社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138 (49)
2009－CSEC－ 44 (49)
2009／3／6

統計的解析による不正アクセスの検出

小島 俊輔 †, 中嶋 卓雄 ††, 末吉 敏則 ‡

† 八代工業高等専門学校情報電子工学科 †† 東海大学産業工学研究科 ‡ 熊本大学自然科学研究科

DoS(Denial of Service) 攻撃からサーバを防御するためには，サーバまたはIDSにおいてアクセスフィルタリング策を講じる必要がある．しかし，一般には到達パケットが攻撃パケットかそうでないかを区別するためのフィルタリングを行うことは困難である．そこで本研究では，不正アクセスが行われているか否かを早期に検出するための仕組みについて，統計的な手法を用いた検討を行っている．特に本稿では，送信先ポート番号を確率変数として χ 二乗を定義し，本校に到達する全パケットについて解析を行った．その結果，χ 二乗手法を用いた場合において，1) 確率変数としてはソースIPアドレスより送信先ポート番号が有効であることがわかり，2) 偽装されたソースIPアドレスやDDoS攻撃に対して不正パケットの検出ができることが分かった．

DDoS 攻撃，統計解析，IDS，χ 二乗手法

Extraction of Anormaly Access using Statistical Analysis

Shunsuke OSHIMA † Takuo NAKASHIMA †† Toshinori SUEYOSHI ‡

† Department of Information and Electronic Engineering,
Yatsushiro National College of Technology

†† School of Industrial Engineering, Tokai University

‡ Graduate School of Science and Technology, Kumamoto University

To defend DoS (Denial of Service) Attacks, the access filtering mechanism is adopted on the end servers or the IDS (Intrusion Detection System). The difficulty to define the filtering rules lies where normal and anomaly packets have to be distinguished in incoming packets. The purpose of our research is to explore the early detective method for anomaly accesses based on statistic analysis. In this paper, we define the chi-square analysis, and then analyze the amount of incoming packets focusing on the destination port number. As the results, we were able to extract the following features for the chi-square analysis. Firstly, the destination port number was more suitable scale of the packet characteristics in chi-square method than the soruce IP address. Secondly, the chi-square method using destination port number can detect the DDoS attack and spoofed source IP address.

DDoS Attack, Statistical Analysis, IDS, chi-square method

1 はじめに

インターネットプロトコル自体の設計上の問題やサーバソフトウェアのセキュリティホール，サーバ自体の設定ミスといった問題は，無差別なDoS/DDoS(Distributed DoS)攻撃を生み出す原因となっており，これらの攻撃によって発生する被害は，サーバの停止や組織の情報漏洩といった深刻な事態を引き起こす要因となっている．これらのDoS/DDoS攻撃パケットは，クライアントへのサービスを提供するHTTP, DNS, SMTP[3) 4) 5)]といったいわゆるwell-knownポートと呼ばれるポートに集中しており，攻撃者はサーバを破壊するためのセキュリティホールや踏台となるような設定ミスを持ったサーバがないかを常に探している．

通常，これらのDoS/DDoS攻撃からサーバを守るためには，攻撃パケット列のパターンを基にフィルタリングルールを構築し，ファイアウォール等でブロックする必要があるが，実際，既にいくつかのファイアウォール製品においてはポートスキャン

攻撃などを検出する機能を有したものがある．これらの機能は，ソースIPアドレスが偽装されているかどうかに関係なく，機械的に攻撃者のIPアドレスを収集し，そのIPアドレスを元にした基本的なフィルタリングルールを構築することで実現している．しかし，攻撃パケットのソースIPアドレスが偽装されていた場合は通常のアクセスかそうでないかを機械的に見分けることは非常に困難である[6) 7)]．

本研究の目的は，ファイアウォールを通過する全パケットについて，パケットの持つ特徴を抽出して統計処理を行うことで，不正アクセス時の統計量の挙動を明らかにし，それによって不正アクセスの検出を行うことである．研究で使用したパケットは，実際に八代高専で動作しているファイアウォールに到達した全パケットのログデータである．これらのデータは一旦，学内に設置されたsyslogサーバに転送・保管しており，このデータを用いて今回の解析を行っている．

現在までの研究成果として，我々はファイアウォールを通過するパケットのソースIPアドレスに着目し，この値を確率変数としたχ二乗やエントロピーといった統計値を求め，その結果から不正アクセスを検出できることを示している[1) 2)]．そこで今回は新たに送信先ポート番号に着目し，これを確率変数としたχ二乗による不正アクセスの検出を試みた．これにより，DoS/DDoS攻撃やIPアドレスのスキャンといった不正アクセスが可能なことを示す．

本論文は，以下のように構成している．まず，第2節で実験環境，第3節でχ二乗手法や実験方法にいて述べ，実験結果については第4節で示した．第5節においては結論と今後の方針について述べることにする．

2 実験環境

今回用いた実験環境を図1に示す．ファイアウォールでは基本的に拒否や許可に関係なく，到達したすべてのパケットのログを取ることにしており，それらはすべて学内に設置されたsyslogサーバに転送される．今回は、このsyslogサーバに保管されたログを元に統計処理を行うこととした．syslogサーバのログは，1つの通信の開始から完了までを1行として構成しており，通信の完了時刻の他，送信元や送信先のIPアドレスやポート番号，通信の継続時間，送受信パケット数やバイト数などが記

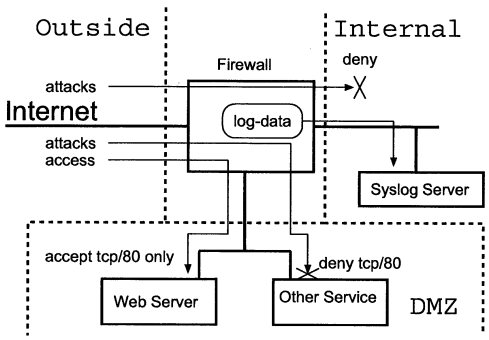

Fig. 1 パケットのログ収集に用いた実験環境

録されている．ファイアウォールでブロックされたパケットについては，通信が確立する1パケット目からパケットをブロックしているため，1つの拒否パケットの受信で1行のログが残る．一方で本校のサーバとの通信が成功した場合については，通常，通信が完了するまでの間は送信元のIPアドレスや送信先のポート番号といったパケットのもつ多くの特徴量が変化しないため，その通信で送受信したパケット数やバイト数と併せて1行のログとして記録される．そこで，今回はこのログ1行につき1件ずつ特徴量を取り出し統計処理を行うこととした．本稿では，便宜上，これを全パケットと表現することにする．

3 χ二乗値を用いた判定方法

ピアソンのχ二乗検定は，期待値と観測値の分布に違いがあるかどうかを検証するための検定方法として利用される．χ二乗値は以下の式で表される．

$$\chi^2 = \sum_{i=1}^{B} \frac{(N_i - n_i)^2}{n_i} \tag{1}$$

ここで，N_iは観測値，n_iは期待値を表す．また，N_iやn_iは比較的大きな値をとるように，到達するパケットの個数をもとに選ぶ．このとき，Nが互いに独立であれば，この結果得られた値は良く知られた$B-1$のχ二乗分布に従うことになる．

χ二乗値を計算するためには，観測を行う区間を設定する必要があるが，本稿ではこの区間をウィンドウと呼ぶことにする．我々は過去の経験[1) 2)]から，今回もウィンドウ幅を500パケットとして実験を行っている．また，ウィンドウ毎に求めた

値について，後述する重み付けを行って集計した値を，期待値 n_i として用いる．

ここで，具体的な N_i や n_i の計算方法について述べておく．ウィンドウに入った標本の数について，期待値としては 5 以上が良いとされており [8]，そのため，Feinstein[6] らによる BIN の考え方を採用して χ 二乗値を計算する．我々が以前行ったソース IP アドレスを確率変数とした χ 二乗値の計算においては，以下の式 (2) を用いて BIN への割り当て数を決定した．

$$\text{BIN}_\text{i}\text{の数} = \begin{cases} \dfrac{n}{3 \cdot 2^{b-i-1}} & (1 \leq i \leq b-1) \\ \text{残り全部} & (i = b) \end{cases} \tag{2}$$

ところで，本校の DMZ には，SMTP, DNS, HTTP といったサービスを提供するサーバのみを設置しているため，もし不正アクセスが一切ないとすれば，到達パケットの送信先ポート番号は非常に限られたものになる．今回はパケットの特徴量のうち相手先ポート番号を確率変数として統計量を算出することにしており，そのため，ウィンドウを数百，あるいは数千パケットという大きな値に設定しておくと，アクセス上位のポート番号のサンプル数が 5 を下回ることは少なくなる．そこで，上位の BIN にはすべて 1 つのポートのサンプル値のみを入れることとした．すなわち，以下の式 (3) のようになる．

$$\text{BIN}_\text{i}\text{の数} = \begin{cases} 1 & (1 \leq i \leq b-1) \\ \text{残り全部} & (i = b) \end{cases} \tag{3}$$

また，現時刻におけるパケットの度数については，Feinstein[6] らと同様に，カウント数に式 (4) による重みをつけて計算することとした．ここで，age[秒] はパケットの受信時刻から現時刻までの時間である．受信時刻が現時刻より離れるほど，カウント数に与える影響が指数関数的にゼロに近付くようにしている．我々は，攻撃を早期発見するために $halflife$ として，経験的に 1,200[秒] を選択した．

$$\exp\left(\frac{age \cdot \ln(0.5)}{halflife}\right) \tag{4}$$

一方で，期待度数の計算には，これとは逆の重みをつけて計算を行う．以下に重み関数の式 (5) を

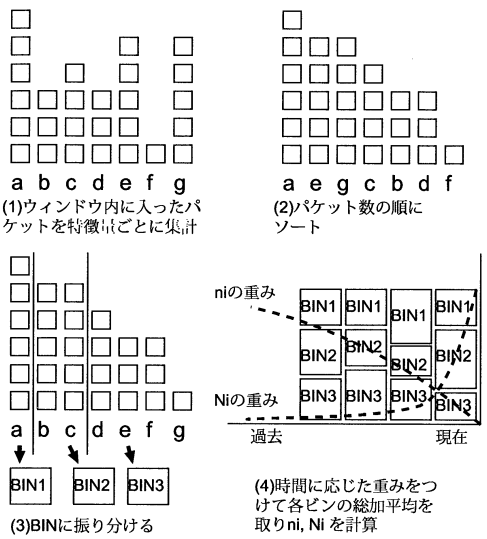

Fig. 2　χ 二乗値に必要な N_i, n_i の算出方法

示しておく．これは，観測対象が現時刻であるという観点から，過去の状態については重く，現時刻に近いほど軽くするという効果がある．

$$1 - \exp\left(\frac{age \cdot \ln(0.5)}{baseline\_halflife}\right) \tag{5}$$

ここで，$baseline\_halflife$ としては，経験的に $halflife$ の 30 倍の値である 36,000(sec) という値を用いて計算を行った．図 2 に全体の計算の流れを示すとともに，以下にその手順を記しておく．

Step 0: ウィンドウ幅となる受信パケットの数，パケットを振り分けるビンの数をあらかじめ決めておく．ビンに入れるポートの数は式 (3) に従う．

Step 1: Step0 で決めたウィンドウ幅 (パケット数) に達するまでパケットを収集し，発信先ポート番号毎に到達パケット数を集計する．

Step 2: 集計した個数をキーとしてポート毎にソートする．

Step 3: パケットの多い方から順にビンに 1 ポートずつ振り分ける．最後のビンには残りのパケットをすべて割り当てる．

Step 4: 各ビンに入ったパケット数を基に，式(4)，および式(5)に従った時間による重みをつけたものを平均し．N_i，およびn_iを算出する．

Step 5: 式1によりχ二乗値を求め，Step1に戻る．

4 実験結果

今回の実験は，2008年11月2日4:00から11月9日4:00までの計7日間，および2009年1月11日4:00から1月18日4:00までの計7日間という2つのデータについて掲載した．この期間とした理由は，本校のDNSサーバに対してDDoS攻撃やDoS攻撃が仕掛けられた形跡があったためであり，χ二乗値の挙動を知るための良いサンプルとなると考えたからである．

4.1 2008年11月2日-11月9日のログによる実験結果

まず，ファイアウォールへ到達した全パケットについて，IPアドレスの第1オクテット毎に集計したものを図3に示す．この図の太線で示した場所に，第1オクテットの広い範囲にわたって大量のパケットが到達している様子がわかる．ログの調査により，これが53/udpに対するDDoS攻撃であることが確認されたが，ここではあえてudp/tcpやicmp, ファイアウォールでの許可や拒否といった区別無しに，ファイアウォールに到達した全パケットを用いたχ二乗値を計算し，その効果を確認した．結果を図4に示す．この図はファイアウォールに到達した全パケットについて，ウィンドウごとに計算したχ二乗値を時間の経過とともにプロットしたものである．さらにビンの数をパラメータとし，χ二乗値の挙動の変化についても調査した．

これらの結果から，DDoS攻撃が行われている間，χ二乗値が非常に大きな値を示していることが分かる．この間のログを詳細に調査したところ，1秒間に数十個の攻撃パケットが到達しており，これに連動するようにχ二乗値も十数秒程度で反応し始めていることが分かった．ソースIPアドレスを確率変数としたχ二乗値と比べた場合，明らかにその特徴がはっきりと現れたため，送信先ポート番号を確率変数としたχ二乗値はDDoS攻撃の検出に対して有用であることが確認できた．

この図では，これ以外にも大きな値を示したところが二箇所あったため(矢印の箇所)，これらについてもログを調査した．1件はある大学(JP)から大量のicmpが到達した形跡があり，またもう1件は海外(BR)からの139/tcpのポートスキャンであった．どちらも数十秒程度の短い攻撃であったためか，χ二乗値は極端に大きな値とはならなかった．

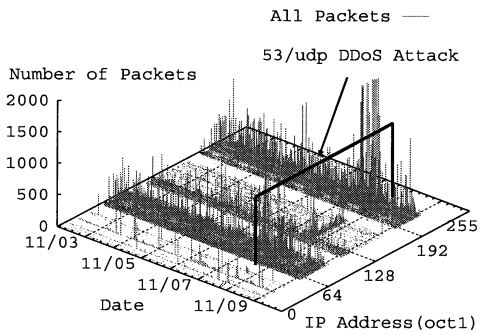

Fig. 3 ファイアウォールに到達した全パケットの分布

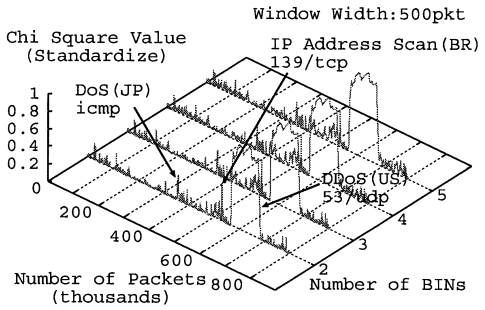

Fig. 4 送信先ポート番号を確率変数としたχ二乗値(2008年11月2日4:00-11月18日4:00)

さらに，ビン数の違いによるχ二乗値の変化についても調査した．図からも分かるとおり，ビン数を増やした場合でも，χ二乗値はほぼ同じ応答を示すことが分かった．一方で，ビン数が多くなるに従い，ノイズが多くなる様子も観測された．これは，八代高専のファイアウォールで許可された送信先ポート番号が80/tcp, 25/tcp, 53/udp, icmpの4種類と少なく，正常時であればほとんどのパケットの宛先がこの値を取るが，各ビンに1ポートずつ入れて計算しており，必要以上にビン数が増えた場合，中間のビンにノイズの基となるパケットが入るためではないかと考えられる．

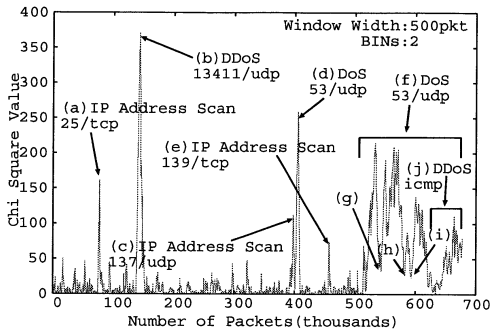

Fig. 5　送信先ポート番号を確率変数とした χ 二乗値 (2009 年 1 月 11 日 4:00-1 月 18 日 4:00)

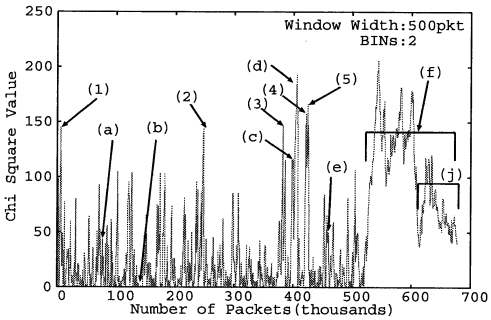

Fig. 6　ソース IP アドレスを確率変数とした χ 二乗値 (2009 年 1 月 11 日 4:00-1 月 18 日 4:00)

4.2　2009 年 1 月 11 日-1 月 18 日のログによる実験結果

図 5 に 2009 年 1 月 11 日の調査結果を示す．図中，χ 二乗値が大きな値を示した箇所についてログを詳細に調査したところ，すべてなんらかの攻撃であることが確認できている．特に，図 5 の点 (b) においては DDoS 攻撃が仕掛けられており，χ 二乗値も非常に大きな値を示した．点 (d) は DoS 攻撃，点 (a)(c)(e) は IP アドレススキャンの跡であった．このグラフがいかに鮮明に攻撃を検出しているかを示す為，比較対象として，ソース IP アドレスを確率変数とした χ 二乗値の計算結果を図 6 に併記しておく．図 5 の結果に対応する場所に，同じアルファベットの記号を付している．

この図 6 から，DDoS 攻撃を受けた点 (b) では，ソース IP を確率変数とした χ 二乗値において，ほとんど反応が表れていない．ログを詳細に調査したところ，DDoS 攻撃の IP アドレスの分布自体がそれほど広い範囲ではなく，通常のアクセスと分布が似ていたため，通常アクセスとの区別がつかなかったのではないかと推察される．また，点 (c) と点 (e) の IP アドレススキャンについては，図 6 において比較的大きな値を示しているが，点 (a) についてはあまり大きな値とはならなかった．この両者の違いについて調査したところ，通常アクセスにおいて差が見られた．つまり点 (a) では他のも比較的多い時間帯であり，そのため他のパケットのソース IP アドレスのアクセスに紛れてしまったため，χ 二乗値が小さくなったのではないかと思われる．

一方で，図 6 の点 (1) については 445/tcp やその他ポートに対する全 IP アドレスのスキャンであることが分かった．これは図 5 には現れていないが，1 秒間に 10 回以上と比較的頻度の高い攻撃であったため検出ができたと推察される．また，図 6 の点 (2)-(5) は誤検出であることが判明した．この部分のログを調査したところ，数ヶ所のクライアントから 80/tcp や 25/tcp に対するアクセスが同時に発生していた．ソース IP アドレスを確率変数とした χ 二乗値の計算では，通常，BIN1 のサイズとして数個から数十個といった範囲が割り当てられるがが，先の 80/tcp や 25/tcp のパケットがすべて BIN1 に入ってしまい，DoS 攻撃と同様の反応が出たためと考えられる．

次に，(f) の範囲について調査した．まず，ログから 53/udp に対する長期間にわたる DoS 攻撃が続いていることが分かった．そのため図 5, 図 6 どちらの場合も χ 二乗値に反応が出ている．この DoS 攻撃は (f) の区間ではまったく休むことなく続いており，χ 二乗値も大きな値を取りつづけるはずであるが，図 5 の点 (g)(h)(i) では値が小さくなっており，この原因についても調査した．まず，この区間 (f) における 53/udp の DoS 攻撃自体が 1 秒に 1 回程度であり，さらに，値が小さくなった点 (g)(h)(i) では IP アドレスのスキャンや他のポートへの攻撃が発生しており，500 パケットというウィンドウ内における送信先のポート分布が通常の状態と似通ってしまったためではないかと考えられる．そのため，ゆっくりとした DoS 攻撃については，この実験結果を見る限りでは，ソース IP を確率変数とした図 6 の方が優れていると言える．

さらに，(j) の区間については，8080/tcp に対する別の小規模な DDoS 攻撃が行われていることが分かった．これにより，図 5，図 6 どちらの場合も χ 二乗値が低下している．

この結果からソース IP アドレスより送信先ポート番号を確率変数とした χ 二乗値が IP スキャンや DDoS の検出には有効であり，一方で DoS 攻撃についてはソース IP アドレスが良い結果を返すことが分かった．ただし，ソース IP アドレスを確率変数とした χ 二乗値では誤検出も多数見受けられたため，DoS の規準として，たとえば攻撃が長時間続く場合は DoS とみなす，といった判断が必要になると考える．さらに，どちらの χ 二乗値の場合においても，複数の攻撃を同時に受けている場合は χ 二乗値が小さくなる傾向があることも分かった．

5 結論

今回，ファイアウォールへの入力パケットについて，送信先ポート番号を確率変数とする χ 二乗値の特性について調査した．その結果，ソース IP アドレスを確率変数とした χ 二乗値よりも，DDoS 攻撃や IP アドレスのスキャンの検出がうまくいくことが確認できた．送信先ポート番号を確率変数とした方法では，DoS 攻撃についても同様に検出が可能であるものの，1 秒に 1 回程度の比較的頻度の低い攻撃については，ソース IP アドレスを確率変数とした χ 二乗値のほうが信頼性が高く，様々な攻撃が重なった場合について，その傾向が強いことも確認できた．

DoS 攻撃は通常，長時間にわたることが多いため，ソース IP アドレスを確率変数とした χ 二乗値では，インパルス的な値を無視するといった運用が有効ではないかと考える．今回の実験結果を基に，DoS, DDoS 検出の判断規準をまとめると以下のようになる．

1. 送信先ポート番号を確率変数とした χ 二乗値を見て，値が大きいようであれば Dos, DDoS, IP アドレススキャンといった攻撃を検出した可能性が高い．
2. 次に，ソース IP アドレスを確率変数とした χ 二乗値を見て，もし χ 二乗値が高い値を長時間キープしているようであれば，DoS 攻撃の可能性が高い．

現在，ソース IP アドレスと送信先ポート番号に絞って調査を行ったが，他の特徴量，たとえば送受信バイト数やパケット受信間隔といった特徴量と組み合わせることで，より精度の高い検出が可能になると考える．また，パケットの持つ特徴量について，エントロピーの計算を施したものを組み合わせることで，何かしらの特徴が読み取れるのではないかと考えている．今後，特徴量の統計情報を組み合わせた実験や検討を行っていきたい．

謝辞

最後になりましたが，本実験を行うにあたり，心よくファイアウォールのログを提供頂いた八代高専の情報処理センター長米沢徹也先生とセンタースタッフに対し，心より感謝申し上げます．

参考文献

1) 小島俊輔, 中嶋卓雄, 末吉敏則：χ 二乗手法を使った不正アクセス IP パケットの特徴抽出, 電子情報通信学会技術研究報告, pp.7–12(CPSY-2008-44), Dec. (2008).
2) Nakashima, T., Oshima, S., Nishikido, Y. and Sueyoshi, T.: Extraction of Characteristics of Anomaly Accessed IP Packets by the Entropy-based Analysis, *International Conference on Complex, Intelligent and Software Intensive Systems(CISIS)*,Mar. (2008).
3) C. L. Schuba, I. V. Krsul, M. G. Kuhn, E. H. Spafford, A. Sundaram, and D. Zamboni: Analysis of a denial of service attack on TCP. *In Proceedings of the 1997 IEEE Symposium on Security and Privacy*, pp.208-223 ,1997.
4) A. Magnaghi, T. Hamada, T. Katsuyama.: A wavelet-based framework for proactive detection of network misconfigurations, *IEEE/ACM Proceedings of the ACM SIGCOMM workshop on Network troubleshooting*, pp.253-258, September (2004).
5) Zhang, Z., Fang, B., Hu, M. and Zhang, H.: Security Analysis of Session Initiation Protocol, *International Journal of Innovatie Computing, Information and Control*, pp.457-469, Vol.3, Number 2, April (2007).
6) L. Feinstein, D. Schnackenberg, R. Balupari and D. Kindred: Statistical Approaches to DDoS Attack Detection and Response, *Proceedings of DARPA Information Survivability Conference and Exposition*, Vol.1, pp.303-314 (2003).
7) Carl, G., Kesidis, G., Brooks, R. and Rai, S.: Denial-of-Service Attack-Detection Techniques, *IEEE Internet Computing*, pp.82-89, January-February (2006).
8) Siegel, S. Castellan Jr, N. J. : Nonparametric statistics for the behavioral sciences 2nd edition, McGraw-Hill, 1988.