様式 C-19

科学研究費補助金研究成果報告書

平成　２３年　６月　７日現在

機関番号：12601
研究種目：若手研究（B）
研究期間：２００９　～２０１０
課題番号：　21700316
研究課題名（和文）

統計学及びゲーム理論に基づくモデルの組合せ法の研究

研究課題名（英文）

Statistical and game theoretical approach for combining models

研究代表者
白石　友一（Yuichi　Shiraishi）
東京大学・医科学研究所・特任研究員
研究者番号：70516880

研究成果の概要（和文）：
二値判別機の出力を組合せて多クラスの分類問題を実行する枠組みにおいて，有効な判別機の組合せの方法論の開発に関する研究を行った．また，様々な刺激条件下の遺伝子発現データを組合せ，ネットワークを推定するための有効な方法論を開発した．さらに二つの因子の組合せの効果に着目し，相乗効果を統計学的検定理論の立場から定義し，その統計量をマルコフ連鎖モンテカルロ法によって推定する方法論を提案した．

研究成果の概要（英文）：
A efficient approach for combining binary classifiers for multi-class classification problems was developed. Then, a new statistical method for inferring network from differently stimulated gene expression data was developed. Finally, a rank-based nonparametric statistical test for measuring the synergistic effects between two gene sets was proposed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,100,000	330,000	1,430,000
年度			
年度			
年度			
総　計	2,400,000	720,000	3,120,000

研究分野：情報学
科研費の分科・細目：情報学・統計科学
キーワード：統計的学習理論，スパース正則化学習，多値判別，遺伝子発現，相乗効果，バイオインフォマティクス

１．研究開始当初の背景

統計学や機械学習やパターン認識，時系列解析など様々な分野において，どういったモデルを仮定するかにより手法の精度が大きく変化してしまう状況が数多くある．このような不確実性に対応する一つの方法として，最適なモデルをクロスバリデーションやブートストラップといった手法により選択する方法がある．しかし，これらの手法はデータのばらつきによる影響を受けやすく，必ずしも正確な結果を返すとは限らないという問題点がある．近年新しく研究が進んでいるアプローチに，候補となるいくつかのモデルを何らかの方法で組合せて新しいモデルを構成するという方法論がある．特にパターン認識の分野において，最適なモデルを唯一つ選択する方法論に比べ，こちらの方法論の優位性の検証がなされている．また，多くのモデルの組合せ法の提案がなされており，テキスト解析やバイオインフォマティックスなど数多くの分野での応用が広まっている．

既存の多くの研究はパターン認識や人工知能の分野におけるものであり，必ずしも強い理論的なバックグラウンドを有していない．そのためにモデルの組合せ法に対する多くの結果は数値実験による検証であり，必ずしも理論的な正当性に基づいて手法が開発されているわけではないという問題点がある．こうした理由により，多くの手法が提案されているにも関わらず，どのような場合にどのような組合せ方法が有効なのかについての定性的な議論が進んでいなかった．

２．研究の目的

以上の背景から，本研究では統計学とゲーム理論を理論的土台とした汎用的なモデルの組合せ法の提案を行った．さらに提案した手法をタンパク質の構造解析や，遺伝子の発現量に基づいた判別解析などのバイオインフォマティックス分野で応用することを試みた．複数のモデルを組合せるアプローチの重要な部分問題として，多値判別の問題がある．判別するべきラベル数が三以上ある多値判別問題は，ラベル数が二つの二値判別問題に比べて性能の良い判別機を構成することが難しいことが知られていた．そのために，性能の良い二値の判別機を組合せて多値判別問題を解くという方法論が数多くの応用分野で盛んにおこなわれていた．

申請者は研究開始当初まで，統計学とゲーム理論の観点から，二値判別機の組合せによる多値判別問題の理論的な解析，および有効な手法の開発を行ってきた．本研究では焦点を「二値判別機の組合せによる多値判別問題」から，より一般的に「複数のモデルを組合せる問題」に広げ，これまでに申請者が行ってきた研究を発展させることを目的にした．

３．研究の方法

候補となるモデルの個数が多い場合には，推定に関係のないモデルを取り除き，できるだけ少ない数のモデルを組合せることで，精度の向上，メモリの節約をすることが求められる．そこで，モデルの組合せ法の学習の際に L1 ペナルティという特殊な正則化項を付加することなどの工夫を行うこと

で，モデルを選択しつつ最適化を行う方法論の提案を行う．また，「最適化手法自体の高速化」と，「効率的に最適化できる定式化の考案」の両面から学習のための手法の改良を行う．さらに，この手法の有効性を数値実験により検証することや，理論的な正当性を与えることを試みる．

さらに，バイオインフォマティックス分野においても，複数のモデルを組合せるというアプローチは広まりを見せている．においては，タンパク質の構造予測問題を，特徴量ごとに判別機を構成し，重みを付けて組合せるという方法論が提案されている．こうした問題でも，本研究で提案する方法論は有効である．例えば，スパースなモデルの組合せ法による方法論を用いることでタンパク質の構造にどの特徴量が効いているかを見ることが可能となり，生物学的に意味のある情報が得られる．本研究では，タンパク質の構造予測問題や遺伝子の発現量に基づいた判別分析，遺伝子ネットワーク推定の問題に本研究課題の過程で得られた手法を応用する．

４．研究成果

二値判別機の出力を組合せて多クラスの分類問題を実行する枠組みにおいて，有効な判別機の組合せの方法論の開発に関する研究を行った．提案手法は，既存の組合せ法と比べて性能の面で優れていること，group lasso typeのペナルティ項を加えた最適化行うことで最終的なクラスラベルの決定に必要でない判別機を除去できることなどの特徴がある．当研究の内容は正式に国際学術誌に採択されている．

また，複数のモデルの組合せという観点から，遺伝子の転写制御ネットワークを推定する統計手法の開発を行った．発生や細胞の癌化などのメカニズムを理解するためには，遺伝子同士の相互関係，すなわちネットワークの推定が不可欠である．単一の環境下において採取された遺伝子の発現プロファイルからは，どうしてもそれを説明するネットワークに冗長性が生じ，正確な推定が実行できないという問題点があった．そこで，複数の環境下における発現データを同時に扱い，それらを組合せてネットワークを推定するための統計モデルを開発した．また，提案手法を実際の生物データに適用し，生物学的な観点においても合理的な結果を導くことを示した．本研究の内容は，国際会議International conference on Intelligent Systems for Molecular Biologyや統計関連学会連合大会で発表され，国際学術誌に採録されている．

近年，ゲノム配列や転写因子結合部位，ヒストン修飾やクロマチン構造などのエピジェネティクス情報など多様なゲノム情報が大量に蓄積されつつあり，これらの因子がどのような協調性を持って遺伝子発現に機能しているかを解明することに期待が高まっている．本研究においては，多様なゲノム情報から，転写におけるゲノム上の因子の役割の解明に貢献することを目指した．特に二つの因子の相乗効果，すなわち単一の因子だけではなく，二つの因子が同時に存在して初めて機能する効果，に着目して研究を進めた．最初に二つの因子の相乗効果を統計学的検定理論の立場から定義し，その統計量をマルコフ連鎖モンテカルロ法によって推定する方法論を提案した．また提案手法を用いて，転写因子結合部位やヒストン修飾間の相乗効果を同定することを試みた．当研究の内容は，分子生物学会年

会で発表され，また国際学術誌に条件付き採録されている．

５．主な発表論文等

〔雑誌論文〕（計２件）
（査読あり）

1. Yuichi Shiraishi and Kenji Fukumizu, Statistical approaches for combining binary classifiers for multiclass classification, Neurocomputing, 2011, 74, 680-688.
2. Yuichi Shiraishi, Shuhei Kimura and Mariko Okada, Inferring cluster-based networks from differently stimulated multiple time-course gene expression data, Bioinformatics, 2010, 26, 1073-1081.

〔学会発表〕（計６件）

1. 白石　友一, A rank-based statistical test for measuring synergistic effects between two gene sets, 第33回日本分子生物学会年会・第83回日本生化学会合同大会, 2010年12月10日, 神戸.

2. Yuichi Shiraishi, Predicting regulatory networks from large scale time-course microarray data under several stimulation conditions on MCF-7 cells, 9th European Conference on Computational Biology, 2010年9月27日, Ghent, Belgium.

3. 白石　友一, Statistical approaches for combining binary classifiers for multiclass classification, 2009 年度統計関連学会連合大会, 2009年9月8日, 京都.

4. 白石　友一, Inferring cluster-based networks from differently stimulated multiple time-course gene expression data , 2009 年度統計関連学会連合大会, 2009年9月7日, 京都.

5. Yuichi Shiraishi, Inferring cluster-based networks from differently stimulated multiple time-course gene expression data, 17th Annual International conference on Intelligent Systems for Molecular Biology, 2009年6月29日, Stockholm, Sweden.

〔図書〕
なし
〔産業財産権〕
なし

〔その他〕
特になし

６．研究組織
(1)研究代表者
白石　友一（Yuichi　Shiraishi）
東京大学・医科学研究所・特任研究員
研究者番号：70516880