ベイズ推定によるインターネット通信の特徴抽出と適応的制御

但馬　康宏† 　小谷　善行† 　寺田　松昭†

† 東京農工大学大学院 共生科学技術研究院 システム情報科学部門
東京都小金井市中町 2-24-16
E-mail: †{ytajima,kotani,m-tera}@cc.tuat.ac.jp

あらまし　本研究において，ベイズ推定を用いてインターネット上の通信を識別する一手法を提案し，実験による評価を行う．インターネットにおけるベイズ推定の応用としては，SPAM メールの判別や，侵入検知などの研究がすでに行われているが，本研究では，パケットをベイズ推定により分類することにより，通信プロトコルに依存せずに，ある特定の局面を識別することが可能であることを示す．この手法により，HTTP プロトコルによる通信において，一般のウェブページの閲覧と，ファイルのダウンロード中を識別する実験を行い，良好な結果を得た．さらに，SSH プロトコルによる通信においても，ログインによる操作とファイル転送の識別を行った．
キーワード　ベイズ推定，トラフィックフロー，データの特徴抽出

Characterization and control of the Internet traffic flows via bayesian method

Yasuhiro TAJIMA†, Yoshiyuki KOTANI†, and Matsuaki TERADA†

† Division of Systems and Information Technology, Institute of Symbiotic Science and Technology,
Tokyo University of Agriculture and Technology
Naka-chou 2-24-16, Koganei, Tokyo, Japan
E-mail: †{ytajima,kotani,m-tera}@cc.tuat.ac.jp

Abstract　In this study, we propose a characterization method of traffic flows on the Internet via bayesian method. In addition, a flow classification experiment has done for evaluation. There are some Internet technologies using bayesian method such as SPAM mail filtering, Intrusion Detection Systems and so on. Our method is a packet classification technique using bayesian method, and it is suitable for recognition of a specific traffic in a sequence of packets. In our experiment, we apply this technique to the HTTP and SSH protocol. We can obtain a fine classification of packets into a class of normal page view and a class of file download.
Key words　bayesian method, traffic flow, characteristics of data

1. はじめに

近年，WWW の普及や P2P ファイル交換ソフトの登場などにより，インターネットにおける個々のホストの通信量増加が著しい．しかし同時に，VPN などの通信の暗号化や SIP [5] や WebDAV などに代表される多機能なプロトコルの利用など，事前に具体的な制御内容を明記した制御方法では効率的な制御が難しい状況が増している．今後，ユビキタスコンピューティングの実現に向けて，末端ホストでの通信量はますます増加すると見込まれ，状況変化に柔軟に対処できる適応的なネットワーク制御の必要性が増加する．

従来のネットワークにおける帯域制御は，ネットワーク中継のルータにおいて，IP アドレスやポート番号などプロトコルに依存したデータを利用して，事前にポリシーを定めて制御する方法が主である．すなわち，HTTP プロトコルの帯域は最大 5Mbps などとポリシーを定める必要がある．したがって，あるプロトコルで送受信されている通信の中で，特に大規模なファイル転送中にのみ帯域を制御するなどといった，局面に応じた制御を行うことは難しい．また，事前にネットワーク利用特性を分析した上でポリシーを定めなくてはならないので，未知のプロトコルに対する適応的な制御にも限界がある．

そこで本研究では，ルータにおいて，通信をパケット単位で分類し，あるプロトコルでの通信の中の，ある特定局面を識別する手法を提案する．あわせて，実験によりその有効性を確かめる．すなわち，ルータにおける入力パケットについて，パケット長とパケットの到着時間間隔を利用したベイズ推定により，個々のパケットのクラス分けを行い，通信の特定局面の識別とした．この手法を用いて，ルータを通過する HTTP プロト

コルによる通信において，一般のウェブページ閲覧と，ファイルのダウンロード中を識別する実験を行い，良好な結果を得た．さらに，SSH プロトコルによる通信においても，ログインによる操作とファイル転送の識別を行った．

インターネットの通信におけるベイズ推定の応用は，SPAM メールの識別 [6] や，IDS における攻撃検知 [3] [1] などの研究が行われている．本研究は，ベイズ推定の新たな応用例であり，特定のプロトコルに依存しない適応的な制御手法である．

2. ベイズ推定によるパケットの分類

2.1 ベイズ推定

事象の確率変数 H, E に対して，ベイズの定理

$$P(H|E) = \frac{P(H,E)}{P(E)} = \frac{P(E|H)P(H)}{P(E)}$$

を用いて，推定を行う手法をベイズ推定と呼ぶ．ここで，H は仮説の確率変数であり，E は状況の確率変数である．H と E の取り得る値をそれぞれ $\{h_0, h_1, \cdots h_n\}, \{e_0, e_1, \cdots e_m\}$ としたとき，ある状況 $E = e_0$ が観測されたとき，仮説 $H = h_0$ が成り立つ確率は $P(H|E)$ である．学習用のサンプルから $P(E|H)$ および $P(H)$ をあらかじめ求めておく．ある状況 $E = e_0$ に対して，$P(H = h_0|E = e_0), P(H = h_1|E = e_0, \cdots, P(H = h_n|E = e_0)$ を比較することにより，H を推定できる．すなわち，

$$\frac{P(H = h_0|E = e_0)}{P(H = h_1|E = e_0)} = \frac{\frac{P(E=e_0|H=h_0)P(H=h_0)}{P(E=e_0)}}{\frac{P(E=e_0|H=h_1)P(H=h_1)}{P(E=e_0)}} = \frac{P(E = e_0|H = h_0)P(H = h_0)}{P(E = e_0|H = h_1)P(H = h_1)}$$

より，$H = h_0$ と $H = h_1$ のどちらが良いかの比較ができる．

2.2 提 案 手 法

図 1 のような環境において，本手法による分類は行われる．本研究では，ルータに入力されるパケットについて，以下の確率変数を考える．

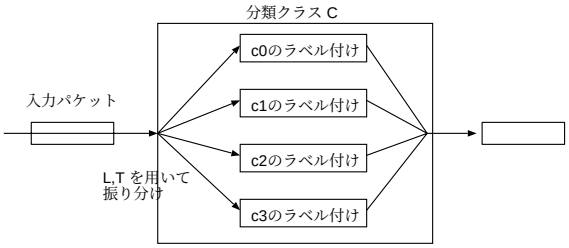

図 1 システム概要

- C : パケット分類のためのクラス
- L : パケット長のクラス
- T : 直前のパケットとの到着時間間隔のクラス

これらを用いて，

$$P(C|L,T) = \frac{P(L,T|C)P(C)}{P(L,T)}$$

として入力パケットがどの分類に属するか，すなわち C がどんな値をとるかを推定する．これらの変数は，次章の実験におけるファイル転送の識別に有効と思われるので選択した．一般に，ファイル転送では，パケットのバースト的な受信が行われるため，到着間隔の利用が有効だと判断した．それぞれの確率変数の詳細は以下の通りである．

- パケット分類のクラス C の取り得る値は，一般のパケット c_0 とファイルダウンロードのパケット c_1 など識別したい局面をそれぞれのクラスとする．
- パケット長のクラス L は，0 から最大パケット長をいくつかのクラスに分類したものである．本研究においては，Ethernet (最大パケット長 1500 bytes) を用いたネットワークを前提とし，表 1 の分類とした．

表 1 パケット長の分類

L の値	パケット長 (bytes)
l_0	1500
l_1	1000 以上 1500 未満
l_2	500 以上 1000 未満
l_3	100 以上 500 未満
l_4	100 未満

- 直前のパケットとの到着間隔 T は，識別分類を行う一連のパケットの列において，入力パケットの到着時刻と，直前のパケットの到着時刻との差をいくつかのクラスに分類したものである．本研究においては，表 2 の分類とした．

表 2 パケット到着間隔の分類

T の値	パケット到着間隔
t_0	1sec 以上
t_1	500ms 以上 1000ms 未満
t_2	100ms 以上 500ms 未満
t_3	1ms 以上 100ms 未満
t_4	100us 以上 1ms 未満
t_5	100us 未満

学習サンプルは，1 つのパケットにつき

$$(C, T, L)$$

なる形式で表され，事前に通信記録を分析し，C の値を各パケットごとに決定しておく．さらに，学習サンプルの集合より，$P(L,T|C)$ および $P(C)$ を求めておく．ルータに到着したパケットは，L および T の値を決定され，C の値を推定される．ルータより出力される際に，C の値に応じて帯域制御などの処理が行われるが，本研究では，C による分類まで取り扱うものとする．

3. HTTP 通信におけるファイルダウンロード識別

3.1 推定対象の局面とサンプルデータ

前章の分類推定を実際の HTTP 通信に対して実施した．一般のウェブブラウジング中に，プログラムファイルなどをダウンロード可能なページに行き当たることがある．このとき，ファ

イルのダウンロードを FTP などのファイル転送用のプロトコルを用いて行うウェブページならば，経路途中のルータにおいて FTP プロトコルに対して帯域制御を行うことで，ファイルダウンロード時にのみ帯域制限を行うことができる．しかし，HTTP プロトコルの中でファイルのダウンロードまで行うウェブページの場合は，ファイルダウンロードに使われているパケットを適切に選別して制御しなければ，一般のページ閲覧時まで帯域制限がかかってしまい，利用者に不便を強いる．ところが，HTTP 通信の中身については，パケットの IP ヘッダや TCP ヘッダの情報からは判別できず，パケットのデータ部分に書かれた HTTP プロトコルのメッセージをを解釈して対応しなければならない．このようなパケットのデータ内容までチェックする作業は，ルータの処理能力を著しく必要とする上，経路上のルータで行った場合は，通信の秘匿が守られない事態ともなり得る．したがって，一般的なルーティングで用いられるような，パケットのヘッダ情報などだけを利用し，帯域制御をかけられることが理想である．

本章では，HTTP 通信においてファイルダウンロード時のパケットのみを識別することを目標に，本研究の提案手法を適用し，実際のデータでその効果を測定した．確率変数のうち，分類クラス C は，以下の 2 つの値をとり得るとした．

- c_0 : 一般的なページ閲覧によるパケット
- c_1 : ファイル転送中のパケット

すなわち，c_1 に分類されたパケットのみを優先度の低いキューに入れるなどの帯域制御を行えば，一般的なウェブページ閲覧に影響を与えることなく，ファイルのダウンロード時にのみ制御が行われることとなる．このような制御を従来の制御方法で行うことは簡単ではない．確率変数 L および T については，前章の定義と同じである．

学習用サンプルデータは，以下の要領で収集した．

- 実際のウェブブラウザ (netscape) を用いて通信されるパケットを収集.
- 収集情報は，パケット到着時刻 (最小計測単位 1us)，パケットの IP ヘッダ，および TCP ヘッダの情報である．
- 収集した各パケット情報のうち，発信元 IP アドレスとポート番号，および宛先 IP アドレスとポート番号すべてが一致する一連のパケットを 1 つのフローと定義する．あるパケットの T は，そのパケットの到着時刻とそのパケットが属するフローにおける直前のパケットの到着時刻との差から表 2 を用いて定められる．ここで，あるフローの先頭のパケットは，$T = t_0$ であるとする．
- 各パケットの L は，そのパケットのパケット長から表 1 を用いて定められる．
- 各パケットが $C = c_0$ であるか $C = c_1$ であるかは，利用者が手作業でラベル付けを行う．本実験では，ダウンロードファイルの大きさは数百 KB から数 MB である．

表 3 に，準備したサンプルデータの仕様を示す.

サンプルデータは，図 2 に示されるようなテキスト形式とした．サンプルデータにおける 1 行がパケット 1 つに対応し，それぞれのデータは，以下のような意味を持つ．

- sec : パケット到着時刻の秒以上の桁
- usec : パケット到着時刻の秒以下の桁 (最小時間単位はマイクロ秒)
- pkid : パケット id (宛先ポート番号と発信元ポート番号の組合せ)
- size : パケットサイズ (bytes)
- class : パケットの分類 (0 ならば c_0，1 ならば c_1)
- src : パケットの発信元 IP アドレス
- dst : パケットの送信先 IP アドレス

これらの情報から (C, T, L) を構成し，

$$P(L, T|C)$$

および

$$P(C)$$

を求めれば，学習過程の完了である．評価用のデータも同じサンプルデータの形式であるが，“class=” の項目を利用せず推定を行う．推定は，以下の要領で行われる．

（1） 評価用データの入力パケットに対して L および T の値を求める．その値をそれぞれ $L = l_i, T = t_i$ と仮定する．

（2） $C = c_0$ として，P_0 を

$$P_0 = P(L = l_i, T = t_i|C = c_0)P(C = c_0)$$

として求める．

（3） $C = c_1$ として，P_1 を

$$P_1 = P(L = l_i, T = t_i|C = c_1)P(C = c_1)$$

として求める．

（4） $P_0 > P_1$ ならば推定は $C = c_0$ であり，$P_1 > P_0$ ならば推定は $C = c_1$ である．

3.2 実験結果

サンプル集合の組み合わせを変化させ，ファイルダウンロードの識別を行い，手で付したラベルとの合致 (識別率) を調べた．すなわち，学習サンプルを用いて $P(L, T|C)$ および $P(C)$ を算出し，その結果を用いて評価サンプルの各パケットに C のラベル付けを行う．推定により付けられたラベルが，あらかじめ手で付したラベルと一致しているパケット数を評価サンプル全体のパケット数で割り，識別率とした．

表 4 に学習サンプル数を変化させたときの識別率を示す.

表 4　識別率 (学習サンプルのパケット数の影響)

識別率 (%)		評価サンプル集合 No.		
学習サンプル集合 No.	サンプル数	5	6	7
1	4987	84.8	87.0	82.8
1 ∪ 2	9965	84.8	87.0	82.8
1 ∪ 2 ∪ 3	14945	84.8	87.0	82.8
1 ∪ 2 ∪ 3 ∪ 4	19928	84.8	87.0	82.8

学習サンプルをサンプル数 5000 から 20000 まで変化させた

表 3　サンプルデータの仕様

サンプル集合 No.	サンプルパケット数	c_0 パケットの数	c_1 パケットの数	c_1 パケットの割合	パケットの内訳
1	4987	2502	2485	49.8%	閲覧ページ数:24, ダウンロード数:1
2	4978	4978	0	0%	閲覧ページ数:22, ダウンロード数:0
3	4980	2995	1985	39.9%	閲覧ページ数:37, ダウンロード数:1
4	4983	2145	2838	57.0%	閲覧ページ数:11, ダウンロード数:1
5	32305	2622	29683	9.2%	閲覧ページ数:44, ダウンロード数:7
6	30918	26277	4641	15.0%	閲覧ページ数:129, ダウンロード数:1
7	34148	24645	9503	27.8%	閲覧ページ数:189, ダウンロード数:3

```
sec=1105474658:usec=805190:pkid=8f4d0050:size=1500:class=1:src=165.93.176.91:dst=165.93.126.48
sec=1105474658:usec=805314:pkid=8f4d0050:size=1500:class=1:src=165.93.176.91:dst=165.93.126.48
sec=1105474658:usec=805703:pkid=8f4d0050:size=1500:class=1:src=165.93.176.91:dst=165.93.126.48
sec=1105474658:usec=805826:pkid=8f4d0050:size=1500:class=1:src=165.93.176.91:dst=165.93.126.48
sec=1105474658:usec=805949:pkid=8f4d0050:size=1500:class=1:src=165.93.176.91:dst=165.93.126.48
...
sec=1105474960:usec=207896:pkid=8f8e0050:size=1454:class=0:src=219.166.163.202:dst=165.93.126.48
sec=1105474960:usec=207938:pkid=8f8e0050:size=619:class=0:src=219.166.163.202:dst=165.93.126.48
sec=1105474960:usec=208068:pkid=8f8e0050:size=1454:class=0:src=219.166.163.202:dst=165.93.126.48
sec=1105474960:usec=209045:pkid=8f8d0050:size=1454:class=0:src=219.166.163.202:dst=165.93.126.48
sec=1105474960:usec=209084:pkid=8f8d0050:size=551:class=0:src=219.166.163.202:dst=165.93.126.48
sec=1105474960:usec=209357:pkid=8f8d0050:size=1454:class=0:src=219.166.163.202:dst=165.93.126.48
```

図 2　サンプルデータの形式

が，同一の評価サンプル集合に対しては，すべて同じ識別結果が出力された．学習用サンプル集合がさらに小さなものの場合，識別結果に差が出るものと思われるが，5000 パケットのサンプルはおよそ 40 分程度のウェブブラウジングにより得られたものであり，学習用サンプル集合の構成は，十分容易であると言える．

表 5 に学習サンプル集合における c_1 ラベルを持ったパケットの割合を変化させたときの識別率の違いを示す．

表 5　識別率 (c_1 ラベルを持ったパケットの影響)

識別率 (%)		評価サンプル集合 No.			
学習サンプル集合 No.	c_1 割合	1	2	3	4
6	16%	91.5	92.8	94.2	80.8
7	28%	91.5	92.8	94.2	80.8
3	40%	91.5	92.8	94.2	80.8
4	58%	88.9	88.8	77.5	81.4
5	92%	78.6	40.2	92.2	77.2

学習サンプルにおける c_1 ラベルのついたパケットの割合が，評価サンプルにおける c_1 ラベルのついたパケットの割合に近いほど高い識別率を示し，逆にその割合が食い違うほど識別率は低くなる．特に，学習用サンプル集合，評価用サンプル集合どちらにおいても c_1 ラベルの付いたパケットの割合が高いものほど識別率が低くなる傾向がある．学習用サンプルにおける c_1 ラベルつきパケットの割合が 9 割を越えているサンプル集合 No. 5 と，評価用サンプル集合を No. 2 とした場合では，識別率が 40.2% とランダムな識別 (50%) よりも悪い結果となっている．

4. SSH 通信におけるファイルダウンロード識別

4.1 推定対象の局面とサンプルデータ

本手法による分類推定を SSH プロトコルによる通信に対しても実施した．SSH は，遠隔ホストへのログインの手段として telnet に代わる手段となったが，その暗号化の基本部分は，HTTP における安全なファイル転送においても利用されている．本研究では，HTTP プロトコルにおける実施と同様に，利用者がファイル転送に利用したパケットを c_1 と分類し，それ以外のパケットを c_0 と分類した．学習用サンプルと評価用サンプルを用いて，分類の精度を計測した．

表 6 に，準備したサンプルデータの仕様を示す．

HTTP プロトコルにおける実施との相違点は，パケットの分類がセッション単位となっている点である．すなわち，c_1 に分類されるパケットはすべて，

```
%ssh remotehost cat foo
```

なるコマンドを実行した場合にネットワークに流れるパケット

表 6　SSH サンプルデータの仕様

サンプル集合 No.	サンプルパケット数	c_0 パケットの数	c_1 パケットの数	c_1 パケットの割合	パケットの内訳
1	5948	1216	4732	79.6%	ダウンロードファイル数:3
2	4996	4354	642	12.9%	ダウンロードファイル数:4
3	5978	1249	4729	79.1%	ダウンロードファイル数:2
4	4981	4427	554	11.1%	ダウンロードファイル数:4

であり，リモートホストにログインした場合のパケットはすべて c_0 に分類される．

4.2 実験結果

表 6 に示すサンプル集合を，それぞれ学習サンプルとした場合の識別率の違いを表 7 に示す．

表 7　識別率 (SSH プロトコル)

識別率 (%)		評価サンプル集合 No.			
学習サンプル集合 No.	サンプル数	1	2	3	4
1	5948	97.0	68.0	94.7	96.3
2	4996	21.3	87.5	23.0	89.1
3	5978	80.4	18.0	97.9	61.6
4	4981	91.0	71.4	93.7	97.1

いずれも，学習に用いたサンプル集合と評価に用いたサンプル集合における c_1 パケットの割合が食い違う場合に低い識別率となっている．しかし，c_1 パケットの割合が近いサンプル集合に対しては，良好な結果を得ることができた．

5. まとめと今後の課題

本研究において，パケット長と到着間隔からインターネット通信の特徴抽出を行う一手法を示し，HTTP プロトコルの通信において，一般のウェブ閲覧とファイルダウンロードのパケット単位での識別を行った．さらに，SSH プロトコルについても同様の実験を行った．

その結果，学習に用いるサンプル集合において，ファイルダウンロードと分類されるパケットの割合が大きくなく，かつ評価サンプルにおけるその割合と近い場合に良好な識別結果を得た．今後の課題として，以下のような点がが挙げられる．

- 確率変数 T を直前パケットのみでなく，n 個前まで参照して分類を行う．直前のパケットとの到着間隔のみでは，HMM によるトラフィック解析 [1] なども研究が行われている．
- 確率変数 L, T のクラス分けの定義をより細かくしたり，粗くするなど，変化させたときの識別率への影響調査．
- 多くのユーザのブラウジングの癖を学習させる．本研究では，学習用サンプル，評価用サンプル，ともに 1 人のユーザのウェブブラウジングにより得たデータを利用しているため，ユーザの癖に影響されている可能性がある．
- ベイジアンネットワーク [2] [4] による，より高次な因果関係による制御．ベイジアンネットワークは，ベイズ推定を多段に組合せて複雑な因果関係から推定を行う手法であり，本手法への応用が期待できる．

文　献

[1] C. Wright, F. Monrose, G. M. Masson, “HMM profiles for network traffic classification”, Proceedings of the 2004 ACM Workshop on Visualization and data mining for coputer security, pp.9–15, 2004.

[2] J. Pearl, “Probabilistic reasoning in intelligent systems: networks of plausible inference”, Morgan Kaufmann Publishers, SanFrancisco CA, 1988.

[3] 石田 千枝，島田 栄一，荒川 豊，竹森 敬祐，渡邊 晃，笹瀬 巌, “ベイズ推定を用いた不正侵入イベント増減予測”, 信学技報, NS2003-287, IN2003-242, pp.175–178, 2003.

[4] 田中 和之，“確率的情報処理と確率伝搬アルゴリズムの基礎”, 信学技報，COMP2004-38, pp.25–32, 2004.

[5] SIP: Sessioin Initiation Protocol, RFC3261.

[6] “A plan for spam”, http://www.paulgraham.com/spam.html