論 文 の 内 容 の 要 旨

論文題目

Machine Learning with Weak Supervision from Risk Minimization Perspective
（リスク最小化の観点からの弱教師付き機械学習）

氏　　名　　坂井　智哉

序論 (1 章)： コンピュータの性能向上, インターネットの普及に伴い, 膨大なデータが日々生み出されている．これらのデータから背後に潜む規則や構造などの有用な情報を抽出することで, 得られたデータを有効活用できるが, 膨大なデータを人間が扱うことは難しい．そこで, 人間のような知的情報処理をコンピュータに学習させて実現することを目指す**機械学習**と呼ばれる分野が注目を集めている．例えば，画像（入力）とその画像に写っている動物の名前（出力）とを集めたデータに対して機械学習を用いることで，その**入出力関係を分析**することや，入力画像を出力値に応じて**分類**することができる．

入力と出力の対応関係を，**教師付き学習**では，入力と出力の組（ラベル付きデータ）を多数用意して学習する．教師付き学習は非常に有用であるが，実用上は，出力となる**教師情報**を収集することに多額の費用がかかるという問題がある．そこで，教師情報を使わずに，入力のみのデータ(ラベルなしデータ)から上手く学習を行おうとする**教師なし学習**を用いることが考えられる．しかし，教師なし学習では，データに対する何らかの仮定，例えば，データがクラスタ構造を持つという仮定（クラスタ仮定）を置くため，仮定が成り立たない場合には高い性能が期待できない.

そこで，少ないラベル付きデータと多くのラベルなしデータ(ラベル付きデータの量が限られた弱い教師情報)を上手く用いて高い性能を得ようとする**半教師付き学習**が研究されてきた．半教師付き学習はラベル付きデータの収集に費用がかけられない際の有用な手段である．しかし従来法では，教師なし学習と同様に，クラスタ仮定のような，**データ分布に対する強い仮定**を置くため, 仮定が成り立たない場合に望んだ性能が得られないという問題があった．そのため，そのような**性能劣化のない半教師付き学習手法**が望まれている．

一方で近年，正例とラベルなし(PU: positive and unlabeled)データのみが与えられる状況から分類器を訓練する方法が盛んに研究されている．通常の教師付き二値分類とは異なり，PU データからの学習(PU 学習)では，分類したいクラスの内，どちらか一つ（これを正クラスとする）にしか教師情報が与えられず，もう片方のクラス（負クラスとする）には

教師情報が与えられない(正例だけの弱い教師情報)．代わりに，答えがわからないラベルなしデータが与えられる．例えば，ソーシャル・ネットワーキング・サービスでの友達関係は，友達であること（正例）は友達申請からわかるが，友達でないこと（負例）は申請しないため分からない．実用上は，しばしば，片方のクラスの教師情報しか与えられない状況が起きうるため，PU データから精度良く分類やデータ解析ができる手法が求められている．

そこで本論文では，上記に挙げた課題を解決するため，以下のような貢献をした．

1. PU 学習において，データ内の正例と負例の数に大きく偏りがある，不均衡データに対して分類器を訓練する方法を示した．従来法は分類器を訓練する際の学習基準であるリスク関数にバイアスがあるのに対して，バイアスのないリスク関数を導き，計算機実験によりその有用性を確認した．［3 章］

2. PU データのみが与えられる状況で利用可能なデータ解析手法を示した．入出力関係を測るために，二乗損失相互情報量（SMI: squared-loss mutual information）と呼ばれる統計的従属性尺度を利用した．PU データのみから SMI を推定する方法を示し，その推定量を次元削減や独立性検定と呼ばれる課題に適用する方法を示した．加えて，計算機実験を行い，手法の有用性を確認した．［4 章］

3. クラスタ仮定のような，データ分布に対する強い仮定を必要としない半教師付き学習の方法を示した．理論解析により，分布に対する強い仮定を一切用いずに，経験リスクの最小化が汎化誤差の上界を最小化することを導き，経験リスクが分散の意味で安定であることを明らかにした．加えて，計算機実験を行い実用における手法の挙動を詳しく調査した．また，手法の性能を多数のデータセットを用いて評価し，その有用性を示した．［5 章］

準備と関連研究（2 章）： 3 章，4 章，および 5 章で必要となる事項を説明する．

前半では，リスク最小化に基づく機械学習手法を復習する．その知識を基に，二値分類問題における教師付き学習手法(正例と負例からの学習，PN 学習とも呼ぶ)を説明する．続いて，PU データから学習する手法について復習する．また，PU 学習と対称関係にある，負例とラベルなし(NU; negative and unlabeled)データからの学習(NU 学習)についても紹介する．その後，分類問題において，PU 学習と PN 学習とを，それぞれの手法の推定誤差上界を基に理論的に比較した結果を紹介する．

後半では，まず，従来の半教師付き学習手法を紹介し，分布に対する様々な仮定が機械学習手法としてどのように定式化されていくのかを見ていく．次に，ラベルなしデータに含まれる正例と負例の比率（クラス事前確率）を推定する手法を説明する．クラス事前確率の推定法は，本論文の計算機実験で頻繁に利用される．その後，入出力関係を測る指標となる統計的従属性尺度を紹介する．

正例とラベルなしデータからの不均衡分類（3 章）： PU 学習の設定で，正クラスと負クラスのデータ数に大きく偏りがある不均衡データに対して分類器を訓練する手法を示す．

PU 学習において，これまでに，誤識別率最小化に基づく手法が提案され，その有用性が示された．一方で不均衡データに対しては，誤識別率を用いるよりも，AUC を用いて分類器を訓練する方が良いことが知られている．

まず, PU データから AUC に基づいたリスク(PU-AUC リスク)を導き, そのリスクが PN 学習のリスクと等価となることを示す．続いて，PU-AUC リスクを最小化することにより分類器を訓練する PU-AUC 最適化法を説明する．これまでに，PU データから AUC に基づいて分類器を訓練する方法が提案されていたが，リスクの観点から，従来法のリスクにはバイアスがあることを明らかにする．更に，PU-AUC リスクの汎化誤差上界を導き，経験 PU-AUC リスクの最小化が AUC により定義される汎化誤差上界の最小化に対応すること，その上界が最適なパラメトリック収束率で減少することを示す．

正例とラベルなしデータからの統計的従属性解析（4 章）： 二乗損失相互情報量（SMI）を用いた PU データのためのデータ解析手法について議論する．

PU データのみが与えられる状況で入出力関係を解析可能な手法はほとんどないため，PU データに対する実用的なデータ解析手法が望まれる．そこで，入出力関係を解析する方法として，変数間の統計的従属性を測ることができる SMI に着目した．SMI は，通常よく用いられる相互情報量よりも雑音や異常値に対して頑健であるという，実用上望ましい性質を持っている．

まず, PU データのみから, PN 学習における SMI と等価となる量(PU-SMI)を導出する．データから PU-SMI を推定する方法を示し，適当な条件の下で，PU-SMI 推定法の収束率を導く．続いて，従来の SMI を用いたデータ解析手法に倣い，PU-SMI 推定量を用いた次元削減と独立性検定の方法を示す．特に，次元削減法は，これまでの PU データから分類器を訓練する手法に必要不可欠な，クラス事前確率推定が不要となることを示す．クラス事前確率推定は，高次元データに対しては難しいことが知られている．しかし，PU-SMI 推定量を用いた次元削減法を用いることで，有用な低次元表現を得た後にクラス事前確率推定法を適用することができ，その精度向上が期待される．

正例，負例，ラベルなしデータからの学習（5 章）： クラスタ仮定といった，データ分布に対する強い仮定を必要としない半教師付き学習の方法に関して議論する．

従来の半教師付き学習手法の多くは，ラベルなしデータを学習に利用するために，クラスタ仮定や多様体仮定といった，データ分布に対する強い仮定を置いた．その上で，分布

に対する仮定を反映した正則化項を設計し，少数のラベル付きデータのみで推定された不安定なリスクを正則化することで分類器を訓練する．しかしながら実応用では，分布に対する強い仮定が成り立たないことが多いため，従来法は望んだ性能が出ず，場合によっては，少ないラベル付きデータだけで学習するよりも性能が劣化することがある．

そこで，データ分布に対する強い仮定を必要としない半教師付き学習の手法を開発する．鍵となる方法は，教師付き学習（PN 学習）におけるリスクと PU 学習におけるリスクとを組み合わせることである．ラベルなしデータを正則化に用いた従来法に対し，PU 学習に基づく手法では，ラベルなしデータをリスクの推定に用いることができる．特に，PN 学習に PU 学習または NU 学習を組み合わせた PNU 学習を提案し，その有用性を理論解析に基づいて議論する．提案した方法を用いて半教師付き分類法を開発し，その汎化誤差上界を導く．特筆すべき点として，データ分布に対する強い仮定なしに，汎化誤差上界が正例，負例，およびラベルなしデータの数に応じて最適な収束率で減少していくことが挙げられる．加えて，開発したリスクに関して，適当な条件の下，ラベル付きデータとラベルなしデータの両方を用いて推定したリスクの分散が，ラベル付きデータのみを用いて推定したリスクの分散よりも小さくなることを示す．

まとめと展望（6 章）： 3 章，4 章，5 章の内容に関する結論と，今後の展望について述べる．

（3,979 文字）