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　5 月 18 日現在

機関番号：１１３０１
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：21700247
研究課題名（和文）項目応答理論とマルコフ確率場を用いた確率的データマイニング理論の体系化

研究課題名（英文）Probabilistic data mining theory using item response theory based on Markov random field

研究代表者
安田　宗樹（YASUDA MUNEKI）
東北大学　大学院情報科学研究科・助教
研究者番号：20532774

研究成果の概要（和文）：

項目応答理論とは社会科学や心理学で発展してきた現代の統計的テスト理論である。当該研究計画は、従来独立に扱われてきた項目間に相関を持ち込み、項目間が複雑に関連し合う新しい項目応答理論モデルの構築とそのモデルに対する統計的近似計算アルゴリズム・統計的機械学習アルゴリズムを提案した。提案した項目応答理論モデルはニューラルネットワーク分野や機械学習分野などで知られているボルツマンマシンに類似の数理構造をもっているため、当該研究計画により得られた成果は項目応答理論の発展のみならず、更なる応用への展開が今後期待される。

研究成果の概要（英文）：

An item response theory (IRT) is a recent statistical test theory which has been mainly developed in social science and psychology. In conventional models of IRT, each item has been statistically independent of each other. In this research program, I have proposed a new probabilistic model of an IRT including correlations among items, and have proposed approximate techniques and machine learning algorithms for the proposed model. Since the model is mathematically equivalent to Boltzmann machines which are well known in the area of neural networks and the area of machine learning, the proposed methods can be applied to not only the IRT but also applications in those areas.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,100,000	330,000	1,430,000
2010 年度	1,300,000	390,000	1,690,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・感性情報学
キーワード：確率的情報処理・統計的機械学習理論・情報統計力学・項目応答理論・データマイニング・統計的近似計算理論・アルゴリズム

１．研究開始当初の背景

近年の通信インフラの整備による高速インターネット・携帯電話の個人レベルでの急速な普及により個々のユーザーが大量のデータを容易に入手することが可能となってきた。これによりユーザーは止めどなくやってくる膨大なデータからの必要な情報の抽出、すなわちデータマイニングという新たなる問題に直面することとなった。必要とされることは、膨大にあるデータから重要な情報を取り出し、その本質の理解を助けるような体系化された情報処理技術であり、そういったニーズによりデータマイニングや統計的学習理論をはじめとした情報処理技術の開発と体系化の学術的・社会的要請が高まってきている。また同時に得られた情報をいかに有効に活用するかという問題もある。データからの学習理論やその活用技術にはこれまで様々な手法が適用されてきているが、その一つとして近年注目を集め、盛んに研究されている確率的情報処理技術がある。確率的情報処理とは問題を確率という数学的枠組みを用いてモデル化し，各々の要素を確率的に取り扱うため柔軟な処理が可能となりそれゆえ頑健な結果を得ることができると期待されている。実際、確率推論の手法の一つであるベイジアンネットワークは現在様々な応用が模索され、また実用的なアプリケーションへと広く利用されつつある。

確率的情報処理の強みの一つは各要素間の確率的な関連性すなわち相関を自然な形で考慮できるというところにあり、要素間の関連性に不確実性がある場合には有効な表現方法の一つとなる。しかしながら、より複雑・大規模化していくと考えられるこれからの情報処理課題に対してその有用性が見込まれるような複雑な相関構造をもつ大規模な確率モデルは一般に厳密な解析が困難であり、また計算機を用いたとしても現実的な時間内で厳密に解決するための系統的な方法が開発されていないため、実用的なアルゴリズム開発やシステムの定性的性能評価などまだその手法が十分に確立されているとは言い難い。そういった状況を打開すべく最近統計物理などの他分野で開発されてきた統計的近似解析手法を確率的情報処理へと転用し、ビリーフプロパゲーション等の具体的な近似アルゴリズムの設計やシステムの定性的なパフォーマンスの評価に役立てるという流れが注目されてきている。

２．研究の目的

研究代表者がこれまで取り組んできた統計的学習理論は端的に言えば、データの背景に存在する本質的な相関構造を推定することであり、これは見方を変えればデータの本質の抽出というデータマイニング技術の目指すところと近いと考えられる。すなわち、研究代表者やその関連分野によって築かれてきた統計的学習理論の知見はデータマイニング理論に転用可能であると期待される。

近年、項目応答理論（または項目反応理論）という理論体系が世界で注目を集めつつある。これは評価項目郡への応答（回答）に基づいて、複数の被験者の特性（能力等）や評価項目の難易度などを同時に推定するための試験理論であり、実際に TOEIC や TOEFL においてはこの理論を用いたデータの解析が行われている。

とくに最近ではインターネット上でテストを実施することで大量のデータを回収する試みも様々の用途に応じて進められているため、この理論を用いたデータマイニング技術への期待はますます高まることが予想される。しかしながら項目応答理論はデータ間の相関を十分に考慮しているとは言い難い。そこで、マルコフ確率場と統計的学習理論か

ら得られた知見の項目応答理論への転用によるデータ間の相関を考慮した新たなデータマイニング技術の構築・体系化が可能であると考えられ、その確立を当該研究の最終的な目的とする。

本研究では本研究期間内に以下の点について具体的に明らかにしていく。

（1）**データ間の相関を考慮できる確率的データマイニング理論体系の構築と整備**

項目応答理論で与えられるデータマイニングモデルを基礎とし、そこにデータ間の相関をマルコフ確率場の知見を参考に導入し、データマイニングのための新たな確率モデルを提案する。本提案モデルはデータ間の関連性を考慮できる強力で汎用性の高いデータマイニングモデルになると期待できる。また、このモデルにおけるデータからのパラメータ推定問題にも取り組む。これは受け取ったデータに応じてモデルのパラメータを適応的に決める手続きであり、統計的学習理論で培われてきたこれまでの知見を活かし，確率的情報処理の観点からの定式化を目指す。また、データに応じてモデル自体の構造も適応的に決定するための確率推定理論の確立も目指す。

（2）**頑健な結果を得るための統計的近似理論の整備**

一般にデータ間の相関をもつ確率モデル（マルコフ確率場もその一つ）は厳密な解析が困難になり、問題クラスとしては NP-hard 問題として分類されるため、適当な近似によるアルゴリズムの設計が必須である。統計物理学による平均場理論を中心にある程度系統的な近似手法が提案されているものの、各々の近似手法の有効性はシステムに依存する部分が大きく、どのような近似手法がある特定の問題に対して適切かを選択するためには経験に頼ることが多いのが現状である。そこで、既存の統計的近似法の数理的構造をより詳しく解析して本質的な構造を抽出し、統一的な視点から統計的近似理論体系の整備を行う。それと同時に様々なシステムに対して頑健な結果を与えることのできる新たな手法の開発を目指す。

３．研究の方法

本研究計画ではデータ間の相関を考慮した確率的データマイニングモデルの構築とその理論体系の整備を目標としている。それを達成するために代表者は以下のプロセスが必要であると考える。

（1）現在のデータマイニング技術の中心の一つである項目応答理論を基礎にマルコフ確率場の概念を組み合わせ、相関を入れたデータマイニングを達成する確率モデルを構築し、モデルに対する数理的構造を詳しく解析することを通して確率的データマイニング理論の整備に取り組む。

（2）マルコフ確率場に対する統計的学習理論の枠組みを本提案モデルに転用し、データ間の相関を有意義に考慮することのできるデータ学習アルゴリズムを構築。ここではモデルパラメータの推定に留まらず、モデル構造の推定まで視野に入れたより強固な統計的学習理論体系の構築を目指す。

４．研究成果

本計画研究課題の遂行で、代表者は主に以下の項目に対する顕著な研究成果を得た。

（1）**データ間の相関をもつ確率モデル上での強力な統計的近似計算アルゴリズムの開発**

代表者はこの項目に対する２つの大きな成果を得た。１つ目は、確率伝搬法とよばれる統計的近似計算手法に線形応答近似とよばれる物理学由来の近似手法を組み合わせた、感受率伝搬法という近似計算アルゴリズムの更なる拡張に成功した。提案法は従来の手法の精度を大きく上回り、システムの構造に対しても比較的頑健な結果を与える。２つ目は、スピングラス理論の中でのレプリカ解析とよばれる解析手法を取り入れることによる、ランダムの影響が強く効いているシステム上での新しい確率伝搬法の提案である。提案法は従来の手法では解くことの難しいシステム上で有効に働く近似計算アルゴリズムである。

これらの研究成果は当該研究計画の項目応答理論の発展のみではなく、マルコフ確率場一般に適用可能であり、確率モデルの統計量を求めるというもっとも基礎的な問題を扱っているため、工学・情報科学・物理学などの多方面への貢献が期待される。

（2）**データ間の相関をもつ確率モデル上での強力な統計的機械学習アルゴリズムの開発**

代表者はこの項目に関して大きく２つ成果を得た。１つ目は、先に述べた感受率伝搬法の拡張手法を統計的機械学習に応用し、新しい学習アルゴリズムの開発に成功した。提案法はシステムの構造に対し頑健であり、性能も従来法を越える。また、データの生成モデルがスパースな構造を持っている場合、提案法による学習は極めて高い性能をもち、生成モデルの構造を抽出（すなわちデータマイニング）することが可能である。２つ目は密な構造のシステムに対する統計的機械学習のアルゴリズムである。相関等式とよばれる統計学的等式を利用して、学習の問題を扱いやすい凸最適化問題に近似的に変換し、密なシステムに対して有効にはたらくアルゴリズムを提案した。

これらの提案法は現在の機械学習分野の主流のひとつを成しているマルコフ確率場上で一般的に利用できるものであるので、得られた結果は当該研究の範囲を容易に超え得るものであると考えられる。また、最近物理学分野で始まってきている逆イジング問題も同様の数理構造をもっているため、今後の物理学への寄与も期待される。

（3）**項目間の相関をもつ項目応答理論モデルの提案**

代表者は当該研究計画の中心を担う新しい項目応答理論の確率モデルを提案した。提案モデルは従来独立に設計されていた異なる項目間に相関を導入し、機械学習分野で知られているボルツマンマシンと類似のモデルとして定式化した。さらに提案モデルに対する学習アルゴリズムの設計にも成功している。

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 10 件）

(1)Muneki Yasuda, Yoshiyuki Kabashima and Kazuyuki Tanaka: Replica Plefka Expansion of Ising systems, Journal of Statistical Mechanics: Theory and Experiment, P04002, 2012.(査読有)
(2)Muneki Yasuda, Shun Kataoka and Kazuyuki Tanaka: Inverse Problem in

Pairwise Markov Random Fields using Loopy Belief Propagation, Journal of the Physical Society of Japan, Vol.81, No.4, pp.044801, 2012. (査読有)
(3)Muneki Yasuda and Kazuyuki Tanaka: TAP Equation for Nonnegative Boltzmann Machine, Philosophical Magazine, Vol.92, Nos.1-3, pp.192-209, 2012. (査読有)
(4)Muneki Yasuda and Kazuyuki Tanaka: Deterministic Approximate Learning Algorithm for Boltzmann Machines using Correlation Equality, IEICE Transactions (D), Vol.J93-D, No.11, pp.2446-2453, 2010. (査読有)
(5)Muneki Yasuda and Kazuyuki Tanaka: Approximate Learning Algorithm in Boltzmann Machines, Neural Computation, Vol.21, No.11, pp.3130-3178, 2009. (査読有)

〔学会発表〕(計19件)

(1) 安田宗樹, 田中和之：項目間の相関を考慮した項目応答理論, 電子情報通信学会技術研究報告, Vol.111, No.483, NC2011-187, pp.387-391 (2012 年 3 月 16 日, 玉川大学)
(2)Muneki Yasuda: Replicated Plefka Expansion and its Application to Inverse Ising Problems, Interdisciplinary Applications of Statistical Physics & Complex Networks, (March4, 2011, KITPC, Beijing, China)
(3)安田宗樹, 田中和之: 改良された感受率伝搬法, 電子情報通信学会技術研究報告, Vol.110, No.265, IBISML2010-94, pp.59-63, 2010, 第13回情報論的学習理論ワークショップ (IBIS 2010) (IBISML 主催, 2010年11月4日, 東京大学)

〔図書〕(計2件)

(1)安田宗樹, 片岡駿, 田中和之(分担執筆)：コンピュータビジョン最先端ガイド3, アドコム・メディア株式会社, pp.137-179, 2010.
(2)田中和之, 安田宗樹 (分担執筆)：映像情報メディア工学大事典(映像情報メディア学会編), オーム社, pp.233-235, 2010.

〔産業財産権〕
○出願状況 (計0件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況 (計0件)

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
HP:
http://www.smapip.is.tohoku.ac.jp/~muneki/
研究業績等はすべて上記のホームページで公開している。

６．研究組織

(1)研究代表者
安田　宗樹 (YASUDA MUNEKI)
東北大学　大学院情報科学研究科・助教

研究者番号：20532774

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：