論 文 の 内 容 の 要 旨

論文題目　Machine Learning from Limited Information:
Approaches Based on Information Sharing
(限られた情報からの機械学習：情報共有に基づくアプローチ)

氏　名　山根　一航

序論（１章)： 近年，情報通信技術や計測技術の高度な発達や携帯端末の普及などにより，あらゆる場所で多種多様のデータが生み出されては蓄積されるようになった．また，計算機および計算機科学のめまぐるしい進歩により，従来実現できなかった様々な情報処理がより高速に，大きな規模で実行できるようになってきている．このような背景に後押しされ，コンピュータを使って自動的かつ効率的にデータを有用な知識に変換するための技術である**機械学習**がその重要性を増しており，活発な研究の対象となっている．囲碁・ポーカーの人工知能や機械画像分類などにおいて人間を超える性能が記録されたという象徴的な研究成果をはじめとして，コンピュータには苦手だとされてきた情報処理も最近の機械学習の発展により自動化・高速化・精密化が可能になりつつある．

これまでの機械学習技術の成功の裏には，質の高い多量のデータの存在がある．しかしながら，すべての実応用領域において，十分な量と質を兼ね備えたデータが用意できるとは限らない．さらなる応用分野の拡大のためには**限られた情報**から精度良く対象を学習する手法の開発が必要不可欠である．

そこで本論文では，訓練データが限られた情報しかもたない状況下での機械学習について議論する．このような状況としては，**量的に**情報が限定されている場合と**質的に**情報が限定されている場合の2種類が考えられる．ここで，量的に情報が限られている状況とは，データが十分にあれば学習可能な対象を，通常想定されるものよりも少ないデータで学習することが求められる状況である．一方，質的に情報が限られている状況とは，情報の種類が限定的であるがために，データが無数にあったとしても（さらなる条件が与えられない限りは）学習対象が部分的にしか特定され得ない状況のことである．

前者は様々な実問題において遭遇しうる普遍的な課題である．３章と４章では，量的に情報が限定されている状況において，**多次元対数密度勾配推定**（３章）や**マルチタスク**

主成分分析（４章）をどのように解くべきかについて議論する．後述するように，これらの問題に対しては**マルチタスク学習**の考え方に基づいた**情報共有のアプローチ**によるが有力であると考えられる．
質的情報の限定がある場合の学習は一見すると不可能に思われるが，５章で紹介する**不対ラベルを用いた向上作用モデリング**では，またも情報共有のアプローチにより問題を解決することができる．
本論文の貢献を以下にまとめる．（１）各出力次元をタスクとみなすことで多次元対数密度勾配推定をマルチタスク学習としてとらえ，タスク間で互いに情報の共有をしながら同時に解く方法を提案した．マルチタスク学習の適用においては，タスク間の情報共有方法の設計が非常に重要であるが，本研究では各次元が同じ関数の偏微分であるという事前知識を利用したモデルを使うことにより，データ分布に依存しない情報共有方法を可能にした．この内容は３章に記述した．（２）複数の似通った主成分分析タスクを，情報を共有しながら同時に解くマルチタスク学習手法を提案した．主成分分析では学習対象が射影行列であるため，従来マルチタスク学習で考えられてきた対象とは異なる幾何構造をもつ．本研究では射影行列全体がなす多様体上での最適化アルゴリズムを適用することで，精度良く，効率的に問題が解けることを実験により示した．この内容は４章に対応する．（３）処置と結果の２種類の教師ラベルのうちのどちらかが欠けたデータしか与えられない状況においても，ある条件を満たす２つのデータセット間で情報共有することで向上作用モデリングの問題が解けることを示した．さらに，多段階推定を避けて直接的に対象を学習する方法を提案した．解析では，提案法が実験的にも理論的にも高い精度を達成することを示した．この結果は５章にまとめた．

準備と関連研究（２章）： ２章では，論文全体にわたって必要となる機械学習の基本的な概念や定義を具体例とともに導入する．また，３章，４章，５章の内容に関連する様々な問題にも簡単に触れる．具体的には，まず，学習アルゴリズム，仮説関数，訓練データやテストデータ，および損失関数や期待損失などの重要な概念を導入しつつ関連する記号を整理し，学習の目的や結果の評価方法が一般的にどのように定義されるかを述べる．これらの基本的枠組みを整理した後，より具体的な機械学習問題を紹介するとともに，それぞれの問題においてどのような形式のデータが与えられ，どのような損失関数を使われるかを説明し，問題のより形式的な定義を与える．
また，本論文の主題である限られた情報からの学習の枠組みに深く関連する，**半教師付き学習**，**弱教師付き学習**や，**転移学習**の話題を紹介する．さらに，３章と４章で重要な役割を果たすマルチタスク学習の枠組みを紹介する．**マルチタスク学習**とは互いに関係のある学習タスクが複数ある場合に，情報を共有しながら同時に学習を進めることで結果の改善を図る枠組みである．

多次元対数密度勾配推定（３章）： ３章では，クラスタ数が未知の場合でも適用可能な**最頻値探索クラスタリング**やデータ分布の正規性の尺度の推定などに応用を持つ**対数密度勾配推定**について述べる．データ変数が多次元ベクトル値をとる（つまり多変量の）場合，学習対象である対数密度勾配は多次元ベクトル値関数になる．この問題は各次元（つまり偏微分）を別々の関数とみなして１つ１つ推定することで従来法が適用できる．しかし，これらの次元はすべて同じ原始関数に偏微分演算を施して得られるものであるという意味で互いに関連があり，ある次元の学習結果が別の次元の学習にとって有用な情報をもつということが期待される．そこで本研究では，各偏微分をタスクとみなした上で，情報を共有しながら同時に学習を進めることで，高い精度で対数密度勾配を推定するマルチタスク学習手法を提案する．具体的には，各次元が同じ関数の偏微分であるということから導かれる一般的な性質に基づいて設計されたモデルを使うことにより，データ分布に依存せず，タスク間の関係性について強い仮定を必要としない情報共有方法を提案する．
実験では，提案手法が高い精度で対数密度勾配を推定できることを人工データや実データにより示し，また，提案法に基づく最頻値探索クラスタリングでも同様に人工データや実データでの実験により高い性能を示すことを確認する．

マルチタスク主成分分析（４章）： ４章では，互いに似通った複数の主成分分析タスクがある場合に，マルチタスク学習のアプローチを用いてタスク間で情報を共有しながら同時に問題を解く手法を提案する．主成分分析の特殊な性質として，学習対象である射影行列の空間がユークリッド空間とは異なる幾何構造を持っていることが挙げられる．したがって，他の問題で良く行われるようにタスク同士の学習結果をユークリッド距離に基づいて近づけると，射影行列の類似性をうまく反映できない恐れがある．また，通常の行列全体の空間は射影行列全体の空間よりも大きな次元を持つため，射影子の行列表現を素朴に扱うと推定や最適化の効率性の観点からも無駄が生じてしまう可能性がある．そこで本研究では，射影行列全体がなす多様体上で直接的に最適化する手法を提案する．また，人工データや脳・コンピュータ・インターフェイスの実データ実験により，提案法が精度良く，効率的に問題を解くことができることを確認する．
不対ラベルからの向上作用モデリング（５章）： ５章では，限られた情報からの**向上作用モデリング**について考える．向上作用モデリングでは，特定の処置（例えば広告配信）による結果の変化（例えば購買量の増加など）に対する効果（**処置効果**）を推定することが中心的な課題となる．向上作用モデリングは，特定の処置をどの個体に対して行えば全体としての処置効果が大きくなるかという点に関心があるため，個体ごとに特徴量で条件づけた場合の処置効果，**個体処置効果**の推定が特に重要である．従来法では通常，処置と結果を表す２つのラベルの組が付与されたデータの存在を想定する．
しかし，このようなデータを集めるのは技術的，または倫理的な理由で難しい場合があ

る．例えば，Eメールによる広告配信の個体処置効果を調べたい場合，どのような顧客に広告を配信したのかは配信者にはわかるはずであるが，広告の受信者が実際に商品を購買するかどうかはマルウェアなどで追跡しない限り難しいだろう．一方で，実際に購買行動が起こる時には，顧客の情報が得られるかもしれないが，それが広告の結果なのかどうかわからないことも多い．

そこで本研究では，**処置とその結果の両方がわかっているデータは一つも与えられない**，質的に情報の限定された状況を考える．このような設定においては，1つの母集団から得られたデータから個体処置効果を推定することは一般的には不可能である．しかし，一定の条件を満たす２つの母集団から得られたデータが手元にあれば，推定可能であることが示される．

さらに，この場合，個体処置効果は簡単な回帰により得られる４つの推定量を組み合わせることで推定可能であるが，回帰問題の推定精度を改善しても，最終的な個体処置効果の推定精度が改善されるとは限らない．実際，実験ではこのような２段階推定は推定結果が非常に不安定になることが観測された．そこで本研究では，多段階推定をすることなく，個体処置効果を直接的に推定する手法を提案する．提案法の推定量は解析解により効率的に計算することができ，また理論的にも実験的にも高い精度を持つことが示される．

まとめと今後の展望（６章）： 最後に６章では，５章までの結論をまとめて述べ，今後の展望を考察する．

（4000字）