LL_009

マルチスケールブートストラップ法によるパケットサンプリング

Packet Sampling using Multi-scale Bootstrap Method

本多 泰理†
Hirotada Honda

1．まえがき

パケットサンプリングは、採取された標本集合としてのパケットデータから、特定のトラヒックフローの統計的性質を抽出し、特定の性質を有するフローの検出を実現する技術である。近年、IDS 等の普及に伴い、パケットサンプリングの重要性は向上している一方、リンク速度の向上や回線の高帯域化に対応した汎用的手法の確立が求められており、IETF psamp WG でもその枠組みが既定されつつある等、検討が活発化している。本稿では、少数の標本から母集団の母数を推定する技術として、マルチスケールブートストラップ法を用いた統計的検定のパケットサンプリングへの適用を提案する。更に、同手法において従来課題とされていた、棄却領域が錐点を有するケースにおける検定も実現し、その応用としてパケットサンプリングにおける複合的検定への応用を行なう。

2．パケットサンプリングについて

パケットサンプリングの既往研究は数多く存在する。ネットワーク上、特にインターネット上において膨大なトラヒック情報を統計的に処理し、特定の性質を有するトラヒック、例えばフロー数やフローサイズが所定の比率以上であるトラヒックを検出するためには、ネットワーク上のトラヒック情報を取得し、母数の推定を行う手法が必要である。近年では、森らによるベイズ推定を用いた Elephant Flow の検出手法[1,2,3]や、ブートスラップ法を適用した手法[4,5]が発表されている他、IETF psamp WG においてその手法の枠組みが検討されている[9]。これらの手法においてはいずれも、限定された標本数のサンプルを利用し統計的な推定手法を用いて母数の推定を行う。[1,2,3]においては、フロー数の分布が Pareto 分布に従うものとの前提に基き、その事前分布を取得データから同定し、ベイズの定理の適用により事後分布を算出するプロセスが提案されている。しかしながら、該手法においては、事前分布として用いるためのトラヒックフローの分布を得るため、事前に同定を行う必要があった。またブートストラップ法による検定[4,5]においては、棄却領域の境界が平面でなく滑らかな曲面である場合、推定母数と実際の母数の差分であるバイアスが、標本数 N に対し 1 次のオーダー、即ち $O(N^{-1/2})$となることが知られている[7,8]。

本稿では、下平[7,8]により提案されたマルチスケールブートストラップ法と、筆者によるその改善手法を適用することにより、事前に同定を行なうことなく少数の標本データに基くパケットサンプリングが可能となることを示す。また該手法がバイアスを 2 次のオーダーに抑制し、かつ所望の統計的性質を有するトラヒックフローの検出が FNR5%未満の精度で可能となることを示す。

以下、次節では既往研究と課題、3 節ではマルチスケールブートストラップ法および筆者の提案手法、4 節では該手法のパケットサンプリングへの応用とケーススタディについて記載する。また 5 節でケーススタディの結果に対する考察を行う。

†慶應義塾大学大学院

3．提案手法について

前節ではパケットサンプリングに関する最近の既往研究と、その課題点について概説した。上記の課題に鑑み筆者は、事前に分布の同定が必要なくかつ推定精度の向上を図る手法として、マルチスケールブートストラップ法およびその改善手法の適用によるパケットサンプリングを提案する。初めに、マルチスケールブートストラップ法およびその改善手法のパケットサンプリングへの適用について説明する。

3.1 マルチスケールブートストラップ法

（1）ブートストラップ法の概要

マルチスケールブートストラップ法は下平[7,8]により提案されたブートストラップ法の派生手法である。ブートストラップ法は、標本からのリサンプリングによりサンプルの複製を多数用意し、各サンプルから得られた統計値により母数を推定する手法である。以下、ブートストラップ法による検定の概要を説明する：

（1）標本サイズ N の標本が得られているものとする。この時、該標本から再びサンプリングを行い標本数 N_1 (N_1<N)のサンプルを Bk 個用意する。ここでリサンプリング時に重複を許可する。

（2）各複製において、所与の帰無仮説 H に対する検定を実施し、仮説が支持される回数 Cr を記録する。

（3）全ての複製に対して（2）の過程が実施された後、得られた Cr の複製数に占める割合 Pr=Cr/Bk を以って、該仮説に基く母数の推定 p 値とする。有意水準 5%で検定を実施する場合、推定 p 値が 5%以上であれば帰無仮説 H は支持され、それ以外の場合帰無仮説 H は棄却される。

以上のブートストラップ法による検定の課題として、帰無仮説の棄却領域 ∂H を座標空間上で表現した場合に、∂H の境界が平面で無く滑らかな曲面である場合、推定される p 値のバイアスが $O(N^{-1/2})$ となる事が知られている[7,8]。これは境界が曲面のため p 値の理論値の分布とブートストラップ法で得られる推定値の間に差分が生じてしまうためである。そこで下平は、棄却領域の境界が滑らかな曲面である場合においてより高精度の検定を実現するマルチスケールブートストラップ法を提案した[7,8]。該手法においては、標本数を変更しながらリサンプリングを行うことにより推定精度の向上を図るが、母集団の分布によりいくつかの手法が考案されている。本稿では指数

型分布族を対象とし、2ステップマルチスケールブートストラップ法と呼ばれる方法を採用する。以下、該手法による検定の過程を説明する（図1）：

（1）サイズ N の標本が得られているものとする。この時、該標本から再びサンプリングを行い、データ長の比 r_{11}、r_{12}、$\cdots r_{1m}$ および各 r_{1i} に対応する複製の個数 B_{11}、B_{12}、$\cdots B_{1m}$ を決定し、各 i に対して標本数 $r_{1i}N$ 個の標本を B_{1i} 個ずつ生成する。これをそれぞれ $\{X^*_1(r_{11}), \cdots, X^*_{B11}(r_{11}); X^*_1(r_{12}), \cdots, X^*_{B12}(r_{12}); \cdots, X^*_1(r_{1m}), \cdots, X^*_{B1m}(r_{1m})\}$ とする。

（2）各 $X^*_{B1i}(r_{1j})$に対して、更にスケール比 τ_2 でリサンプリングを行う。即ち、各 $X^*_i(r_{1j})$から、ランダム抽出によりデータサイズ $\tau_2 N$ 個の標本 $X^*_1(r_{1i}, r_{2j})$ (l=1,$\cdots$, B_{1i}、i=1,$\cdots$, m、i=j,$\cdots$, n)を作成する。

（3）各 $X^*_1(r_{1i}, r_{2j})$に対して帰無仮説 H の支持/棄却を調べ、仮説が支持された回数を記録する。データ $X^*_1(r_{1i}, r_{2j})$ における支持回数を $Cr(r_{1i}, r_{2j})$ (i=1,2,$\cdots$K)とする。

（4）各 i に対して、通常のブートストラップ法により算出された推定確率値 $p(r_{1i}, r_{2j})$を p 値の理論値

$\pi(s_1,s_2,s_3;\gamma_1,\gamma_2,\gamma_3)=1-\Phi(s_1\gamma_1(1+s_2\gamma_3)-(\gamma_2+s_2\gamma_3)/s_1\gamma_1)$ に当てはめ、最小二乗法により以下の各回帰係数を推定する：$\hat{\gamma}_1=\hat{v}-2\hat{a}\hat{v}^2, \hat{\gamma}_2=\hat{v}(\hat{a}-\hat{c}), \hat{\gamma}_3=\hat{v}\hat{a}^2$

ここで $s_1=(\tau_1+\tau_2)^{-1/2}$、$s_2=\tau_1^2\tau_2^2 s_1^4$。

各回帰係数は指数型分布族のパラメータを表している。Φは標準正規分布関数である。

（6）（5）で推定された回帰係数により、計算式 $\hat{p}=1-\Phi(\hat{\gamma}_1(1+\hat{\gamma}_2)+\hat{\gamma}_2/\hat{\gamma}_1)$ により、補正した推定 p 値を算出する。

ここで上記の記号につき補足する。一般に指数型分布族とは、分布関数が $P(X_i=x_i)=\exp\{\theta' x_{ij}+\Psi(\theta')\}$の形で表される確率分布の集合である。（5）の a および γ_i は指数型分布族の幾何的パラメータを表す数値であり、特に a は加速定数と呼ばれる。

以上の2ステップマルチスケールブートストラップ法により、指数型分布族に対するマルチスケールブートストラップ法を用いた検定が可能となる。なお、上記手順により求められた推定 p 値はバイアス $O(N^{-1})$ となることが知られている。また母集団が正規分布を示すことが事前に判明している場合には、通常のマルチスケールブートストラップ法の適用により、バイアスを $O(N^{-3/2})$ に抑制可能である。いずれの場合においても、母集団の分布は正規分布もしくは指数型分布族に属することが前提であり、それ以外の分布の場合には滑らかな変換により同分布族へ変換可能であることが必要である。2ステップマルチスケールブートストラップ法では2次のオーダー、3ステップの同手法では3次の精度の検定となるが、本稿では計算時間の観点および N が非常に大きいことから、2ステップの手法を採用している。

4.2 マルチスケールブートストラップ法の課題

前節において、母集団が指数型分布族に属する場合のマルチスケールブートストラップ法による検定の手法を概説した。しかしながら、棄却領域の境界が滑らかでない場合においては、上記の主張は成り立たないことが知られている。筆者は上記の課題に対して、棄却領域が錐である場合には、適当な座標変換により指数型分布族に対するマルチスケールブートストラップ法による検定が可能であることを主張する。例えばいま、元の確率変数 $X_1, X_2, \cdots X_n$ の分布が各々 $P(X_i=x_i)=\exp\{\theta^{(i)} x_{ij}+\Psi(\theta^{(i)})\}$ (i=1,$\cdots$,n)と表されるものとする。この時、座標軸の回転 $\hat{X}_i=AX_i$（A は直交行列）および座標変換 $\hat{y}_i=F(\hat{x}_i)$ により、確率変数の分布は $P(\tilde{X}_i=x_i)=\exp\{\theta^{(i)}(A^{-1}x_i)_j+\Psi(\theta^{(i)})\}$ と変換される。これも指数型分布族に属するため、座標変換により元の棄却領域の錐点が滑らかな曲面に変換されれば、$\tilde{X}_i$ に対する $\hat{y}_i$ 平面上でのマルチスケールブートストラップ法による検定が可能となる。

5. 提案手法について

5.1 パケットサンプリングへの適用

本節では前記提案手法のパケットサンプリングへの適用について検討を行う。ここでは、インターネット上のトラヒックフローに対して、所定のフロー数およびフローサイズの閾値を超過するトラヒックの検出を目的としてパケットサンプリングを行う。

前述の通り、インターネット上の非常に大きなトラヒックをサンプリング対象とする場合、フロー数の分布は pareto 分布に従うことが一般的に知られており、本稿ではこれを座標変換により指数型分布族の確率変数へ変換することにより、マルチスケールブートストラップ法を適用する。

提案手法の詳述に先立ち、トラヒックフローの定義を行う。トラヒックフローとは、IP パケットに記述される属性のうち、送信元アドレス、宛先アドレス、送信元ポート番号、宛先ポート番号、およびプロトコル番号のセットとして定義される概念である。これらの属性が同一のパケットは同一のトラヒックフローに属するパケットと見なす。Pareto 分布の定義から、一般にフロー数の分布関数は、

$P(X=x)=\beta\alpha^{\beta}/x^{1+\beta}$ $(x\geq\alpha)$と表される。またパケットサイズは、データ信号のみをサンプリング対象とするとの前提に基き、ここでは所定の区間$[y_1, y_2]$の一様分布を仮定する：$P(Y=y)=1/N_Y$。この時、フローサイズ Z=XY の分布は

$P(Z=z)=c_1/z^{1+\beta}$ $(c_1=\beta\alpha^{\beta}(y_2^{\beta+2}-y_1^{\beta+2})/((\beta+2)N_Y))$

と表すことができる。但し提案手法を適用するため、変数変換 X1=ln(X)、Z1＝ln(Z)により上記の変数を指数型分布族に属する確率変数へ変換する。この場合、

$P(X1=x)=\beta\alpha^{\beta}/e^{(1+\beta)x}$

$P(Z1=z)=c_1/e^{(1+\beta)z}$

により X1 および Z1 はともに指数型分布族に属する分布を有する確率変数であることが確認できる。

5.2 ケーススタディ

本稿では、トラヒックの中で特に大きな統計的性質を具備するトラヒックフロー（Elephant flow）の定義に基き[2]、以下の検定を実施する：

①トラヒックフロー数 0.1%以上のフローの検出

②フローサイズ 0.1%以上のフローの検出

③トラヒックフロー数 0.1%以上かつフローサイズ 0.1%以上のフローの検出

上記の検定を、トラヒックフロー数およびフローサイズをそれぞれ確率変数 X1 および Z1 へ置き換えて帰無仮説を定義する：
帰無仮説①'：$X1<\ln(10^{-4})$
帰無仮説②'：$Z1<\ln(10^{-4})$
帰無仮説③'：$X1<\ln(10^{-4})$または$Z1<\ln(10^{-4})$

以上により当初の問題は、上記の各検定を指数型分布族に対するマルチスケールブートストラップ法を適用して実施する問題に帰着された。

本稿では上記の検定の有効性を確認するため、シミュレーションにより検定結果を確認する。設定として、測定期間に流れた 10^5 個のパケットに対して、10^3 個の標本がサンプリングされた状況を想定し、検定による検出結果と元のデータとの比較を行う。帰無仮説③'については、帰無仮説の棄却領域 H の境界は図２のように尖点を有する領域となるため、筆者の提案手法を適用し座標変換：
$X2=X1-X0$, $Z2=Z1-Z0$、$(X0=Z0=10^{-4})$、更に
$X3=(X2-Z2)/2^{1/2}$、$Z3=(X2+Z2)/2^{1/2}$,
$Z4=Z3^2 (Z3>0)$, $Z4=-Z3^2 (Z3<0)$ を実施する。
この時、新たな変数の同時確率分布は
$P(X4=x, Z4=z) = c_1 \beta\ \alpha^{\beta} \exp\{-(1+\beta)(\mathrm{sign}(z)z^{1/2}/2^{1/2}+(X0+Z0))$
となる。これは再び指数型分布族に属する分布である。
以上の変換により、当初の検定③'の帰無仮説の棄却領域は H3'："Z3<X3^2"に変換される。
スケール比、β の値及び複製数等の設定は表１に記載する。以上の設定の下で、仮想的にパケット数とサイズから構成されるフローの集合を生成しシミュレーションを実施する。

5.3 結果および考察

表２は、オリジナルデータ（母集団）の中で①、②、③の各条件（比率が 10^{-4} より大）に該当するフローの数と、提案方式によりサンプルフローの中から検出されたフローの数を各 β の値について示す。β が 1.25 および 1.75 の時は検出対象フロー数および検出数ともに大きな値となっている。サンプルフローは実際のデータの１％のみの採取であるが、検出数は実際に検出すべきフローの約 40%から 55%程度の値を示した。また表３に、サンプリングされたデータに占める検出対象のフロー数と、その内的確に検出されたフロー数の内訳を記載する。図３は各検定項目に対する検出状況を示す図である。「A.未検出」はサンプルフローに含有される検出対象フローのうち、検出に失敗したフローの数を、「B.検出」は検出対象フローのうち的確に検出したフローの数、「C.誤検出」は非検出対象フローのうち検出したフローの数を示す。従って、A の値と B の値の合計がサンプルフロー数に占める検出対象のフロー数、B の値と C の値の合計が提案方式による検出フロー数の総和を表す。以上より、サンプルされたフローの中の検出対象のフローの多くが検出されていることが確認できる。さて本稿では検定の有効性を確認するための指標として、FNR(False Negative Ratio)および FPR(False Positive Ratio)を使用する。前者は検定において検出すべき対象の非検出率、後者は検出すべきでない対象の誤検出の率をそれぞれ表す。各検定項目に対する FNR の結果を図４、FPR の結果を図５に示す。これらの結果から、β の値により多少の変動はあるものの、FNR は 10%未満に抑制されている一方、FPR は 30-50%とであることが確認される。理論的結果によれば、２ステップのマルチスケールブートストラップ法はバイアス $O(N^{-1})$ の近似的に普遍な検定である。ここで普遍な検定とは、有意水準 α $(0<\alpha<1)$ で検定を実施する時、
$P[\hat{\alpha}<\alpha \mid \eta \in H] \le \alpha$　、$P[\hat{\alpha}<\alpha \mid \eta \notin H] \ge \alpha$
を満たす検定を指す。この時、∂H 上の点 η に対して帰無仮説が棄却される確率 $P[\hat{\alpha}<\alpha \mid \eta \in \partial H]=\alpha$ となるが、２ステップのマルチスケールブートストラップ法の場合、この値は２次のオーダーのバイアス $O(N^{-1})$ を含み $\alpha+O(N^{-1})$ となることが知られている[8]。ケーススタディの場合、α=0.05、$N=10^5$ であるから、FPR は高々0.0501 で抑えられることになるが、本結果においてはこれを上回る FPR が提示された。一原因として、スケール比が小さくリサンプル後のデータ数が少なかったことが考えられる。但し、本稿の結果は既存手法との比較において FPR、FNR ともに低く抑制している。以上の結果から、マルチスケールブートストラップ法に基く提案手法により、事前分布の同定を行うことなく、既存の手法以上の精度での推定を実施可能であることが確認できた。

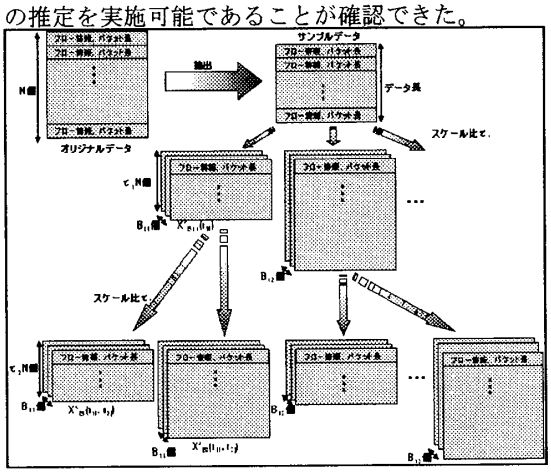

図１．マルチスケールブートストラップ法

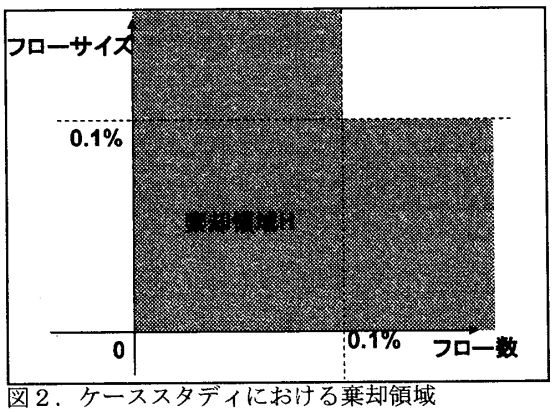

図２．ケーススタディにおける棄却領域

表１．各種パラメータ

複製数 Bk	スケール比 τ1	スケール比 τ2	α	β
50	0.75, 0.85, 1.5, 1.75	0.5, 1.2, 1.75	1.0	1.25 1.5 1.75

表2.実際の特定フローと検出数

β	オリジナルデータにおける検出対象フローの数			提案方式による検出数		
	①	②	③	検出数①	検出数②	検出数③
1.25	1123	1205	1123	484	607	607
1.50	762	803	762	316	316	316
1.75	1123	1205	1123	528	621	621

表3.検出フローの内訳

β	サンプルデータにおける検出対象フローの数			左記の内、提案方式による検出数		
	①	②	③	検出数①	検出数②	検出数③
1.25	242	245	242	163	191	191
1.50	202	205	202	171	174	171
1.75	252	254	252	176	202	200

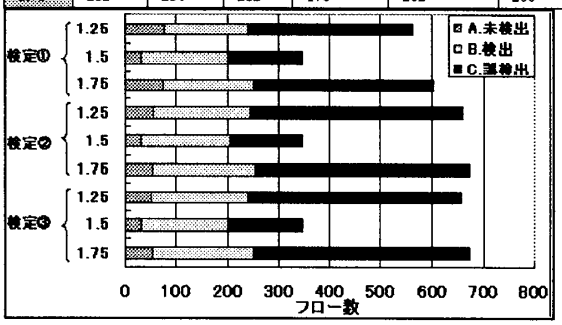

図3．各検定項目の検出数

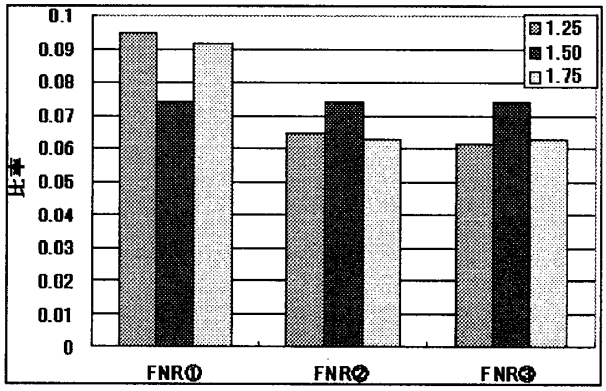

図4．各検定項目の FNR

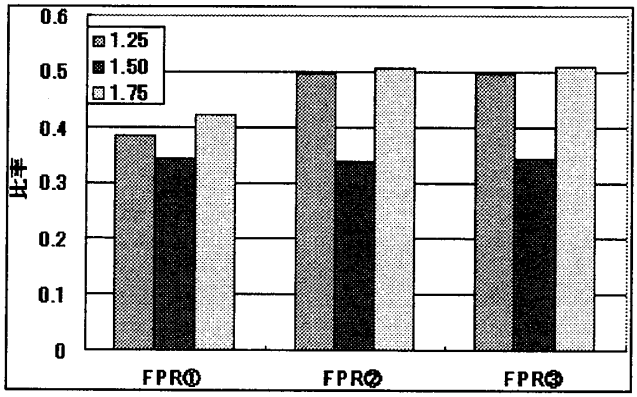

図5．各検定項目の FNR

6. まとめおよび今後の課題

本検討では、パケットサンプリングにおいて取得されたデータに対し所定の変数変換の実施によりマルチスケールブートストラップ法が適用となること、その結果、母集団の 1%の標本数のデータに対して特定の閾値を上回る統計的性質を示すフローの抽出が FNR10%未満の精度で実現されることを確認した。提案方式によれば、事前分布の仮定は必要無いため、分布の同定を行うことなく所望のフローの検出が可能となる。一方で FPR は理論値に比して高い値を提示した。本手法は IDS における異常トラヒック検知や、品質劣化要因の検出等への応用も可能であると考えている。今後の課題として、①本検討で理論値との乖離が顕著であった FPR に対する検討の精緻化②経験分布等、より一般的な分布設定に対する提案手法の可用性の向上と実際のトラヒックデータへの適用④最小二乗法の最適化アルゴリズムにおける精緻化と計算効率の向上⑤PPPoE 等のカプセル化されたトラヒックに対する汎用的パケットサンプリング手法についての検討、を行っていきたい。

7. 参考文献

1．上山憲昭、森達哉、川原亮一「トラヒック計測におけるタイムアウト処理に関する検討」、電子情報通信学会信学技報、IN2005-126、pp.97-202

2．Tatsuya Mori, Masato Uchida, Ryoichi Kawahara, "Identifying Elephant Flows Through Periodically Sampled Packets", IMC'04, October, 25-27, 2004.

3．川原、森、石橋、上山、阿部「サンプルパケット情報を用いた TCP 品質劣化検出のためのフローレート推定法」、電子情報通信学会信学技報 NS2005-114, pp.7-12.

4．Stenio F.L.Fernandes, Tatience Correia, et al. "Bootstrap-Based Estimation of Flow-Level Network Traffic Statistics",

5．Stenio F.L.Fernandes, Tatience Correia, et al. "Estimating Properties of Flow Statistics using Bootstrap",

6．Bradley Efron, "Bootstrap Confidence Intervals for a Class of Parametric Problems", Biometrica, 1985, 72, 1, pp.45-58.

7．下平英寿、「ブートストラップ法の幾何学とスケール変換」、日本数学会 2004 年秋季総合分科会、2004.

8．下平英寿、「ブートストラップ法によるクラスタ分析のバラツキ評価」、統計数理第 50 巻第 1 号 pp.33-44、2002.

9．IETF Packet Sampling (Psamp) Working Group, http://www.ietf.org/html.charters/psamp-charter.html