科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ５ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2012～2013
課題番号：２４７００１３５

研究課題名（和文）確率的潜在変数モデルを用いた分散学習と統合に関する研究

研究課題名（英文）Distributed and Integrated Learning Algorithm for Probabilistic Latent Variable Model

研究代表者
佐藤　一誠（SATO, Issei）

東京大学・情報基盤センター・助教

研究者番号：９０６１０１５５

交付決定額（研究期間全体）：（直接経費）　　3,400,000円、（間接経費）　　1,020,000円

研究成果の概要（和文）：確率的潜在変数モデルは、柔軟なモデル設計能力により様々な科学分野で注目を集めている。例えば、ソーシャルネットワークデータ分析では、隠れたコミュニティ構造を潜在変数として抽出することができる。しかしながら、確率的潜在変数モデルの学習は局所解を多く含む最適化問題として定式化され、一般的に難しい問題である。本研究では、確率的潜在変数モデルに対して、決定論的・確率的の２つの側面から効率的な学習アルゴリズムを提案する。１つ目は、周辺化変分ベイズ法に基づく学習アルゴリズムで、２つ目は、量子アニーリングに基づくアルゴリズムである。学術文書やネットワークデータの解析で提案手法の有効性を確認した。

研究成果の概要（英文）：Probabilistic latent variable models have attracted attention in many scientific fields because of their power and flexibility to model real world phenomena.Latent variable reveal the the underlying structure in data. For example, a probabilistic latent variable model for network such as socia l network enables researchers to analyze latent community in a network. However, learning probabilistic la tent variable model is difficult. Typically, learning probabilistic latent variable model is formulated by an optimization problem which has many poor local solutions. We provided an efficient two learning algori thms to find better local solutions. One is based on a collapsed variational Bayes inference, which is a d eterministic algorithm. Another is based on a stochastic search with quantum annealing, which is a stochas tic algorithm. We found that these algorithms outperformed existing methods in an academic paper analysis analysis and a network data anaysis.

研究分野：総合領域

科研費の分科・細目：情報学・知能情報学

キーワード：確率的潜在変数モデル　周辺化変分ベイズ法　量子アニーリング　Dirichlet Process　Bayesian Nonparametrics

１．研究開始当初の背景

データの生成過程に対し、潜在変数と呼ばれる確率変数を導入することで、データ中に隠れた情報を抽出する確率的潜在変数モデルの研究がデータ解析において幅広く用いられている。
トピックモデル及びその拡張手法として**ノンパラメトリックベイズモデル**と呼ばれるモデリング手法が注目を集めている。これらのモデルは、自然言語処理への応用をきっかけとして、そのモデリングの柔軟性から Web マイニング、評判分析、推薦システム、画像処理、音声・音響処理、バイオインフォマティックス、地理情報解析などさまざまな分野で用いられている。

変数を含むモデルの学習は多くの局所解を含む最適化問題となり一般的に難しい問題である。本研究では、このような問題に対して決定論的・確率的アルゴリズムの側面から効率的なアルゴリズムを提案する。

２．研究の目的

[決定論的方法]

確率的潜在変数モデルとして近年最も応用されているモデルにLatent Dirichlet Allocation (LDA)がある。LDAの学習アルゴリズムは様々な方法が提案されているが、周辺化変分ベイズ法と呼ばれる決定論的アルゴリズムの学習性能が良いことが知られている。周辺化変分ベイズ法は、解析的に解くことが出来ない積分計算を伴うため近似計算を必要とする。これはテイラー展開により計算可能であるが、特定の次数での近似がLDA の学習では性能が良いことが経験的に知られている。本研究では、この現象に対して理論解析すると共に、より効率的な学習アルゴリズムを提案する。

[確率的方法]

一般に確率的潜在変数モデルでは、学習アルゴリズムを構成する最適化問題に多くの局所最適解があり、大域解を得るのが難

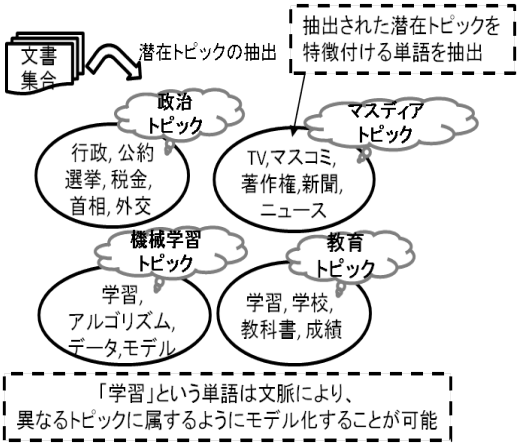

図 1 文書集合からの潜在トピック抽出の例

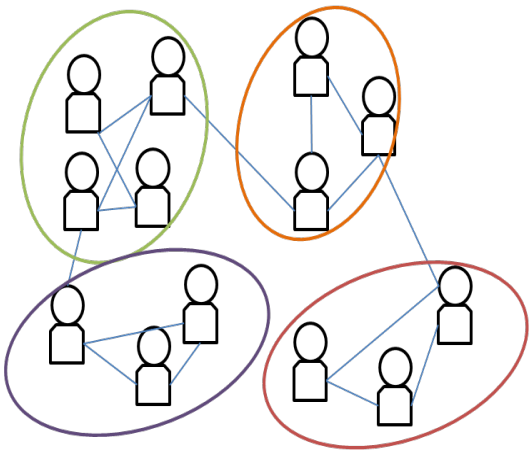

図 2 潜在コミュニティ抽出例

しい。したがって、近似アルゴリズムを用いる必要がある。本研究で扱う近似アルゴリズムは、確率的な探索アルゴリズムである。この手法では、解の候補を探索する際に必ずしも目的関数の値が良くなる方向へ探索するわけではなく、確率的な振る舞いを導入することで局所解から抜け出しより良い局所最適解を見つける手法である。本研究では、この古典的な確率的揺らぎに対して量子揺らぎを導入することで、局所解問題に対してより効率的な探索手法を提案する。

３．研究の方法

[周辺化変分ベイズ法の理論解析と効率化]

変分ベイズ法及び周辺化変分ベイズ法は Kullback-Leibler(KL) 情報量の下での最適化として導出される。しかし、KL 情報量のままでは、周辺化変分ベイズ法におけるテイラー展開の特定の次数の性能が何故容易のかは明らかにすることはできない。そこで、KL 情報量を一般化したα 情報量を用いて学習アルゴリズムを一般化し、パラメータであるα の挙動を考察することで、周辺化変分ベイズ法の挙動を理論的に解析した。

また、LDA のノンパラメトリックベイズ拡張である HDP-LDA の周辺化変分ベイズ法は、データの十分統計量の分散値を保持しておく必要があり、分散の計算は実装の複雑化を助長する。実は、LDA において学習能力の高い次数でテイラー展開の近似を行った場合、これらの分散の計算をする必要がないという利点があった。しかし、HDP-LDA では依然としてこの分散値を計算する必要がある。本研究では、学習効率の高い次数によるテイラー展開の近似で LDA と同程度に実装が容易で、分散値の計算を必要としない周辺化変分ベイズ法を開発する。

[量子揺らぎに基づく確率的探索]

量子情報処理技術の知見を活かし、量子揺らぎを学習アルゴリズムに導入する。量子揺らぎの導入によって導出されるアルゴリズムは、図3で示すように、複数プロセスの並列探索アルゴリズムによって近似できることを数学的に示した。これは、「学習の並列化」を「量子状態の重ね合せ(量子揺らぎ)」という概念で定式化したとも言える。

４．研究成果

[周辺化変分ベイズ法の理論解析と効率化]

周辺化変分ベイズ法においてテイラー展開を 0 次 近似した場合に学習効率が高いことが経験的に知られている。この 0 次周辺化変分ベイズ法は、α 情報量を用いた分析により、α → 1 とした場合の近似になっていることを示すことができた。α ≦ 0 の場合、「Zero-forcing effect」と呼ばれる効果が、学習に寄与することが知られているが、0 次周辺化変分ベイズ法はこの影響を受けないことがわかる

「Zero-forcing effect」は、推定する分布がモードに対してシャープになる傾向にあり、推定精度に悪影響を及ぼすことが知られている。したがって、周辺化変分ベイズ法におけるテイラー近似で特定の次数による近似(0 次近似)の学習性能が良いのは、この「Zero-forcing effect」の影響を受けないからであると言える。この成果は機械学習の最難関国際会議である ICML に採択された。

上記で行った周辺化変分ベイズ法の理論解析を元に LDA の拡張モデルである階層 Dirichlet 過程 LDA (HDP-LDA : Hierarchical Dirichlet Process enhanced Latent Dirichlet Allocation) における周辺化変分ベイズ法の研究を行った。LDA では、潜在変数の次元数を予め決めておく必要がある。HDP-LDA では、この次元数を推定することができる。HDP-LDA の周辺化変分ベイズ法はすでに提案されているが、LDA の場合と比べて近似計算に複雑な数値計算を伴うため、実装が複雑になる。また、データの十分統計量の分散値を保持しておく必要がある。例えば、文書解析においては、各単語毎にこの分散値を潜在変数の次元数だけ保持する必要があり、大規模データでは好ましくない。また、分散の計算は実装の複雑化を助長する。実は、LDA におい「Zero-forcing effect」を軽減する次数での近似を行った場合、これらの分散の計算をする必要がないという利点があった。しかし、HDP-LDA では依然としてこの分散値を計算する必要がある。本研究では、LDA と同程度に実装が容易で、分散値の計算を必要とせず、「Zero-forcing effect」を軽減する周辺化変分ベイズ法を開発した。応用として、学術情報のトピック解析やネットワークデータのリンク解析により従来手法に比べて性能が向上することを実験的に示した。この研究成果は、データマイニングの最難関国際会議である SIG-KDD に採択された。

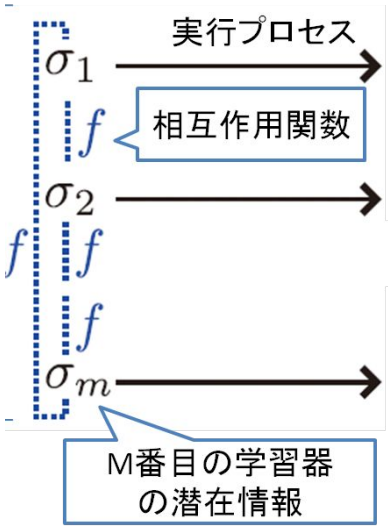

図 3　提案アルゴリズムの概要

[量子揺らぎに基づく確率的探索]

本研究で扱う確率的潜在変数モデルの学習は、図 2 で示す。ネットワークデータのコミュニティ抽出を応用例とする。ここで、ネットワークデータのコミュニティ抽出とは、各ノードをそのリンク関係から、いくつかのグループに分けることである。例えば、ソーシャルネットワークデータやWebリンク解析では、コミュニティとしていくつかのグループに分けられる。また、学術情報処理では、論文間の参照関係や共著者関係をネットワークデータとして記述されることが多いため、このようなネットワークデータ解析は重要であると考えられる。

また、ここで扱うネットワークデータの確率的潜在変数モデルでは、潜在コミュニティ数を予め指定する必要がなく、データに合わせて自動的に決定される。したがって、コミュニティ数決定も含めた最適化になるため、従来のコミュニティ抽出手法よりも難しさが 1 段増した問題となっている。

局所解を含む問題の解の探索では、初期値(今回の場合は初期クラス) を変えたm個のプロセスを走らせ、最もデータの対数尤度が高いクラス分けを解とすることが通常

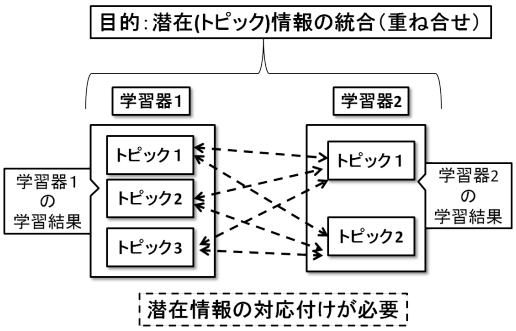

図 4 潜在情報の対応付け

行われる。このように従来の手法では、m個のプロセスを独立に走らせるのに対し、本研究で導出されたアルゴリズムでは、相互作用させながら探索を行う。複数プロセスで学習状態が複数の状態を同時に持つ量子揺らぎの効果をシミュレートすることで、量子効果としての相互作用により効率的な解を探索することができる。

確率的潜在変数モデルにおいて、このような複数並列の学習機の状態の重ね合わせをそれらの相互作用としてモデル化する場合、潜在変数の名寄せ問題が起こる。潜在変数の名寄せとは、複数の学習器が抽出した潜在変数の情報を統合する際に、同一の意味を持つ潜在情報と、異なる潜在情報を区別することを意味する。この問題は、潜在変数が非観測データであり、並列に実行した学習器ごとに結果が異なるため、潜在変数モデルの並列学習全般に常に生じる問題である。また、潜在変数モデルの学習が初期値に依存することにも起因する。

より具体的に以下説明する。潜在変数は非観測情報であり、その区別は図 1 で示したようにデータによって特徴付けられるため、潜在情報の対応付けが必要となる。図 4 に例を示す。例えば、学術論文のネットワークデータを考えよう。ここでは潜在コミュニティとして研究トピックが考えられる。学習器ごとに学習したトピックにはトピッ

ク番号が割り振られる。このトピック番号に対応付けられたトピックの情報は、データから抽出され、学習器ごとに異なるため、学習器間でのトピック情報(番号)の対応付けが必要となる。また、トピック数も、学習器ごとで異なる場合もあり、これらの対応付けはさらに難しくなる。例えば、「教育トピック」が存在する場合に、学習器１はトピック 3 を「教育トピック」として抽出し、学習器２は、トピック 1 を「教育トピック」として抽出する可能性がある。したがって、学習器１と２のトピック情報を統合する際には、トピック 3 とトピック 1 の対応付けが必要となる。

この問題に対し、本研究では、「各データがどの潜在変数の値を持つか」という従来のモデリングに対して、「各データがもっている潜在変数は、どのデータと同じか」というモデルに置き換えることで、潜在変数の名寄せ問題を解決した。具体的には、量子状態を記述する密度行列をデータ×潜在変数というモデリングから、データ×データというモデリングに対して構築することで、名寄問題を解決するアルゴリズムが数学的に導出されることを示した。

以下、実験結果について述べる。ネットワーク科学分野の論文共著者ネットワークデータセット(Netscience)に対してコミュニティ抽出を行った。各ノードはネットワーク科学分野の研究者でノード数は 1,589 である。また、論文参照データセット(Citeseer)に対してもコミュニティ抽出を行った。ノードは、参照・被参照論文でノード数は 2,110 である。

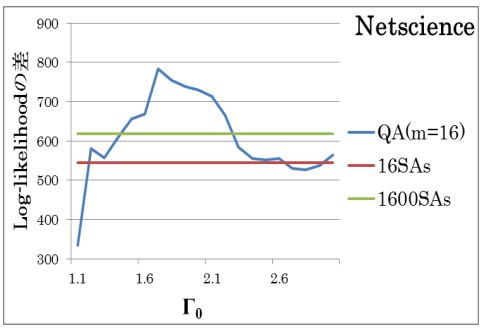

図５　論文共著者ネットワーク

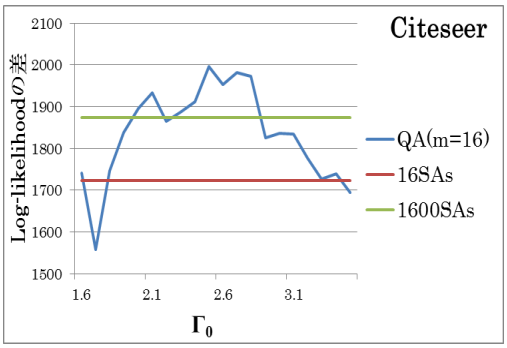

図６　論文参照ネットワーク

図 5,6 に実験結果を示す。本研究の目的は最適化であるため、評価尺度は目的関数である対数尤度を用いる。図 5,6 の縦軸は、Gibbs sampling と呼ばれる確率的探索をした場合の対数尤度と SA(Simulated Annealing)および量子揺らぎを利用した QA(Quantum Annealing)各々の手法の対数尤度の差を表す。したがって、高いほど優れた手法であることを意味する。我々の目的は出来るだけ早い計算時間で最適化を行うことであるため、Gibbs sampling の反復回数は 100 とした。また、1600 プロセス実行し結果の中でもっとも対数尤度が高い結果を選んだ。SA(Simulated Annealing)は、逆温度βによる確率的揺らぎの制御を行う手法でベースランとした。βは複数のスケジューリングを試し、最も良い結果となるものを選

んだ。16SAs は、初期状態の異なる 16 プロセスによる実験結果で得られた 16 の結果のうちもっとも対数尤度が高くなった結果を選ぶ。1600SAs は、1600 プロセスの実行結果の中でもっとも対数尤度が高い結果を選んでいる。SAs の反復回数は 30 とした。QA (Quantum Annealing)は、複数プロセスを相互作用させる我々の手法である。βは、SA と同じスケジューリングとした、つまり、SAs は、相互作用なし(f=0)の QA と見なせる。QA の反復回数は SAs と同様に 30 とし、並列数 m=16 とした。量子揺らぎの強さを制御するパラメータΓ_0を変えて実験を行った。Γ_0が大きくなるにつれ相互作用の効果が出る時間が遅くなる、つまり、反復回数固定のもとでは、相互作用が小さくなるため、QA の結果は同並列数の SAs に近づくことに注意されたい。したがって、実際の実験では、ある程度大きいΓ_0を設定して、徐々に減らしていくことで効果を確認することができる。図 5 からも分かる通り、この実験では$\Gamma_0 = 3$で、SAs と同程度の結果となり、減少させることで、100 倍の 1600SAs よりも対数尤度の高い結果が得られることがわかる。また相互作用の効果が反復回数の少ない段階で出てしまう場合($\Gamma_0 = 1.1$)、SAs よりも性能が悪くなることもわかる。

この研究成果は、機械学習の国際論文誌である Neurocomputing に採択された。

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）

Issei Sato, Shu Tanaka, Kenichi Kurihara, Seiji Miyashita, Hiroshi Nakagawa. Quantum Annealing for Dirichlet Process Mixture Models with Applications to Network Clustering. pp. 523–531. Neurocomputing, Vol. 121, 10.1016/j.neucom.2013.05.019, 2013

〔学会発表〕（計 2 件）

Issei Sato, Hiroshi Nakagawa. Rethinking Collapsed Variational Bayes Inference for LDA. 29th International Conference on Machine Learning (ICML 2012). pp. 999-1006. Beijing/China. June 26-July 1, 2012.

Issei Sato, Kenichi Kurihara, Hiroshi Nakagawa. Practical Collapsed Variational Bayes Inference for Hierarchical Dirichlet Process. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(KDD 2012). pp. 105-113. Edinburgh/Scotland. August 12-16, 2012.

〔図書〕（計 0 件）

〔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 なし

６．研究組織
(1)研究代表者

佐藤　一誠（SATO, Issei）
東京大学・情報基盤センター・助教
研究者番号：90610155

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：