様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月３１日現在

機関番号：１２６０８
研究種目：基盤研究（C）
研究期間：　２００８年　～　２０１０年
課題番号：　２０５００２５４
研究課題名（和文）ベイズ的なマルチスケール・ブートストラップ法の理論とその応用

研究課題名（英文）Theory of Bayesian Multiscale Bootstrap and its Applications

研究代表者
下平　英寿（SHIMODAIRA HIDETOSHI）
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：００２９０８６７

研究成果の概要（和文）：
データ解析の信頼度を高精度で計算するために，ランダムネスのスケーリング則を利用したリサンプリング・アルゴリズムの研究を行った．先行研究で提案したマルチスケール・ブートストラップ法では頻度論の立場で統計的仮説検定の不偏な信頼度（p-値）を近似計算した．本研究では頻度論だけでなくベイズ法の立場で事後確率の計算を行う方法を提案し両者の関係を明らかにした．さらにランダムネスのスケーリング則を利用して機械学習の信頼度計算および能動学習を行った．

研究成果の概要（英文）：
A resampling algorithm based on a scaling-law of randomness has been studied for computing a highly accurate confidence level of data analysis. In a previous study, we have proposed multiscale bootstrap method which computes an approximately unbiased confidence level (p-value) of statistical hypothesis testing in a frequentist sense. In this study, we proposed a method for computing the posterior probability in a Bayesian sense. We have shown a connection between the frequentist and the Bayesian confidence levels. We also studied a confidence level, as well as an active learning, for machine learning using the scaling-law of the randomness.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,600,000	480,000	2,080,000
２００９年度	1,300,000	390,000	1,690,000
２０１０年度	600,000	180,000	780,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：数理統計学
科研費の分科・細目：情報学・統計科学
キーワード：リサンプリング，マルチスケール，スケーリング則，仮説検定，信頼度，機械学習，バイオインフォマティクス，モデル選択

１．研究開始当初の背景

（１）近年ゲノム科学など様々な分野で膨大なデータが蓄積されている．これから知識発見を行うために，データマイニングによって非常に多くの仮説が同時に探索されることがある．このような状況では，データに内在するランダムネスの影響が増幅されてバイアスを生じ，誤った発見に導かれやすくなる．これは仮説検定の多重性と呼ばれる効果で

ある．分子進化学を例にして説明する．生物進化の分岐順序を表すラベル付き木は系統樹とよばれ，DNA 配列の生物間差異から推定される．系統樹推定は統計的モデル選択の一例であって，仮説となる系統樹を確率モデルで表現し，情報量規準を最小にする系統樹を選ぶ．比較する生物種の個数が増えると系統樹の個数は指数的に増える．結果として，多数の誤った系統樹のうち一つが「まぐれ」で過度にデータへ適合し，新たな発見をしたように見えてしまう確率が高くなる．

（2）そこでデータのランダムネスが推定値に与える影響を正しく評価することが重要になる．このための一般的な確率シミュレーション技法がブートストラップ法であり，アルゴリズムはきわめて単純で応用が容易である．データからランダムに要素を取り出し複製データを１万回程度生成する（図１）．これらに通常のデータ解析を繰り返し適用して得られた系統樹の集合を調べ，データが仮説を支持する頻度（これをブートストラップ確率と呼ぶ）を信頼度として利用する．並列化が容易なので，計算量は本質的な問題ではない．ところがブートストラップ法にはバイアスがあり信頼度の精度が不十分であることが分かってきた．

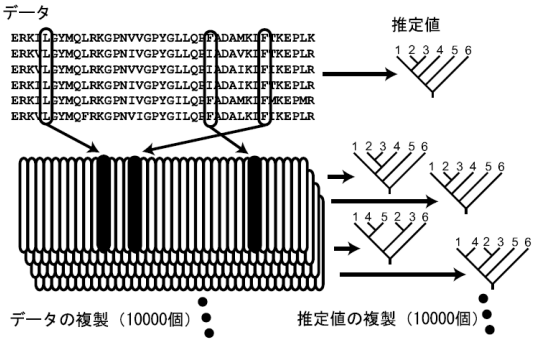

図１：ブートストラップ法の概念図

（3）下平は頻度論の立場でブートストラップ確率のバイアス補正を行い，信頼度を高精度で計算するアルゴリズムを開発してきた．キーとなるアイデアは，データのスケールを変化させたときのブートストラップ確率の変化率から，仮説領域の幾何学的な情報（データ点までの距離や仮説境界の曲率）を引きだすことである．これらの情報が得られれば，直ちに信頼度を精確に計算できる．ここでデータのスケールというのはランダムネスの程度を表すスカラー量であり，サンプルサイズの平方根に反比例する．

（4）観測データのサンプルサイズを n，ブートストラップ法でランダムに生成する複製データのサンプルサイズを m とする．通常のブートストラップ法では $m=n$ である．マルチスケール・ブートストラップ法では m を変化させる．このような方法は一般に m-out-of-n bootstrap 法と呼ばれる．データからランダムに要素を取り出して複製データを生成するときに，通常は n 回反復して n 個の要素を取り出すが，それを m 回反復して m 個の要素を取り出すようにブートストラップ法のプログラムを修正するだけであり，容易に実装できる．ブートストラップ法は重複を許したリサンプリング，すなわち，同じ要素を複数回取り出すことを許す方法である．従って任意の整数 $m>0$ に変更可能である．m を変更すると，複製データのランダムネス（バラツキ）が変化する．大きな m の複製データから計算した推定値は分散が小さくなり，小さな m の複製データから計算した推定値は分散が大きくなる．この分散は m に反比例するので，観測データから計算した推定値の分散を調べるために本来は $m=n$ にする必要がある．

（5）ところが下平の提案したマルチスケール・ブートストラップ法の理論では，精度の高い信頼度を計算するためには，負のサンプルサイズ $m=-n$ とすることが証明される．m は取り出す要素の個数であるから，正の整数と考えるのが普通であるが，理論上は負の値が最適と証明される．この常識外れなところがマルチスケール・ブートストラップ法のオリジナリティといえる．実際の計算では，複数の m（もちろん正）においてブートストラップ法を実行してその結果を $m=-n$ へ外挿する．このときブートストラップ確率をそのまま外挿するのではなく，

$$p(\sigma^2)=\Phi\left(\sigma\Phi^{-1}\left(BP(\sigma^2)\right)\right)$$

を $\sigma^2=-1$ へ外挿する．ただし，

$$\sigma^2=\frac{n}{m}$$

は推定量の相対的な分散（$m=n$ のとき $\sigma^2=1$），$BP(\sigma^2)$ はブートストラップ確率，$\Phi()$ は正規分布（平均 0，分散１）の累積分布関数である．$m=n$ の通常のブートストラップ確率は $p(1)$，$m=-n$ の精確な信頼度は $p(-1)$ と書ける．

２．研究の目的

（1）本研究の目的はベイズの立場で仮説の事後確率を高精度で計算するブートストラップ法の開発である．もちろん事後確率は事前分布に依存する．もし平坦な事前分布ならブートストラップ確率が事後確率である．図

２の左図のような状況においては，事後確率が頻度論の信頼度に一致するような「マッチング事前分布」を与えることもできる．ところがおなじマッチング事前分布を用いても右図では両者が一致しない．Efron and Tibshirani（1998）は，右図の状況でもマッチング事前分布を用いた事後確率の利用を提唱したが，現実的な計算法が与えられていなかった．本研究ではマルチスケール・ブートストラップ法を利用して，このような事後確率の計算を可能にする．

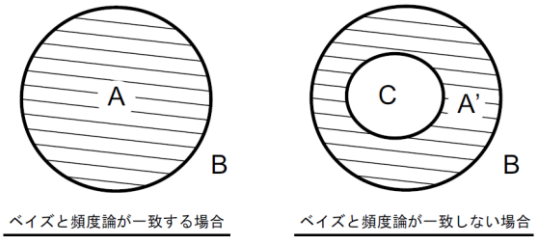

図２：斜線の領域が仮説（A または A’）

（２）これまでに提案していたマルチスケール・ブートストラップ法は頻度論的な信頼度を計算する手法である．この方法は近似計算に基づいており，図２の左図の状況では一般に精度の高い信頼度を計算するが，右図の状況では頻度論的にも精度が悪くなる場合がある．新たに提案を試みた手法は，ベイズのみならず頻度論的にもマルチスケール・ブートストラップ法を拡張して適用範囲を広げ，右図の状況でも精度を上げることが期待される．このような拡張は応用上の適用範囲を広げるだけでなく，ベイズと頻度論の両者を同じ枠組みで記述するので，より合理的な帰納推論へ何らかの示唆が得られる可能性もある．

（３）マルチスケール・ブートストラップ法は応用を限定しない一般的な手法であるが，これまでは分子進化学における系統樹推定やバイオインフォマティクスにおける遺伝子発現データの階層的クラスタリングに利用されることが多かった．本研究ではこのほかの応用も試みる．様々な応用を通して理論的な課題が明らかになる．

３．研究の方法

（１）理論的な考察，とくに数理統計学の手法によって課題の解決を試みる．m を変化させたときのブートストラップ確率の変化は一種のスケーリング則である．マルチスケール・ブートストラップ法の実現には，まずスケーリング則を明らかにすることが必要である．これまでに得られている結果は図２の左図の状況を想定していたが，本研究では右図の状況でも有効なスケーリング則を明らかにする．

（２）一方で，データ点から仮説境界までの距離や仮説境界の曲率など幾何学的なパラメータをつかって精確な信頼度を導出する．これについても図２の左図の状況では分かっていて，頻度論とベイズの結果が一致する．本研究では図２の右図の状況を考察する．

（３）上記の（１）と（２）の結果を結びつける．ブートストラップ確率の変化にスケーリング則を適用して，この当てはめから幾何的パラメータを推定する．このようにして得られた幾何的パラメータをつかって精度の高い信頼度を計算する方法を考察する．

（４）理論的な結果はソフトウエアに実装する．これは手法を応用するために必要であるばかりでなく，理論的な結果をシミュレーションによって検証するために重要である．

４．研究成果

（１）図２の右図において，とくに領域 A’が薄く領域Ｂと領域Ｃが近い状況でスケーリング則を導出した．左図におけるスケーリング則を表す式を二つ利用して，それらの差の形式で表現できる．A=A’+C とおけば右図は左図になり，これまでのスケーリング則が適用できる．また A’+B に対しても同様にこれまでのスケーリング則が適用できる．この二つの結果を一つの式にまとめれば，右図のスケーリング則が得られる．

（２）右図の A’を帰無仮説とする頻度論的な信頼度 p(A’)を，B を帰無仮説とする頻度論的な信頼度 p(B)および C を帰無仮説とする頻度論的な信頼度 p(C)を使って表すと，$p(A') = 1 - |p(B) - p(C)|$ となることが分かった．これは A’を帰無仮説，B+C を対立仮説とする両側検定の信頼度である．同様にベイズ的な信頼度は $1 - \min\{1, p(B)+p(C)\}$ と書けることが分かった．

（３）上記（１）のスケーリング則をマルチスケール・ブートストラップ法で得られるブートストラップ確率の変化に適用すると，図２の左図の状況における従来理論を適用することによって，p(B)と p(C)を計算できる．この結果を上記（２）に代入することにより，右図の A’に対する頻度論的，およびベイズ的な信頼度が計算できる．

（４）上記（２）では頻度論的な A’の両側

検定を示した．A’の片側検定の信頼度は1-p(B)または1-p(C)であるから，そのうちとくに小さい方を使うことにすると，1 - max{p(B), p(C)}と書ける．この片側検定(one-sided, s=1)の頻度論的信頼度と先ほどの両側検定(two-sided, s=2)の頻度論的信頼度を s に関して外挿すると，A’のベイズ的信頼度は s=0 に相当して，いわば無側検定(zero-sided, s=0)の頻度論的信頼度と解釈できる．通常このような検定は存在しないが，頻度論とベイズの関係を表すと考えられる．

（５）大規模シミュレーションにより，ここで導出した頻度論的およびベイズ的信頼度が，理論上予想される性質をもつことを確認した．

（６）機械学習への応用として判別の信頼度の計算を行った．判別のブートストラップはバギングと呼ばれる．これにマルチスケール･ブートストラップ法を適用した．$m=-n$ によって高精度の信頼度を計算するほかに，$m\to\infty$ とすると能動学習の効果が得られることが分かった．

（７）因果分析(LinGAM 法)への応用を行い，因果の信頼度を計算できることを確かめた．またブートストラップ法の計算を高速化するための技術を新たに考案した．

（８）ブートストラップに関連した問題として，ＭＣＭＣ法を利用してグラフのスケールフリー性を仮定したベイズ推測を行った．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① P. Sheridan, T. Kamimura, H. Shimodaira, A scale-free structure prior for graphical models with applications in functional genomics, PLoS ONE, 5, e13580, 2010, 査読あり
② Hidetoshi Shimodaira, Frequentist and Bayesian measures of confidence via multiscale bootstrap for testing three regions, Annals of the Institute of Statistical Mathematics, 62, 189-208, 2010, 査読あり

〔学会発表〕（計 16 件）

① H. Shimodaira, Multiscale Bagging with Applications to Classification and Active Learning, The 2nd Asian Conference on Machine Learning (ACML2010), 2010.11.10, 東工大（東京）
② M. Aoki, Multiscale-bagging with Applications to Classification, The 2nd Asian Conference on Machine Learning (ACML2010), 2010.11.10, 東工大（東京）
③ Y. Komatsu, Assessing statistical reliability of LiNGAM via multiscale bootstrap, International Conference on Artificial Neural Networks (ICANN2010), 2010.9.15, Thessaloniki (ギリシャ)
④ Hidetoshi Shimodaira, Approximately unbiased tests for cone shaped regions via multiscale bootstrap, 2009 Annual Meeting, Statistical Society of Canada, 2009.6.3, The University of British Columbia （カナダ）
⑤ Hidetoshi Shimodaira, Frequentist and Bayesian measures of confidence via multiscale bootstrap for testing three regions, A Bayesian Approach to Statistical Inference and Its Related Topics, 2008.10.22, 京都大学数理解析研究所

〔図書〕（計 1 件）

① 竹村彰通, 北川源四郎, 藤越康祝, 久保川達也, 塚原英敦, 田中勝人, 内田雅之, 下平英寿, 渡辺美智子, 古澄英男, 生駒哲一, 東京大学出版会, 21 世紀の統計科学ＩＩＩ　数理・計算の統計科学, 2008, 209-238

〔その他〕
ホームページ
http://www.is.titech.ac.jp/~shimo/index-j.html

大田区の区民大学（東京工業大学連携講座）
「ＤＮＡ情報からよみとる生物進化とランダムネス」
http://www.is.titech.ac.jp/~shimo/kumin2010/index-j.html

６．研究組織
(1)研究代表者
下平　英寿（SHIMODAIRA HIDETOSHI）
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：００２９０８６７

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：