課題番号 12-DA04

量子アニーリングを用いた機械学習およびデータマイニングの並列アルゴリズム開発

研究課題代表者　宮下　精二（東京大学）

概要

潜在変数を含む機械学習の学習アルゴリズムは、学習対象のデータを用いた非線形最適化問題として定式化される。この最適化問題には、局所解が数多く存在し、大域的な解を得ることは計算量的に不可能である。そのため解の近似的な探索方法が学習器の性能に大きな影響を与える。本研究では、量子揺らぎを用いた効率的な探索アルゴリズムを提案する。特に、このアルゴリズムは並列学習アルゴリズムとして実装することができるため、PC クラスタを用いた並列化の効果を明らかにすることを目的とする。

1. 研究の目的と意義

［研究目的］

本研究は、機械学習における量子揺らぎを利用した並列学習アルゴリズムを PC クラスタによって実現し、その効果を明らかにすることを目的とする。

統計的機械学習は、過去に蓄積されたデータに潜む有用な知識を、機械が利用可能な形で抽出し、新たな問題解決に役立てることを目的とする。ここで、データに潜む「有用な知識」をどのようにモデル化するのかが重要な鍵となる。統計的機械学習では、「有用な知識」としてデータ間の類似性に着目する方法が１つのアプローチとしてしばしば用いられる。この類似性を統計的に抽象化するために、「潜在変数」と呼ばれる確率変数を導入した確率的潜在変数モデルを用いる。データ間の類似性を、確率変数として記述し統計的に推定することで確率的変数モデルの学習が定式化される。本研究では、この潜在変数に対して、確率的な揺らぎとは別に量子揺らぎを導入した確率的潜在変数モデルの効果的な統計的機械学習理論の構築を目的としている。

確率的潜在変数モデルの学習アルゴリズムは、学習対象のデータを用いた非線形最適化問題として定式化される。ここで問題になるのは、解析的に解く事ができず、局所解が数多く存在するため、解の探索が学習器の性能に大きな影響を与えることである。

我々はこれまで、「量子揺らぎ」の概念を利用して、複数の学習プロセスを相互作用させ実行するアルゴリズムを提案した(図１参照)[1]。提案手法は、複数の学習器を実行し、互いに情報を相互作用させる手法である。このような学習手法は、古くはマルチエージェントによる学習として研究されてきたが、我々は、量子情報の理論を基に、複数の学習器のもつ情報を情報の重ね合わせとして定式化することで学習アルゴリズムを導出したという点で特色あるものである。

これまでの我々の研究では、実際に並列化実装を行わず、各々の学習を１プロセスで順次行う擬似的な並列化であった。アルゴリズムの性質上、実際に並列化実装を行った場合の評価は大変興味深いものである。特に、我々の手法は近似アルゴリズムであり、1/並列数のオーダーで近似精度が抑えられているため、並列数増加による振る舞いの解析が重要である。また、周辺分野への幅広い適用を考えると、スケーラビリティを向上させることが重要である。

本研究では、東京大学情報基盤センターの FX10 を用いることで実際の並列化を行い、提案アル

ゴリズムの効果を実データにおいて解析することを目的とする。応用分野としてソーシャルネットワークなどのネットワーク構造の解析を扱う。

図 2 にネットワーク構造解析の例を示す。ソーシャルネットワークなどの人間関係を表すネットワーク構造にはコミュニティ構造が潜んでいると考えられるこのようなコミュニティ構造を潜在変数として抽出することを目的とする。図 2 では色分けされたグループがそれぞれのコミュニティを表している。このような問題はグラフ分割としても考えることができる。確率的潜在変数によるアプローチでは、それぞれのノード（人）は、どのコミュニティに属しているかを示す潜在変数を持っていると仮定する。この潜在変数を推定することで、それぞれのノードがどのコミュニティに属しているかを推定することができる。また、この潜在変数の確率分布を推定することで、それぞれのコミュニティに属する確率分布を推定することができるため、各ノードを 1 つのコミュニティに限定しない推定を行うことができる。このような情報は単純なグラフ分割の手法では得ることができない。

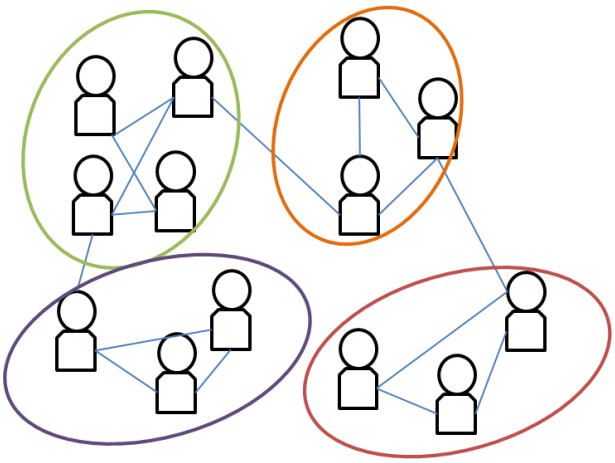

図 2 ネットワークのコミュニティ抽出

[研究の意義]

本研究は、「学習の並列化」を「量子揺らぎ」という概念で定式化している点で特色のある研究であると考えられる。そして、次のような研究の意義があると考えられる。

1. 並列化による学習効率の向上
2. 「機械学習」と「量子情報処理」の掛け橋となる学際的な拡張
3. 実データ解析への応用

すでに、擬似並列化環境では提案アルゴリズムによる学習の効率化を確認しており、実際の並列化環境で更なる効果が期待できると考えている。

さらに、本研究は、潜在変数を含む確率モデル全般に適用可能な学習アルゴリズムを目指しており、さまざまな応用が可能である。たとえば、隠れマルコフモデルのような系列構造を持つモデルに対しても適用可能である。

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名
東京大学（情報基盤センター）
(2) 共同研究分野
超大規模データ処理系応用分野
(3) 当公募型共同研究ならではという事項など
FX10 における並列アルゴリズム開発

これまでの研究では、擬似的な並列環境による実験を行なってきたため、実際の並列環境における効果については十分把握できていなかった。東京大学情報基盤センターの FX10 を用いることにより、並列環境での実験が可能となった。また、並列化やアルゴリズムの効率化に関して、東大情報基盤センターの研究者の方々と議論することで効率的に研究を進めることができていると考えている。

3. 研究成果の詳細と当初計画の達成状況

(1) 研究成果の詳細について

我々が学習対象として扱っているモデルは、確率的潜在変数モデルという機械学習の中でも学習することが難しいモデルである。確率的潜在変数モデルでは、データ中の隠れた性質(関係性など)を抽出することができるため、データ解析分野では幅広く用いたれているモデルであるが、潜在変数という非観測の確率変数を含むため、多くの局所解が存在し、学習の難易度が高い。このようなモデルに対して、「量子揺らぎ」を用いることで効率的に学習可能であることが我々の研究により明らかになっている[1]。

ネットワークデータの場合を例に潜在変数について以下説明を行う。図 2 の例のように各ノードが潜在的に属しているコミュニティを推定する問題を考える。各ノードが、コミュニティのインデックスを意味する離散値を取る潜在変数をもっており、その潜在変数の値が同じノードは、同じコミュニティに属するという仮定(モデル化) を行う。例えば、4 個のノード 1,2,3,4 の潜在変数をσ= (*z1; z2; z3; z4*) とする。σ= (1; 2; 2; 1) と推定された場合、ノード 1,4 がコミュニティ 1 に属し、ノード 2,3 が、コミュニティ 2 に属すると知ることができる。ここで、潜在変数が取りうる離散値の最大値を潜在変数の状態数と呼ぶことにする。このような推定は、ネットワークデータ *D* が与えられた下での、事後確率 $p(\sigma|D)$を最大にするσ を求めることに対応しており、具体的には以下の最適化問題に帰着される。

$$\sigma^* = argmax_\sigma \log p(\sigma|D)$$

ただし、多くの場合、$p(\sigma|D)$は解析的に求まることができず数値計算が必要となる。また、上記の最適化問題は局所解を多く含む非線形最適化問題となっているため、近似的により良い局所解をいかに求めるかが重要となる。

本研究では、この近似解法として、(1)確率的探索法、及び(2)決定的探索法(変分ベイズ法)に関して研究を行なっている。以下、各々について説明を行う。

「確率的探索法」

確率的探索では、解の候補を探索する際に必ずしも目的関数の値を改善する方向へ探索するわけではなく、確率的な揺らぎを導入することで局所解を回避しより良い局所最適解を見つける手法である。データ(ノード)i における潜在変数の割り当てをσ_i、その他のデータ（ノード）の潜在変数の割り当てを$\sigma^{\neg i}$と表現する。本研究では、$\sigma^{\neg i}$の値を固定した下で、ノード i が、潜在変数 k の値を取る確率$p(\sigma_i = k|\sigma^{\neg i} D)$は、解析的に求めることができるようなモデルを扱う。これにより、他のデータの潜在変数の値を固定し、1 つのデータの潜在変数の値をサンプリングすることで確率的探索を行う。このような方法は Gibbs sampling と呼ばれ、統計的機械学習で扱う確率モデルの多くが、この種の探索が可能なモデルである。実際の探索では、確率的な揺らぎを制御するパラメータを導入し最適化を行う Simulated Annealing を用いる。

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。我々の手法では、この量子揺らぎを複数プロセスの並列探索アルゴリズムによって近似的に導入する。

並列数を m, σ_j を j 番目のプロセスにおける潜在変数の状態とすると、量子揺らぎを考慮した確率的探索は、以下の最適化問題として定式化される。

$$\{\sigma_j{}^*\}_{j=1}^m = argmax_{\{\sigma\}_{j=1}^m} L\left[\{\sigma_j\}_{j=1}^m\right],$$

$$L\left[\{\sigma\}_{j=1}^m\right] = \sum_{j=1}^m \frac{\beta}{m} \log p(\sigma_j|D) + f(\beta,\Gamma) R\left[\{\sigma\}_{j=1}^m\right]$$

βは確率的揺らぎを制御するパラメータ、Γは量子揺らぎを制御するパラメータ、$R[\{\sigma\}_{j=1}^{m}]$は複数プロセスにおける潜在変数間の類似度を表す尺度、$f(\beta,\Gamma)$はこの類似度の影響力を制御する関数で、図１における潜在変数間の相互作用である。この最適化問題は、密度行列を用いて確率的潜在変数モデルを量子系に拡張することで導出される。したがって、$f(\beta,\Gamma)$及び$R[\{\sigma\}_{j=1}^{m}]$は、数学的に導出された具体的な関数であることに注意されたい。$L[\{\sigma\}_{j=1}^{m}]$は、$\frac{\beta}{m}$で重み付けされた古典系の目的関数の和と$\{\sigma\}_{j=1}^{m}$に関する制約項からなる目的関数とみなすことができる。

通常、確率的揺らぎを導入した場合でも局所解の問題は回避できないため、初期状態の異なるプロセスを複数実行し、その中で最適なものを選ぶ。これは、我々の最適化問題において関数$f(\beta,\Gamma)=0$ と置いた場合に相当する。これに対し我々の手法では、複数プロセスを実行する際に独立に実行するのではなく関数$f(\beta,\Gamma)$をβ及び Γによって制御することで相互作用させて実行させる方法である。

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)[2]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)[3]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。モデルとして Newman らのモデル[4]を採用した。また、近年統計的機械学習分野で注目を集めている Dirichlet 過程[5]を用いることで、コミュニティ数の自動決定を行った。

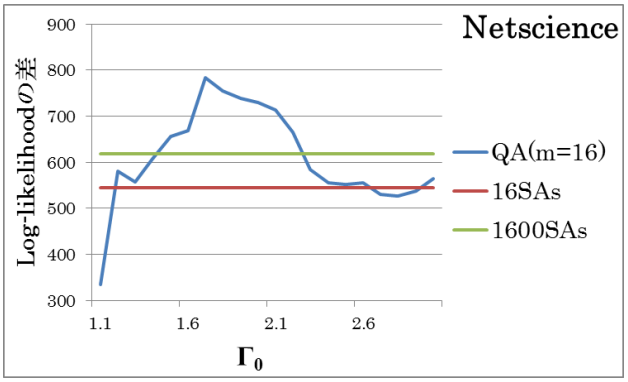

図３　論文共著者ネットワークにおける実験結果

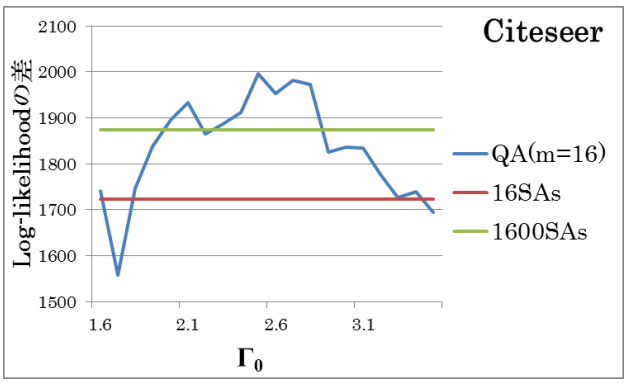

図４　論文参照ネットワークにおける実験結果

図 3,4 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 3,4 の縦軸は、Gibbs sampling を用いて確率的探索をした場合の対数尤度と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、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 3 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。本研究は、国際論文誌に投稿中のため、詳しくは採択後の論文を参照されたい。

「決定的探索法（変分ベイズ法）」

決定的探索法では、事後確率分布$p(\sigma|D)$を最大にする潜在変数σを求めるのではなく事後分布$p(\sigma|D)$の近似事後分布$q(\sigma)$を求める。確率分布間の類似度としてはカルバックライブラー情報量$KL[q||p]$が用いられ、解くべき最適化問題は以下のようになる。

$$q^*(\sigma) = arg\min_{q(\sigma)} KL[q(\sigma)||p(\sigma|D)]$$

近似事後分布$q(\sigma)$の構造を、その最大確率を容易に計算可能な分布と予め限定することで、事後確率最大の潜在変数の値も容易に求めることができる。ただし実際には、この最適化問題は$p(\sigma|D)$を含んでいるため解くことができないが、以下の関係を用いることで解析可能な最適化問題に帰着できる。

$$\log p(D) = F[q(\sigma)] + KL[q(\sigma)||p(\sigma|D)]$$

$\log p(D)$は、データの対数尤度で$q(\sigma)$に依存しないことから、$F[q(\sigma)]$を最大化することで$KL[q||p]$を最小化することが可能である。したがって、以下の最適化問題を解けばよい

$$q^*(\sigma) = arg\max_{q(\sigma)} F[q(\sigma)]$$

この方法は変分ベイズ法[6]と呼ばれ統計的機械学習で最も用いられている手法の 1 つである。$F[q(\sigma)]$は、$\log p(D) \geq F[q(\sigma)]$を満たすことから、変分下限と呼ばれている。または、物理における自由エネルギーのアナロジーから$-F[q(\sigma)]$は変分自由エネルギーと呼ばれている。$F[q(\sigma)]$に対する最適化問題は、EM アルゴリズムと同様に解析的に求まる各ステップ（E ステップと M ステップと呼ばれる）繰り返し計算によって解くことが可能である。ただし、$q(\sigma)$の初期状態によって求まる解が異なる局所解問題がある。

我々はこの変分ベイズ法に対して量子揺らぎの効果を導入する手法を提案した[1]。確率的探索法の場合と同様に複数の変分ベイズ法を相互作用させながら実行することで量子揺らぎを導入することができる。$q(\sigma_j)$を j 番目のプロセスにおける変分近似事後分布とすると、以下の最適化問題に帰着される。

$$\{q^*(\sigma_j)\}_{j=1}^m = arg\max_{\{q(\sigma_j)\}_{j=1}^m} F^{(m)},$$

$$F^{(m)} = \sum_{j=1}^{m} \frac{\beta}{m} F[q(\sigma_j)] + f(\beta,\Gamma) R[\{q(\sigma_j)\}_{j=1}^m]$$

この最適化問題は、密度行列を用いて確率的潜在変数モデルを量子系に拡張することで導出できることから、$f(\beta,\Gamma)$及び$R\left[\{q(\sigma_j)\}_{j=1}^m\right]$は、数学的に導出される具体的な関数である。また、実際には確率的探索法の場合とは異なる関数になっている。特に、$R\left[\{q(\sigma_j)\}_{j=1}^m\right]$は潜在変数間の類似度ではなく、事後分布間の類似度を表す尺度になっている。

ただし、我々の従来手法[1]では変分ベイズ法と同様の繰り返し反復法を導出するためには、以下の最適化問題を解く必要がある。

$$k'^{*} = argmax_{k'} \sum_{i=1}^{n} q(\sigma_{j,i,k}) q(\sigma_{j+1,i,k'})$$

$q(\sigma_{j,i,k})$は、プロセス j で、データ i の潜在変数が状態 k をとる確率を意味する。この最適化は潜在変数の状態数 K に対して $O(K^2)$の計算量となるこのため我々の従来手法では K の値を大きく取ることができなかった。そこで、本研究では以下の様な近似を行うことでこの問題を回避する。

$$k'^{*} = argmax_{k'} \sum_{i=1}^{n} \delta(\sigma_{j,i,k}) \delta(\sigma_{j+1,i,k'})$$

$\delta(\sigma_{j,i,k})$は、$q(\sigma_{j,i,k})$を最大にするkの場合に１その他を0とする指標変数である。この最適化は、元の最適化問題が事後分布によって定式化されているのに対しMAP(Maximum a posteriori estimation)に置き換えることで近似したことを意味する。これにより各事後分布のMAP解に変更があった場合のみこの最適化問題を解けばよく、変更前の結果を利用することで最適化問題自身も効率的に解くことができる。また、逐次的にこの更新を行うことができるため、分散環境では非同期にアップデートすることができる。

以下、実験結果について述べる。E-mailによるコミュニケーションネットワークデータセット[7]に対してコミュニティ抽出を行った。各ノードはE-mailによりやりとりのあったユーザーでノード数は4,793ノードである。なお、解析対象は、25人以上の他のユーザーとコミュニケーションをとっているユーザーに限った。ネットワーク構造の確率的潜在変数モデルとしてNewmanらのモデル[4]を採用した。変分ベイズ法を用いる場合は、予めコミュニティのクラス数を決定しておく必要がある。今回は、クラス数10と25で実験を行った。並列数は、16及び64で実験を行った。

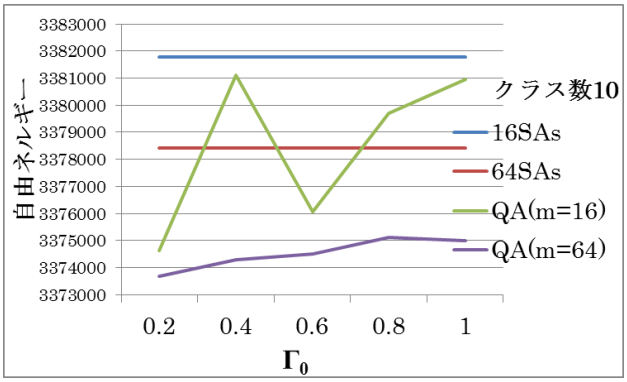

図５　Enron コミュニケーションネットワークにおける実験結果（クラス数１０）

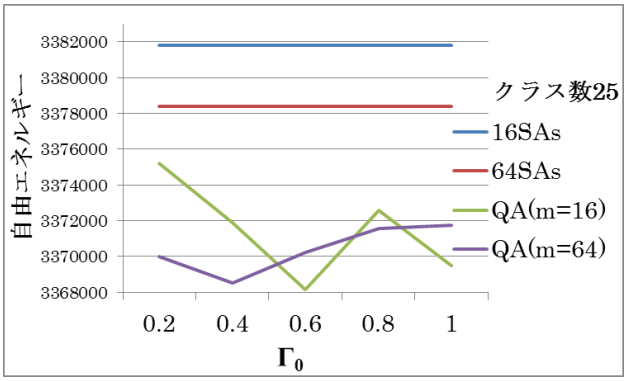

図６　Enron コミュニケーションネットワークにおける実験結果（クラス数２５）

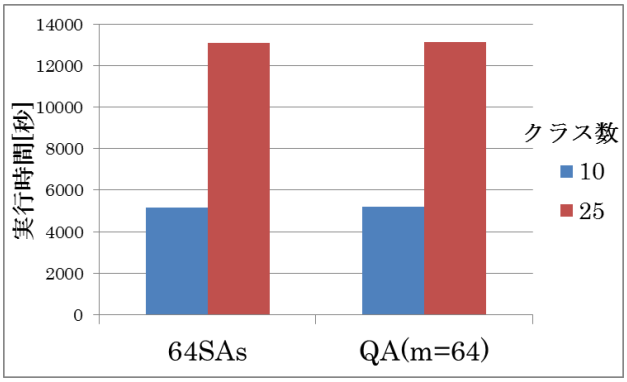

図７　実行時間の比較

図5,6に実験結果を示す。縦軸は、変分ベイズ法の目的関数であるFに対して符号を逆転させた値で自由エネルギーと呼ぶ。つまり、この値が小さいほど性能が良いことを示す。図5のクラス数10では、同一並列数で各々、従来手法(QA)が従来手法(SA)よりも性能が良いことが確認できる。また、量子効果を表すパラメータΓ_0の値によっては、並列数16の提案手法が、並列数64の既存手法を上回る性能となる場合があること

が確認できた。

図6のクラス数25の場合、並列数16の提案手法でも並列数64の従来手法を上回る結果を確認できた。しかし、提案手法間では、並列数64が必ずしも並列数16よりも性能がよいとは限らない結果となった。この理由として、クラス数と最適化問題の難しさの関係が考えられる。クラス数を増やすと、取りうる状態数が増えることから、最適化問題の解の探索が複雑になり問題そのものが難しくなる。取りうる状態数はクラス数に対して指数的に増えるのに対して、提案手法は近似精度が並列数に線形なため、並列数を線形に増やしたとしても有効に探索できる空間が限定されてしまうのではないかと考えられる。

関連して、図5,6を比較すると次のことが読み取れる。図5における64並列の従来手法と16並列の提案手法の関係のように、クラス数が少ない場合は、最適化問題は容易になるため、単純に並列数を増やせば従来手法は提案手法と同程度の性能を出すことができる。しかし、クラス数が大きくなると、問題の解の探索が難しくなり、並列数を増やしただけの従来手法と量子効果を導入して並列化した従来手法では差が顕著に現れる。

図7に実際の実行時間の結果を示す。縦軸が実行時間[秒]で、並列数が64の場合の結果をSA,QA各々比較した。従来の変分ベイズ法(SA)と同程度の速度で実行可能であることが確認できる。また、クラス数を増加させても、従来手法と提案手法の実行時間の差はほとんど見られないことがわかる。これは[1]で提案手法の計算量が$O(K^2)$であったことに比べて線形の計算量になっているため大きな進歩といえる。

これまでの議論をまとめると、量子効果を導入すると、問題を難しくした場合（取りうる状態数を多くした場合）に従来手法よりも効果が顕著になるが、並列数による差は小さくなると言える。このような考察はFX10を用いた並列計算によって初めてわかった知見である。

(2) 当初計画の達成状況について

これまで我々の研究では、シングルプロセスによって複数のプロセスを逐次的に繰り返す擬似的な並列化であった。当初の計画通り、複数プロセスを並列処理することによって、実際の振る舞いを観察することができた。また、これにより並列数を増やした場合の振る舞いとして、並列数とクラス数（潜在変数の取りうる状態数）との関係に対して新たな知見を得ることができた。更に、これまで我々のアルゴリズムがクラス数に対して２乗の計算量であったのに対して、線形で計算可能なアルゴリズムを開発できた。したがって、当初の目的は達成されたと考えられる。

4. 今後の展望

今回の研究では、主にネットワーク構造の潜在変数モデルによる評価を行ってきたが、今後の課題として、他のモデル(例えば、系列データのモデルである隠れマルコフモデルなど)にも適用を行い、量子効果導入の効果を確認することが考えられる。また、今回の研究で得られた、最適化問題の難しさと並列数の関係に関する知見は、クラス数（潜在変数が取りうる状態数）と並列数の関係であったが、確率モデルそのものが持つモデルの複雑さと並列数の関係を考察することは興味深い課題であると考えられる。

5. 研究成果リスト

学術論文

Issei Sato, Shu Tanaka, Kenichi Kurihara, Seiji Miyashita, Hiroshi Nakagawa. Quantum Annealing for Dirichlet Process Mixture Models with Applications to Network Clustering. Neurocomputing , 2013. (採録決定)

参考文献

[1] Issei Sato, Kenichi Kurihara, Shu Tanaka, Hiroshi Nakagawa, and Seiji Miyashita. Quantum Annealing for Variational Bayes Inference. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, 2009.

[2] Netscience dataset http://www.casos.cs.cmu.edu/computational tools/datasets/external/netscience/

[3]Citeseer dataset http://www.cs.umd.edu/projects/linqs/projects /lbc/index.html

[4] Mark Newman and Elizabeth Leicht. Mixture models and exploratory analysis in networks. In *Proceedings of National Academy of Sciences of the United States of America*, 2007.

[5] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*,, 1974.

[6] Hagai Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI1999)*,1999.

[7]Enron dataset http://snap.stanford.edu/data/email-Enron.ht ml