様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月７日現在

機関番号：62603
研究種目：基盤研究(C)
研究期間：　2010～2012
課題番号：22500217
研究課題名（和文）マルコフ連鎖モンテカルロ法による仮想データ生成と非線形情報処理への応用
研究課題名（英文）fictitious data generation using Markov chain Monte Carlo and application to nonlinear information processing
研究代表者 伊庭 幸人（IBA YUKITO）
統計数理研究所・モデリング研究系・准教授
研究者番号：30213200

研究成果の概要（和文）：
「仮想データ生成問題」に対してマルコフ連鎖モンテカルロ法による解法を提案し、その具体例としてサロゲートデータ生成とプレイメージ生成の問題を研究した。前者に関しては、非線形時系列の問題にマルチカノニカルモンテカルロ法（マルコフ連鎖モンテカルロ法の一種）を適用し，提案手法がうまくいくことを示した。後者については、創薬に関係した化学構造式の判別問題に対応するプレイメージ問題を研究し，有望な結果を得た。

研究成果の概要（英文）：
A Markov chain Monte Carlo solution for “fictitious data generation” is proposed; it is applied to data surrogation and preimage generation. Surrogation of nonlinear time series is successfully treated by multicanonical Monte Carlo. A preimage problem for drug design corresponding to discrimination of structural diagrams is also studied.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2011 年度	600,000	180,000	780,000
2012 年度	500,000	150,000	650,000
2013 年度	500,000	150,000	650,000
年度			
年度			
総　計	1,600,000	480,000	2,080,000

研究分野：総合領域
科研費の分科・細目：情報学，感性情報学・ソフトコンピューティング
キーワード：確率的情報処理

１．研究開始当初の背景

(1) マルコフ連鎖モンテカルロ法

マルコフ連鎖モンテカルロ法（ＭＣＭＣ）は高次元の複雑な確率分布からの乱数生成（サンプリング）のための汎用の手法である。1950 年代に統計物理で導入されて以来、物理学の分野ではさかんに利用されており、1990 年代からはベイズ統計によるデータ解析で広く使われるようになった。その中でも、1990 年代に導入されたレプリカ交換モンテカルロ法・マルチカノニカル法など拡張アンサンブＭＣＭＣと総称される方法はマルコフ連鎖モンテカルロ法の主要な欠点である「多峰型分布についての収束の遅さ」を改善し、潜在的な応用範囲を大きく広げた。しかし、これらの方法の可能性は上記以外の分野

ではまだ十分に生かされているとは言えず、さらなる応用の開拓が望まれていた。

(2) 仮想データ生成
「仮想データ生成」という概念は本研究計画ではじめて提示されたものであり、従来の研究でそうしたまとめ方で横断的にとらえた例はないと思われる。

(3) サロゲートデータの生成
データの中から読み取ったパターンが実際に存在するものなのか、偶然に起きたゆらぎなのかを判定することは情報処理において基本的な問題であるが、現実のデータの解析にあたっては、教科書的な統計的検定よりも著しく複雑な「帰無仮説」を考える必要がしばしば起きる。たとえば、複数の神経細胞から得られたスパイク時系列に意味のある相関があるか否か、という問題を考えた場合、ランダム系列との比較では不十分であり、系列内部の相関やスパイク潜時などがもとの時系列と一致するような「仮想データ」を考える必要がある。このような仮想データをサロゲートデータと呼び、それを生成することをサロゲーションと呼ぶ。サロゲートデータを生成するために、従来は個々の事例に応じたさまざまな工夫が行われてきた。Schreiber は 1998 年の論文 Constrained randomization of time series data, Physical Review Letters 80 2105-2108 で「サロゲートデータと実験データとの統計量の一致を拘束条件として考え、最適化による統一的な扱いを行う」ことを提案したが、最適化法では独立な多数のサンプルを偏りなく生成できる保証がないという点で問題があった。

(4) プレイメージ生成
「仮想データ生成」のもうひとつの例として、学習ずみの情報処理装置に対して、内部の非線形特徴空間の部分集合を与えたときに、その逆像となるような仮想的なデータを組織的に求める問題がある（プレイメージ生成問題)。たとえば「Aという望ましい性質（たとえば薬効)」「Bという避けるべき性質（たとえば毒性)」があったとする。特徴空間において「Aを持ちBを持たない」集合が学習によって知られているときに、その逆像となる入力データを多数求めることができれば、有用な対象を実験で生成するためのヒントが得られる。こうしたことが組織的に実現できれば、創薬など様々な分野への応用が期待できるが、従来は現実的な系、特に離散構造をデータとして扱う場合についての結果は乏しかった。

２．研究の目的
(1) 一般に確率的情報処理においては、目的に応じた「仮想データ」を生成する技術が重要である。本研究計画では、マルコフ連鎖モンテカルロ法を応用することで「仮想データ生成」に対する統合的なアプローチを開発し、実世界の問題で検証することを目指す。具体例としてはすでに述べた２つの例、サロゲートデータ生成とプレイメージ生成を扱う。

(2) サロゲートデータの生成のために、個別の事例に応じた様々な工夫が行われてきたが、その本質は「与えられた統計量がもとのデータと一致する」ことを拘束条件とする条件つき分布からのサンプル生成であると考えられる。従って、拡張アンサンブルＭＣＭＣを利用すれば、きわめて一般的にこれを扱うことができるはずである。本計画ではこの考え方を実装し検証することを目指す。

(3)　本研究計画では、プレイメージ生成問題を「複数の非線形特徴が与えられたものと（近似的に）一致する」という拘束条件から定まる分布からのサンプリング問題として解釈し、ＭＣＭＣによって扱う。これによって、様々な状況で適用できる統一的な手法が開発できることが期待される。本計画では、この考え方を実装し、実世界の離散構造データへの応用までを行うことを目指す。

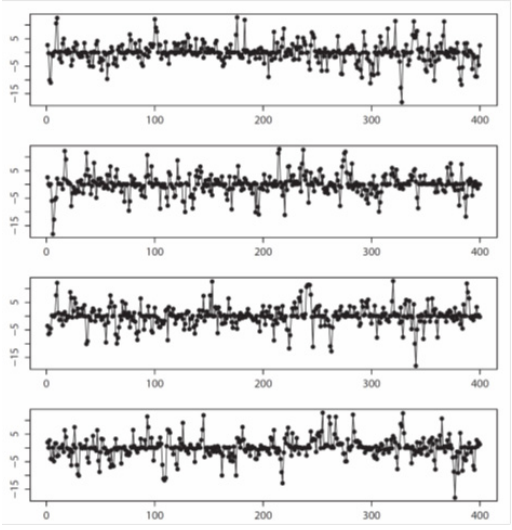
３．研究の方法
(1)研究を開始するにあたっては、研究集会等を開く、海外の国際会議に参加する、などの方法で情報収集を行う。また各分野の専門家と随時議論し、研究のコンセプトに対する反応や手法の適用対象に関する情報を収集する。

(2)サロゲートデータ生成については、非線形時系列の事例でマルチカノニカル法の実装を行う。また神経スパイクデータへの応用について考察する。

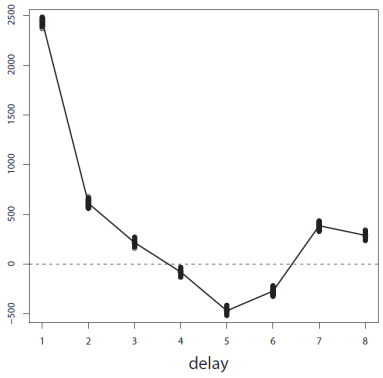
(3)プレイメージ生成については、創薬の専門家との共同研究を行う。

４．研究成果
(1)準備
「仮想データ生成」のうち、「サロゲートデータの生成」「プレイメージ生成」「データ秘匿」の各課題について、それぞれの関連分野の研究者を招聘して、研究集会「仮想データ生成とその周辺 : 逆像問題、サロゲーション、秘匿」を行い、今後の研究に必要な情報収集と討論を行った。見かけ上全く異なる分野に共通の問題意識があることが確認でき、「仮想データ生成」というコンセプトの有用性が確認できた。

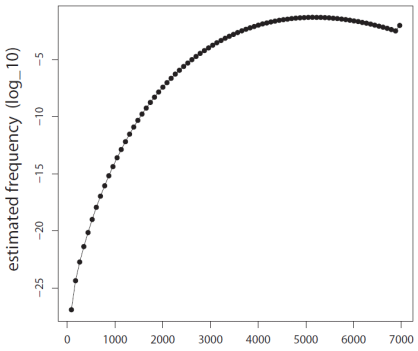
(2) サロゲートデータの生成
①マルチカノニカル法
仮想データ生成問題のうち、サロゲートデータの生成問題についてマルチカノニカルモンテカルロ法を適用することを試みた。この方法のポイントは「与えられた統計量が厳密に保存されたサンプル」に「保存性をある程度破ったサンプル」を適宜混ぜて発生させる点である。適応的な手法でこれをうまく行って、そこから統計量を厳密に保存するサンプルだけを抜き出すことで、理論上は偏りを発生させることなく、この問題に対するマルコフ連鎖モンテカルロ法の収束を桁違いに加速することができる。統計物理や光通信の問題では以前から応用されているが、サロゲートデータの生成問題への適用はこれがはじめてと思われる。
②例題
連続変数の時系列データに対して　(1) 各時刻での値の確率分布を保存する (2) 複数の時間遅れについて２次相関の値が与えられた精度 ε で実際のデータに一致する、の２つの条件をみたす仮想データの集合をマルチカノニカル法で生成する実験の結果を示す。この問題はフーリエ変換と位相のランダム化によっても近似的に解けるが、境界条件の扱いに問題があり、先に述べた Schreiber(1998) でもベンチマークとして用いられている。
③結果
下図に実験で得られたサロゲート時系列の例を示した。一番上がもとの時系列（観測データ。この場合はシミュレーションで作った人工データ）であり、のこりの３つが提案手法で生成されたサロゲート時系列である。横軸は時間。



この例で得られたサロゲート時系列の２次相関関数を多数のサンプルについて重ね打ちにしたものを以下に示す。横軸に時間遅れ、縦軸に相関の値を示す。実線がもとの時系列に対応する。実験では、時間遅れが 8 までの相関がすべて高い精度で一致するサロゲートデータの生成を目指したが、目標がよく達成されていることがわかる。



次のグラフはマルチカノニカルモンテカルロ法の計算過程で求まる「状態密度」の対数を表したものである。横軸は 2 次相関関数の一致度をあらわす量で、最も良い一致が左端に相当する。時系列をランダムに生成した場合には 10 のマイナス 25 乗といったきわめて稀な確率でしか得られない一致度のサンプルが生成されていることがわかる。



提案手法は、最適化法に基づく Schreiber の手法と比較して、 (a) 与えられた集合からの一様なサンプリングが原理的に可能、(b) 精度 ε の関数として仮想データの相対確率がわかる、という２点で優れていると考えられる。
④成果の発表
サロゲートデータ生成に関する研究成果は、統計数理研究所で行った２つの研究会、情報論的学習理論と機械学習研究会 (IBIS-ML)、および日本物理学会大会で口頭発表し、IBIS-ML の会議録に報告が掲載された。また、国際会議 BayesComp2012 (2012 年 6 月, 東京) の講演でも発表した。またこれらの結果を含

むマルチカノニカル法に関する欧文の総合報告を執筆し、2013 年 5 月に投稿した（査読中）。
⑤神経科学への応用については、専門家と議論を行ったが、具体的な結果を得るには至らなかった。さらに具体的に事例を詰めるとともに、必要な統計量の効率的な計算法を開発する必要があると思われる。

(3)プレイメージの生成
①研究協力
「非線形情報処理におけるプレイメージ生成」については、機械学習の手法による創薬手法の研究を行っている山下博史氏(総研大大学院生)、吉田亮准教授(統計数理研究所)との共同研究を行うことができ、より現実的な問題を扱うことが可能となった。
②例題
山下氏らが従来取り組んでいるＳＶＭとグラフカーネルによる化学構造式の判別問題についてプレイメージ生成問題を考えた。化合物の化学式が与えられたときに、その特徴を抽出し、薬効や副作用の有無を判断するのが、通常のデータマイニングである。これに対し、プレイメージ生成の場合は、特徴量や判別の結果を与えて、該当する化学式を逆に生成することを目指す。プレイメージ生成問題の中でも判別対象がグラフのような離散構造である場合は特に困難であると考えられる。
③解法
共同研究の中で、山下氏が従来から検討していた方法を発展させ、化学式の断片の集まりをマルコフ連鎖モンテカルロ法類似の手法によって合成するという手法を開発した。この方法は厳密には詳細つり合いを満たさないが、数値実験の結果は有望である。
④発展と今後の課題
本研究の成果に基づいて、民間企業から資金を導入しての共同研究が開始された。簡単な例の数値実験では有望な結果が得られたが、より高度で実用的な例に発展させる必要がある。また、レプリカ交換モンテカルロ法のような拡張アンサンブル法の導入も今後の課題である。
⑤成果の発表
国際会議 BayesComp2012 と国内会議 IBIS2012 のポスターセッション、日本統計学会で共同研究者が発表した。

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１件）
① 伊庭幸人 マルチカノニカル法によるレアイベントサンプリングとサロゲートデータ生成への応用 電子情報通信学会技術報告（IBISML）111(87), 43-50, 2011-06-1（査読無）

〔学会発表〕（計８件）
① 伊庭幸人 マルチカノニカル法とレアイベント生成について 研究会 「仮想データ生成とその周辺： 逆像問題，サロゲーション，秘匿」平成 23 年 1 月 7 日 統計数理研究所（立川）
② 伊庭幸人 レアイベントサンプリングに関する２，３の話題 統計数理研究所共同研究集会 「マルコフ連鎖モンテカルロ法とその周辺」 平成 23 年 1 月 11 日 統計数理研究所（立川）
③ 伊庭幸人 マルチカノニカル法によるレアイベントサンプリングとサロゲートデータ生成への応用 第 5 回情報論的学習理論と機械学習研究会（IBISML）2011 年 6 月 21 日 東京大学本郷キャンパス
④ 伊庭幸人 マルチカノニカル法によるサロゲートデータ生成 日本物理学会 2011 年秋季大会 2011 年 9 月 22 日 富山大学五福キャンパス
⑤ Yukito Iba Sampling rare events using multicanonical MCMC Bayesian Inference and Stochastic Computation 2012 workshop (BayesComp2012) 2012 年 6 月 22 日 Tokyo
⑥ Hiroshi Yamashita, Ryo Yoshida, Yukito Iba Preimage analysis of chemical structures Bayesian Inference and Stochastic Computation 2012 workshop (BayesComp2012) 2012 年 6 月 23 日 Tokyo
⑦ 吉田亮 山下博史 伊庭幸人 分子設計のカーネル逆像問題について:医薬品開発への応用 日本統計学会 2012 年 9 月 10 日 札幌
⑧ 山下博史 吉田亮 伊庭幸人 創薬を支援するデータ駆動型化合物設計 15 回情報論的学習理論ワークショップ IBIS2012)2012 年 11 月 8 日 東京

〔その他〕
研究会「仮想データ生成とその周辺： 逆像問題，サロゲーション，秘匿」ホームページ
http://www.ism.ac.jp/~iba/ken2011Jan.htm

６．研究組織
(1)研究代表者
伊庭 幸人（IBA YUKITO）
統計数理研究所・モデリング研究系・准教授
研究者番号：30213200