した．

表1 検知結果の比較

・学習データと評価データが同じ場合

Random Forest:

Recall		back	neptune	satan	smurf
back	1:1	0.0%	0.0%	0.0%	0.0%
	1:2	89.9%	0.0%	0.0%	0.0%
	1:3	0.0%	0.0%	0.0%	0.0%
ipsweep	1:1	0.0%	19.1%	95.0%	0.0%
	1:2	0.0%	0.0%	91.4%	99.9%
	1:3	2.6%	0.0%	98.9%	100.0%
portsweep	1:1	81.0%	19.1%	95.5%	0.0%
	1:2	90.6%	19.1%	91.4%	0.0%
	1:3	91.0%	23.1%	92.6%	0.0%
satan	1:1	0.5%	100.0%	100.0%	18.8%
	1:2	0.2%	100.0%	100.0%	0.1%
	1:3	1.3%	99.9%	92.6%	100.0%
teardrop	1:1	0.0%	0.0%	0.0%	0.0%
	1:2	0.0%	0.0%	0.4%	99.9%
	1:3	0.0%	96.2%	87.7%	100.0%
warezclient	1:1	16.6%	50.3%	64.3%	0.0%
	1:2	19.8%	0.0%	2.2%	0.0%
	1:3	1.3%	0.5%	0.1%	0.0%

SVM:

Recall		back	neptune	satan	smurf
back	1:1	0.6%	0.0%	0.0%	0.0%
	1:2	2.2%	0.0%	0.0%	0.0%
	1:3	2.1%	0.0%	0.0%	0.0%
ipsweep	1:1	0.3%	0.0%	0.0%	0.0%
	1:2	0.0%	0.0%	0.1%	0.0%
	1:3	0.0%	0.0%	0.1%	0.0%
portsweep	1:1	0.4%	100.0%	97.7%	0.0%
	1:2	0.0%	99.9%	98.4%	0.0%
	1:3	0.0%	99.9%	98.4%	0.0%
stan	1:1	0.0%	100.0%	99.6%	100.0%
	1:2	0.0%	100.0%	99.6%	100.0%
	1:3	0.0%	100.0%	99.6%	100.0%
teardrop	1:1	0.0%	99.8%	97.2%	0.0%
	1:2	0.0%	99.8%	97.2%	0.0%
	1:3	0.0%	99.8%	97.2%	0.0%
warezclient	1:1	0.2%	0.7%	97.0%	0.0%
	1:2	0.2%	0.4%	97.0%	0.0%
	1:3	0.2%	0.4%	97.0%	0.0%

・学習データと評価データが異なる場合

Random Forest

Recall		back	neptune	satan	smurf
back	1:1	0.0%	0.0%	0.0%	0.0%
	1:2	97.2%	0.3%	0.0%	0.0%
	1:3	0.0%	0.0%	0.0%	0.0%
ipsweep	1:1	0.0%	69.9%	99.5%	0.0%
	1:2	0.0%	0.0%	89.2%	99.9%
	1:3	5.0%	1.4%	99.9%	100.0%
portsweep	1:1	74.8%	70.8%	99.7%	0.0%
	1:2	97.1%	70.7%	95.7%	0.0%
	1:3	97.2%	72.7%	75.5%	0.0%
satan	1:1	0.0%	100.0%	100.0%	100.0%
	1:2	0.0%	99.9%	100.0%	0.3%
	1:3	0.0%	99.8%	99.9%	100.0%
teardrop	1:1	0.0%	0.0%	0.0%	0.0%
	1:2	0.0%	0.1%	0.0%	99.9%
	1:3	0.0%	97.5%	89.0%	99.8%
warezclient	1:1	1.6%	21.4%	33.2%	0.0%
	1:2	0.0%	0.0%	0.4%	0.0%
	1:3	0.0%	0.0%	0.1%	0.0%

SVM

Recall		back	neptune	satan	smurf
back	1:1	2.6%	0.0%	0.0%	0.0%
	1:2	3.1%	0.0%	0.0%	0.0%
	1:3	3.1%	0.0%	0.0%	0.0%
ipsweep	1:1	0.0%	0.0%	0.1%	0.0%
	1:2	0.0%	0.0%	0.1%	0.0%
	1:3	0.0%	0.0%	0.1%	0.0%
portsweep	1:1	0.0%	99.8%	99.8%	0.0%
	1:2	0.0%	99.8%	99.8%	0.0%
	1:3	0.0%	99.8%	99.8%	0.0%
satan	1:1	0.0%	100.0%	99.9%	100.0%
	1:2	0.0%	100.0%	99.9%	100.0%
	1:3	0.0%	99.8%	99.9%	100.0%
teardrop	1:1	0.0%	99.7%	98.9%	0.0%
	1:2	0.0%	99.7%	98.9%	0.0%
	1:3	0.0%	99.7%	98.9%	0.0%
warezclient	1:1	0.0%	0.1%	97.4%	0.0%
	1:2	0.0%	0.0%	97.4%	0.0%
	1:3	0.0%	0.0%	97.2%	0.0%

3　考察

学習データの攻撃と正常通信の割合の違いによる検知率の変化を検証したが，Random Forest では，割合によって検知率が大きく変わることがあり，特に，ある割合のときのみ特別に高い検知率を示すこと（teardrop を用いて neptune や satan を検知する場合など）があったが，SVM の場合には，割合による検知率の変化は見られなかった．この結果を見る限り，SVM を用いる場合には，均衡データ問題を意識しなくても，学習データを用いることができるということになるが，更なる検証が必要と考える．

また，評価データの違いによる検知率の変化も検証したが，今回の実験では，学習データと評価データが同じ場合も異なる場合もほぼ同じ結果となった．この結果を見る限り，評価データによる検知率の差異は少なく，評価データが別途得られない場合には，同じデータセットを学習データと評価データの両方に用いても，一定の結果を得ることができるという結果となった．

最後に，今回検知を試みた 4 種類のすべての攻撃について，ネットワーク型検知において，機械学習の手法と学習データの組み合わせを工夫すれば，自分以外の攻撃を検知することができることが分かった．

4　まとめ

本稿では，マルウエア検知に関して，用いるデータセットについて，KDD CUP 99 Data Set を中心に解説を行った後，学習データと機械学習の手法の最適な組み合わせを明らかにするため，評価実験を行い，その結果について報告するとともに考察を加えた．その結果，機械学習の手法を用いることにより，ネットワーク型の IDS（NIDS）において，シグネチャ型検知の拡張として，自分以外の攻撃を検知することができる可能性を示すことができた．

今回の実験では，機械学習の手法として SVM と Random Forest のみを用いたが，今後更に手法を増やし，マルウエア検知に関するよりよい組み合わせを検証して行く予定である．また，今回の実験では，SVM のカーネルや Random Forest の決定木の数などのパラメータを固定したが，今後，パラメータを変えて実験する予定である．攻撃について，学習データ，評価データともに単独の攻撃での検証を行ったが，今後攻撃を混合させた場合の検知率についても測定する予定である．更に，今回の実験では，特徴値について特に抽出せずに，数値のデータすべてを用いたが，今後特徴値の選別についても検討する予定である．測定基準についても，今回はマルウエア検知を第一に考え，Recall を採用したが，精度（Precision）や正解率（Accuracy），それらの総合指標である F-measure や AUR など，その他の基準につ

いても測定する予定である.

謝辞

この研究の一部は，総務省による「国際連携によ るサイバー攻撃の予知技術の研究開発」の支援を受けています.

参考文献

[1] Vladimir N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Transactions on Neural Networks*, Volume 10, No.5, pp.988-999, (Sep. 1999)

[2] Leo Breiman: Random Frorests, Machine Learning, Volume 45, No.1, pp.5-32, (2001)

[3] Zeon Trevor Fernando, I. Sumaiya Thaseen and Ch. Aswani Kumar: Network Attacks Identification Using Consistency Based Feature Selection and Self Organization, *Proc. First International Conference on Networks & Soft Computing*, pp.162-166, (2004)

[4] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang and Dmitri Alperovitch: Support Vector Machine and Random Forests Modeling for Spam Senders Behavior Analysis, *Proc. IEEE Global Communications Conference Exhibition & Industry Forum* (GLOBECOM 2008), (2008)

[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani: A Detailed Analysis of the KDD CUP 99 Data Set, *Proc. The 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications* (CISDA 2009), (2009)

[6] 市野将嗣, 市田達也, 畑田充弘, 小松尚久:トラフィックの時系列データを考慮したAdaBoost に基づくマルウェア感染検知手法, 情報処理学会論文誌, Volume 53, No.9 pp2062-2074, (2012)

[7] Mohmed Bekkar and Dr. Taklit Akrouf Alitouche: Imbalanced Data Learing Approaches Review, *International Journal of Data Mining & Knowledge Management Process* (IJDKP), Volume 3, No.4, pp.15-33, (2013)

[8] Asieh Mokarian, Ahmad Faraahi, Arash Ghorbannia Delavar: False Positives Rduction Techniques in Intrusion Detection Systems-A Review, *International Journal of Computer Science and Network Security* (IJCSNS), Volume 13., No.10, pp.128-134, (2013)

[9] S. Revathi, Dr. A. Malathi: A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, *International Journal of Engineering Reseach & Technology* (IJERT), Volume 2, Issue 12, pp.1848-1853, (2013)

[10] 神薗雅紀, 畑田充弘, 寺田真敏, 秋山満昭, 笠間貴弘, 村上純一:マルウェア対策のための研究用データセット ～ MWS 2013 Datasets ～, 情報処理学会シンポジウムシリーズ, Volume 2013, CSS2013(MWS2013), (2013)

[11] Mrs. Anshu Gangwar, Mr. Sandeep Sahu: A Survey on Anomaly and Signature Based Intrusion Detection System(IDS), *Journal of Engineering Research and Applications*, Volume 4, Issue 4, pp.67-72, (2014)

[12] Saffa o. Al-mamory, Firas S. Jassin: Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set, *Journal of Babylon University, Pure and Applied Sciences*, Volume 21, No.8, pp.2663-2681, (2013)

[13] Md. Al Mehedi Hasan, Mohammed

Nsser, Biprodip Pal and Shamim Ahmad: Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS), *Journal of Inteligent Learning Systems and Applications*, Volume 6, No.1, (February 2014)
[14] Martina Troesch and Ian Walsh: Machine Learning for Network Intrusion Detection, *Final Report for CS 229*, pp.1-5, (2014)
[15] 荒木雅弘, フリーソフトではじめる機械学習入門, 森北出版株式会社, (2014.3)
[16] Kriangkrai Limthong: Performance of Interval-Based Features in Anonaly Detection by Using Machine Learning Approch, *International Journal of Machine Learning and Computing*, Volume 4, No.3, (June 2014)
[17] Payal P. Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin: Preprocessing and Classification in WEKA Using Different Classifires, *International Journal of Engineering Research and Applications*, Volume 4, Issue 8, pp.91-93, (August 2014)
[18] 井手剛, 入門機械学習による異常検知ー R による実践ガイドー, コロナ社, (2015.3)
[19] マルウェア対策研究人材育成ワークショップ 2014(MWS2014)
<http://www.iwsec.org/mws/2014/about.html>(accessed 2015-08-15)
[20] KDD Cup 1999 Data
<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>(accessed 2015-08-15)
[21] MIT Lincoln Laboratory DARPA Intrusion Detection Evaluation
<http://www.ll.mit.edu/ideval/data/>(accessed 2015-08-15)
[22] Sig KDD bringing together the data mining, data science and analytics community
<https://kddcup2015.com/information.html>(accessed 2015-08-15)
[23] Weka 3 - Data Mining with Open Source Machine Learning Software in Java
<http://www.cs.waikato.ac.nz/ml/weka/index.html>(accessed 2015-08-15)