ニューラルネットワークによる情報通信量の予測

第22回 信号処理シンポジウム
2007年11月7日〜9日（仙台）

ニューラルネットワークによる情報通信量の予測

Prediction of Data Traffic by Neural Network

松島稔 † 平野晃宏 † 中山謙二 †
† 金沢大学大学院 自然科学研究科 電子情報工学専攻

Minoru MATSUSHIMA† Akihiro HIRANO† Kenji NAKAYAMA†
†Division of Electronics and Computer Science
Graduate School of Natural Science and Technology, Kanazawa Univ.
E-mail: {hirano,nakayama}@t.kanazawa-u.ac.jp

アブストラクト 小規模ネットワークを対象とした，近年増え続けている不正通信を自動で検出するためのシステムを提案する．ニューラルネットワークを用いて通信量の予測を行ない，予測結果から大きく外れた通信を異常と判断する．この精度が十分でない．そこで入力形式や出力形式を工夫することで精度の向上を検討した．曜日情報や時間情報は周期関数とみなして Sin と Cos に分解して入力し，出力の形式は連続値で表現するのではなく量子化を行ないバイナリで表現をすると精度の改善がみられた．

Abstract This paper proposes an automatic detection system to an illegal trafic that keeps increasing in recent years intended for a smallscale network. The network traffic is forecasted by using a neural network (NN). The communication that comes off greatly is judged to be abnormal from the forecast result. To improve the accuracy, the the input and the output of the NN is examined. The combination of periodic functions Sin and Cos as the day of a week and the time, and a binary expression as the forecasting output results in a better accuracy.

1 はじめに

近年コンピュータネットワークにおいて不正アクセスやワーム等の不正通信が増加し，より高度なセキュリティの必要性が高まっている．比較的よく使われるセキュリティ機器 (システム) にファイアウォール (FW, 図 1) やアンチウィルスソフト，不正アクセス検出システム (IDS, 図 2)[1]-[2] などが挙げられるが、これらだけでは万全とは言えない．そこで本研究では未知の攻撃にも対応できるような異常検出型のシステムを提案する．ニューラルネットワークに通信量を予測をさせ，予測値と実際の通信量の誤差が大きかった場合にネットワークが異常であると判断し，管理者に知らせるシステムである．

現段階では通信量の予測が十分な精度となっていないので，この予測精度を上げるための検討を行なう．第 2 章では、既存のセキュリティ機器について説明する。第 3 章で本システムを提案する。そして第 4 章で実測データを用いたシミュレーションにより、予測精度を確認する。

2 既存のセキュリティ機器

2.1 FW

FW はネットワークのセキュリティ機器において最もポピュラーなものである．その機能は外部から不正なアクセスを遮断して，ネットワークのセキュリティを守る．また内部から外部へと向かう，許可していないサービスを利用した通信を禁止することもできる．しかし，基本的に FW は通信のデータがどのホストから送信されてどのホストが受信するのか，そのときのサービスは何かという情報を表す IP アドレスとポート番号の 2 つだけを見て通信を許可するかを判断する．この 2 つの情報は簡単に偽装することができるうえ，一見正常な通信の中に悪意のある命令を忍ばせることですり抜けることができるので FW を導入すれば全ての不正アクセスを防げるわけではない．

2.2 アンチウィルスソフト

アンチウィルスソフトはネットワークのセキュリティを守るものというよりは，コンピュータ自身に

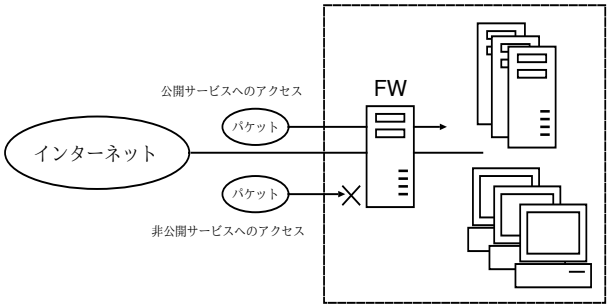

図 1: ファイアウォールの動作

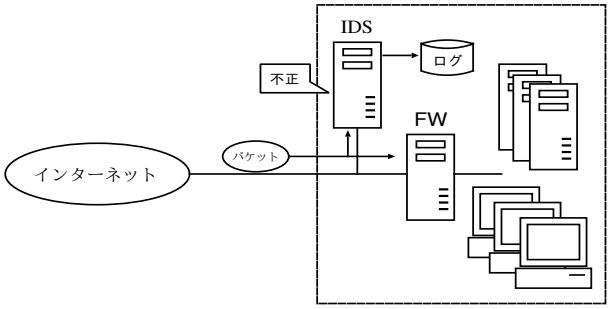

図 2: IDS の動作

ウィルスが感染するのを防ぎ，感染してしまった場合にはウィルスを駆除してシステムを復旧するためのソフトである．コンピュータがウィルスに感染すると，ネットワークにつながっている他のコンピュータを攻撃したり，感染させようとするのでコンピュータを守るということは間接的にネットワークを守るということになる．ただしアンチウィルスソフトは全てのウィルスの感染を防ぐことができるわけではなく，あらかじめウィルスの特徴を集めたデータベースに登録されているウィルスしか予防，駆除をすることができない．従ってデータベースに登録される前の最新のウィルスやデータベースの更新を行なっていると防ぐことができない場合がある．

2.3 IDS

IDS にはホスト型とネットワーク型の 2 種類のタイプがある．ホスト型の IDS はサーバにインストールすることによって，サーバのプロセスに異常がないか，ファイルに改竄がないかなどを監視し異常があった場合に管理者に知らせるシステムである．ネットワーク型はアンチウィルスソフトと同様にウィルスやクラッカーの攻撃の特徴を集めたデータベースを持っていて，ネットワークに流れているパケットを監視する．攻撃のパターンと一致した通信を見つけた場合には直ちに管理者にメール等で知らせる．ただし IDS は攻撃を発見しても管理者に知らせるだけで，FW のように通信を遮断しない．このときに通信を遮断するようなシステムの場合は侵入防止システム (IPS) という．

2.4 既存システムの問題点

一般的に上記のような機器，ソフトを用いてネットワークのセキュリティを守るが，全ての攻撃を防ぐことはできない．アンチウィルスソフトや IDS のように不正検出型システムでは未知の攻撃に対応できないのに対して，ウィルスや攻撃のパターンは毎日新しいものが増えている．ネットワークのセキュリティをより強固にするためには，未知の攻撃にも対応できるような異常検出を導入する必要があると考えられる．

3 提案する異常検出システム

3.1 異常検出までの流れ

異常検出までの流れは以下のようになる．

- 通信量の測定
- 通信量の予測
- 予測結果と測定結果の比較
- 異常の検出

3.2 通信量の測定

ネットワークに通信量を測定するためのサーバを設置する．サーバの設置場所はネットワークや検出対象によって異なるが，いずれにしてもなるべく多くのホストからの通信を観測できる所に設置するのが望ましい．本研究では図 3 のような構成のネットワークの中にある NAT サーバに snort[3]-[4] をインストールして測定を行なった．snort とはオープンソースのネットワーク型 IDS で，自分で管理者に知らせる攻撃のルール等を自由にカスタマイズできる．この機能を利用してサーバが観測できる全てのパケットのヘッダ情報を記録するようにルールを追加した．また IDS による不正検出とニューラルネットワークによる異常検出を組み合わせることも可能である．なお，ここでいう通信量とはネットワークに流れるパケット量のことである．

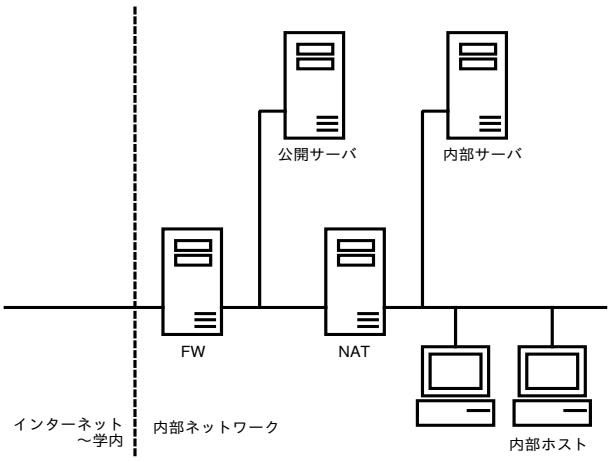

図 3: ネットワーク構成図

3.3 通信量の予測

上記で測定した過去の通信量等をニューラルネットワークに入力して現在の通信量を予測させる．隠れ層が 1 層の 2 層形ニューラルネットワークである．隠れ層及び出力層で用いるユニットの活性化関数はハイパーボリックタンジェントを用いて，学習についてはバックプロパゲーション法を行なった．

3.4 入力データ

ニューラルネットワークには以下の 10 種類の情報を入力する．なお過去 24 時間の通信量とは，25 時間前から 1 時間前までに流れた通信量のことである．これを 3 時間毎に区切ってその平均を入力している．

- 過去 24 時間の TCP の通信量
  TCP によって送受信された通信量である．
- 過去 24 時間の UDP の通信量
  UDP によって送受信された通信量である．
- 過去 24 時間の SSH アクセスによる通信量
  SSH によって送受信された通信量である．
- 過去 24 時間のメールの送受信による通信量
  メールサービス (SMTP, POP2, POP3, IMAP, POP3S) によって送受信された通信量である．
- 過去 24 時間のドメイン名解決による通信量
  DNS サーバがドメイン名から IP アドレスを求める際の通信量である．
- 過去 24 時間のウェブアクセスによる通信量
  ウェブアクセス (http, https) によって送受信された通信量である．
- 過去 24 時間のネットニュースの通信量
  ネットニュースによって送受信された通信量である．
- 過去 24 時間の NetBios
  NetBios によって送受信された通信量である．
- 平日/休日情報
  その 3 時間が平日か休日かを表す情報である．平日であれば 1，休日であれば-1 となる．日にちの変わり目の場合は 3 時間の内，2 時間が属する日の値表す．
- 曜日情報
  その日が何曜日かを表す情報である．7 ビットにそれぞれ曜日を対応させるバイナリ表現にする方法と，7 日間の周期関数とみなし，Sin と Cos に分解して連続値を入力する [5]2 つの方法を検討する．日にちの変わり目についてはバイナリの場合は両方の曜日を 1 にする．連続値の場合は中央である 2 時間目の属する曜日を Sin と Cos に分解する．
- 時間情報
  曜日情報と同様に 24 時間の周期関数とみなし，Sin と Cos に分解して連続値で入力する．ただし曜日情報をバイナリ表現にする場合は，2 つのビットが 1 になっている位置などで時間情報に近いものを表現できるので時間情報は入力しない．

3.5 出力データ

出力データは 1 時間前から現在までに流れた通信量の予測値である．図 4 に 1ヵ月間の通信量の実測値を示す．

出力ユニットを複数用意してそれぞれをバイナリ表現にすることで量子化をする．通信量が 0 のときは全ての出力を-1 にし，通信量が一定量増える毎に出力を 1 にするユニットを増やしていく．今回は通信量が 5000 増える毎に 1 にする出力を増やしていき，90000 パケットを越えたら全ての出力が 1 になるようにした．

また通信量や電力のなどの予測時に用いられる，出力に線形ニューロンを 1 つ用意して結果を連続値で表現する方法 [6] についても検討を行なう．通信量そのものを出力しようとすると値が大き過ぎるため学習がなかなか進まない．しかし図 4 のような通信量

の場合，単純に最大値で線形正規化を行なうと，一部の大きな値のせいで出力パターンがうまく分散されない．そこでデータセットの内 90%が分類される 90000 パケットで線形に正規化する．

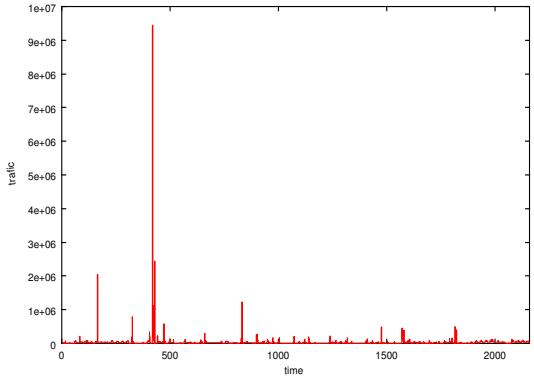

図 4: 3ヵ月間の通信量

表 1: 出力データの形式

予測通信量	出力形式
0 - 4999	0000000000000000000
5000 - 9999	1000000000000000000
10000 - 14999	1100000000000000000
⋮	⋮
85000 - 89999	1111111111111111110
90000 -	1111111111111111111

4 シミュレーション

4.1 シミュレーション条件

図 3 のようなネットワーク構成にホストが 20 台程度，サーバが 16 台設置された環境において，通信量の予測のシミュレーションを行なった．通信量のデータは 2006 年 4 月 1 日から 2006 年 6 月 30 日までの 3ヵ月間の実測データを用意して，その内最初の 2ヵ月を学習データに用いて，残りの 1ヵ月をテストデータとした．学習回数は 100 万回，学習係数は 0.00003，結合荷重の学習アルゴリズムにはバックプロパゲーション法を用いた．またシミュレーションを行なった各方式のパラメータは次の通りである．

4.2 評価方法

整数で表現される実際の通信量，連続値の出力値とバイナリの出力値はそのままでは結果を比較することができない．そこで前者の値もバイナリ表現に変換してやり，実際の通信量の値と予測値で何ビットずれているのかを比較することで表かを行なう．

表 2: 各方式のパラメータ

方式	曜日形式	出力形式	隠れ層数	出力層数
(a)	連続値	連続値	10	1
(b)	バイナリ	連続値	10	1
(c)	連続値	バイナリ	19	19
(d)	バイナリ	バイナリ	19	19

4.3 シミュレーション結果

シミュレーションの結果は表 3 のようになった．出力形式に注目すると連続値の (a)(b) とバイナリの (c)(d) を比較すると，学習時の誤差についてはバイナリ形式よりも連続値形式の方が小さくなっていることが図 5 よりわかる．しかし，予測の正答率については誤差なしから誤差 2 ビット以内まで全てバイナリ形式の方が正答率が良いことが確認できた．誤差なしは 1-5%，誤差 1 ビット以内では 9%，2 ビット以内では 8-10%の向上である．一方曜日情報については，連続値で入力した方が正答率が改善されていることが多いが，(a) と (b) の誤差なしを比較したときだけはバイナリの方が 1.25%良い結果となった．

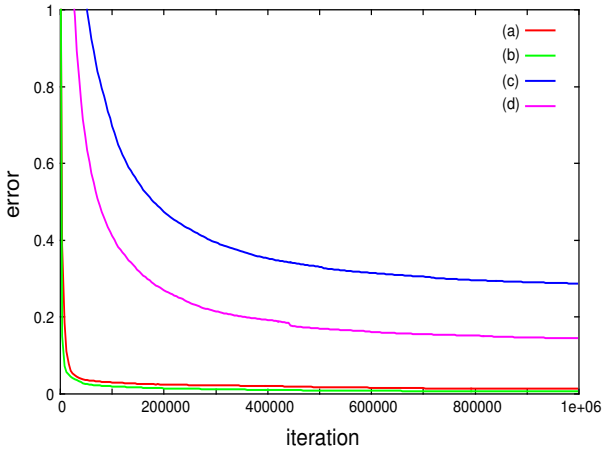

図 5: 学習誤差

5 まとめ

階層型ニューラルネットワークを用いて小規模ネットワークの通信量の予測を行なった．数値を予測する場合によく用いられる出力層を線形ニューロンに

表 3: 各方式の正答率

方式	誤差なし [%]	1 ビット以内 [%]	2 ビット以内 [%]
(a)	16.81	38.06	51.13
(b)	18.06	36.11	48.47
(c)	21.53	47.08	59.86
(d)	19.72	45.83	58.89

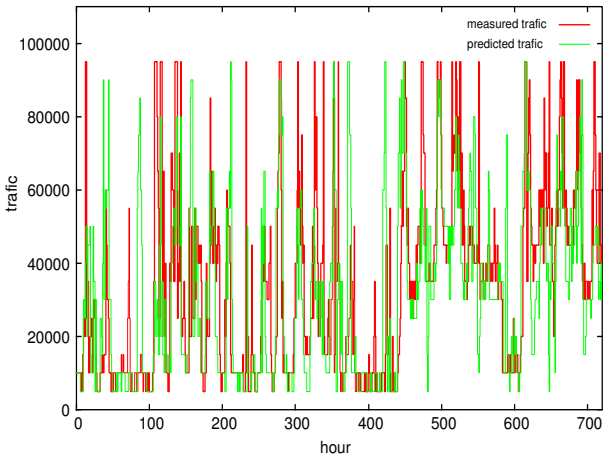

図 6: 方式 (c) の予測結果

する方法よりも，出力値を量子化してバイナリで表現した方が良い結果となることがわかった．

しかし，現段階では一番良い (c) 方式でも誤差なしに予測できる確率は 21.53%，誤差 2 ビット (15000 パケット) 以内でも 59.86%である．予測が難しい原因として，環境が小規模ネットワークのため各ユーザーの行動が反映されやすく，例えば思いつきで動画を見たり大きなファイルをダウンロードをするだけでもネットワークには普段よりも大きな通信量となってしまうことがあるということが考えられる．今後，不正通信を自動で検出するために実測値と予測値との間にしきい値を設けて，誤差がそれ以上になったら管理者に連絡するシステムを想定しているので，このままでは誤検知が多くなってしまうのでさらなる改善が必要である．

また snort 本来の不正検出の機能と今回提案したニューラルネットワークを応用した異常検出を統合したシステムを構築することも課題として挙げられる．

参考文献

[1] 竹森 敬祐，三宅 優，田中 俊昭，笹 瀬巌, “IDS ログから算出される情報エントロピー,” 情報処理学会 CSEC 研究会, Vol.2004, No.54(20040521) pp. 31-36

[2] 安藤 類央，武藤 佳恭, “ニューラルネットワークを用いた学習型 NIDS の開発,” 情報処理学会マルチメディア通信と分散処理研究会報告, Vol.2002, No.12(20020214) pp. 145-150

[3] 日吉 龍, “IDS 入門,” 技術評論社, 2004.

[4] 渡辺 勝弘, 鹿田 幸治, “snort2.0 侵入検知,” ソフトバンクパブリッシング, 2004.

[5] J. Nuno Fidalgo and Manuel A. Matos, “Forecasting Portugal Global Load with Artificial Neural Networks” *Proc. ICANN 2007, International Conference on Artificial Neural Networks* , Porto, Portugal, Sept. 2007

[6] P.Cortez, M.Rio, P.Sousa and M.Rocha “Topology Aware Internet Trafic Forecasting Using Neural Networks” *Proc. ICANN 2007, International Conference on Artificial Neural Networks* , Porto, Portugal, Sept. 2007