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０５月２２日現在

機関番号：１７１０２
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：21700308
研究課題名（和文）　機械学習による高次元小標本データ解析法の開発と暗号解読への応用

研究課題名（英文）　Development of method to analyze the high-dimensional and small-sample data based on machine learning and its application to cryptanalysis

研究代表者
川喜田　雅則（KAWAKITA MASANORI）
九州大学・システム情報科学研究院・情報学部門
研究者番号：90435496

研究成果の概要（和文）：
教師付き学習において例えラベル付きデータが小標本でもラベル無しデータが利用可能であれば推定精度を改善できることが知られている。このような学習は半教師付き学習と呼ばれる。大抵の既存の半教師付き学習は利用可能な情報が増えているにも関わらず、さらに何らかの仮定がなければ教師付き学習を改良できない。またモデル選択についてはラベル付きデータを用いて従来のモデル選択基準が用いられることが多い。しかしラベル付きデータが小標本であるため、標本数が十分多いことを仮定する従来のモデル選択基準はうまく働かない。本課題の主な成果はこれらの問題を解決したことである。
研究成果の概要（英文）：
It is already known that the estimation accuracy of supervised learning can be improved by using the unlabeled data even when the number of labeled data is quite small. This type of learning is called semi-supervised learning. The most conventional semi-supervised learning requires some additional assumptions to dominate the supervised learning even though we have additional information. Further, as for model selection, the conventional criteria (including AIC or CV) are applied to the labeled data. However, because such criteria require a large number of labeled data, they do not work well in this setting. Our main result is that we solved these problems.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００９年度 | 1,000,000 | 300,000 | 1,300,000 |
| ２０１０年度 | 900,000 | 270,000 | 1,170,000 |
| ２０１１年度 | 600,000 | 180,000 | 780,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 2,500,000 | 750,000 | 3,250,000 |

研究分野：総合領域
科研費の分科・細目：情報学･統計科学
キーワード：統計的学習理論，n<<p 問題，変数選択

## １．研究開始当初の背景

高次元小標本データの解析の一つのアプローチとして半教師付き学習が知られていた。半教師付き学習とはラベル付きデータが少数しかないような教師付き学習において、ラベル無しデータが大量に与えられたもとで教師付き学習の推定精度を改善することを目標にした学習である。既存のほとんどの半教師付き学習はクラスター仮定、低密度仮定、多様体仮定、特徴量分割仮定などの何らかの

仮定を必要とする。もしそれらの仮定が破れれば教師付き学習より悪化する可能性もある。半教師付き学習においては利用可能な情報が純粋に増えているのだから、これは奇妙な状況である。また、これらの仮定は一般に判別問題でのみ成立する。回帰においては似たような仮定が成立することは難しい。そのため半教師付き学習の研究は判別に著しく偏っている。
また大抵の既存の手法ではモデル選択に関しては AIC, cross validation などの従来のモデル選択法をラベル付きデータのみに適用している。しかしそれらの従来のモデル選択法の妥当性はラベル付きデータの数が十分大きいときのみ議論されている。半教師付き学習ではラベル付きデータが少数であることを仮定するためそれらの基準がうまく働く保証はない。ラベル無しデータの情報も活用可能でラベル付きデータが少なくともうまく働くモデル選択基準の開発が必要である。

２．研究の目的
半教師付き学習に焦点を当て、上記の問題点を解決することを目的とする。

３．研究の方法
密度比に基づく半教師付き学習のクラスに着目した。このクラスについて上記の問題を解決する方法があることを示した。また密度比に基づく方法は高次元にある程度の耐性を持つため、高次元小標本データの解析法として理に適っている。

４．研究成果
半教師付き学習について下記の成果が得られた。なおこのテーマの進展が顕著で重要であったため、このテーマに特化して成果をまとめたことを付記する。

1. 密度比に基づく半教師付き判別法が、「条件付きモデルが間違っている」とき特に他の仮定を必要とせず教師付き学習を改良できることが知られている。しかしその設定は特徴空間が離散有限、ラベル無しデータの数が∞、ラベルが二値（判別問題）、最尤推定量を用いるなど、様々な制約があった。本研究ではそのような制約を全て外し、上記の考え方がかなりの程度一般化可能なことを示した。またその理論的構造が統計学で知られるパラドックスと関連があることを指摘し、その幾何学的構造を明らかにした。その解析の結果、教師付き学習を改良できる条件として「ラベル無しデータの数がラベル付きデータより多い」ことが必要であることを示した。
2. １で一般化された密度比に基づく半教師付き学習法はラベル無しデータの数がラベル付きデータより多ければ教師付き学習を改良することが保証されていた。しかし、そもそもラベル無しデータが一つでもあれば教師付き学習より情報は増えているのだから、改良できる方法があるはずである。本研究では実際に上記の半教師付き学習法を改良してラベル無しデータが一つでもあれば教師付き学習を改良できる方法を提案した。
3. １や 2 の半教師付き学習法において密度比の推定は分母分子の密度の推定を別々に行っている．しかし近年密度比を推定する際には分母分子の密度を推定するのではなく、密度比そのものを直接推定する方が良いことが知られている。そのような密度比推定のパラメトリックなクラスを用いたとき１と同様に教師付き学習を改善できることを示した。また密度比のパラメータの最適な推定関数を導出した。さらに推定関数をあるクラスに制限したとき、密度比推定に用いるパラメトリックモデルは大きければ大きいほど教師付き学習を推定量の分散の意味で改良することがわかった。
4. 半教師付き回帰についてラベル付きデータが小標本でも有効なモデル選択基準 DEE が Chapelle らによって提案されている。しかし彼らの実験では DEE を Schuurman の ADJ と比較した結果、やや劣ることが示された。しかし我々は DEE の導出に二つの誤りがあることを発見した。一つは技術的な誤りであり、もう一つは論理的な誤りである。この二つの誤りを修正したところ、実験でも DEE は ADJ を上回る性能を出した。
5. 4 の考え方を拡張して半教師付き回帰について新たなモデル選択基準を提案した。４のモデル選択基準はラベル無しデータを興味パラメータの推定に活用していなかった。この研究では興味パラメータの推定に 2 や 3 で提案された密度比推定量を用いる。結果として得られたモデル選択基準は従来のモデル選択基準と異なりノイズの分布、分散が未知でもよく、ラベル付きデータの数が多いことも要求しない（漸近展開を用いていない）。そのため小標本データでも有効に働くことが期待される。実際いくつかの数値実験ではその有効性が確認された。
6. 2,3,5 を組み合わせた半教師付き回帰は真の回帰関数が複雑で、モデルがう

まくフィットできないときは教師付き学習を大きく改良できる。しかし真の回帰関数が十分シンプルでモデルが間違っている場合、理論とは異なり実際の実験では教師付き学習を改良できる確率は 50%、あるいはそれよりやや下がるときがある。2 や 3 の方法は確かに理論上、モデルが正しいときは教師付き学習と等価な性能になるはずであるが、それは漸近理論の結果であり、実際の有限標本ではやや教師付き学習より劣ってしまう。本研究では 2,3 の推定量の重みをデータから適切に推定することを提案する。具体的には推定された密度比のべき乗を重みのモデルとし、何乗すべきかをモデル選択基準によって選択する。このような方法は下平によって共変量シフトの枠組みで既に提案されている。しかし下平のモデル選択基準は漸近理論を用いて導出されているため、半教師付き学習の設定ではうまく働かない。我々は 5 のモデル選択基準を重みの選択に用いることで高い確率で教師付き学習を改良できる方法を開発した。結果として教師付き学習を常に高い確率で改良できる半教師付き学習の開発に成功した。また実験において以下の興味現象が観測された。2,3 の漸近理論からは「モデルが間違っていれば密度比を重みに用いるべきで、モデルが正しければ定数を重みに用いるべき」と示唆されているように見えるが、2,3,5,6 を組み合わせた方法で選ばれた重みはいずれの状況にせよ定数重みや、密度比そのものが選ばれることはまれであった。このことは重み付き尤度に基づく半教師付き学習における最適な重みは我々が想定しているクラス以外に存在する可能性を示唆しているとも考えられる。

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 0 件）

〔学会発表〕（計 11 件）

① 川喜田　雅則 ″情報幾何によるブースティングの性能の考察,″データ科学特別セミナー, 大阪大学, 7 月 15 日, 2009
② 川喜田　雅則, 竹内　純一 ″ベイズ推定を用いない曲指数型分布族の推定量の改善,″第 12 回　情報論的学習理論ワークショップ（IBIS 2009), 九州大学（病院地区百年講堂）, 9 月 19－21 日, 2009
③ Kawakita, M. and Takeuchi, J. ″A new method of improving predictive distribution without Bayes estimation,″情報理論とその応用シンポジウム 2009, 山口県湯田温泉　ホテルかめ福, 12 月 1-4 日, 2009
④ 川喜田　雅則, 竹内　純一 ″ラベル無しデータを利用した回帰の改良,″第十二回人口知能学会　データマイニングと統計数理研究会(SIG-DMSM), 統計数理研究所, 3 月 29 日-30 日, 2010
⑤ Kawakita, M. and Takeuchi, J. ″Semi-supervised learning in view of estimating functions,″ Information Geometry And Its Applications III, Aug, 2nd-6th, 2010
⑥ Kawakita, M., Oie, Y. and Takeuchi, J. ″A note on model selection for small sample regression,″ International Symposium on Information Theory and its Applications 2010, Taichung, Taiwan, Oct, 17th-20th, 2010
⑦ 川喜田 雅則, 竹内 純一 ″大標本仮定を必要としない半教師付き回帰のモデル選択,″ 第 13 回　情報論的学習理論ワークショップ（IBIS 2010, 東京大学駒場キャンパス, 11 月 4 日-6 日, 2010
⑧ 川喜田 雅則, 竹内 純一 ″密度比推定を用いた半教師付き回帰法の改良,″ 情報理論とその応用シンポジウム 2010, 長野県長野市信州松代ロイヤルホテル, 11 月 30 日-12 月 3 日, 2010
⑨ 川喜田 雅則 ″半教師付き回帰のためのモデル選択,″ 情報幾何学と統計多様体上の一般化共形構造の周辺, 東北学院大学多賀城キャンパス, 12 月 17-18 日, 2010
⑩ Kawakita, M. ″A class of semi-supervised regression based on density ratio estimation″, ENSEEIHT-Kyushu University Workshop on Data Mining and Media Processing, ENSEEIHT, Nov. 24-25, 2011
⑪ 川喜田雅則 ″安全な半教師付き回帰のクラスとそのモデル選択″, ミニワークショップ 統計多様体の幾何学とその周辺(3)／ 幾何学と諸科学の連携調査, 北海道大学, 12 月 1-4 日, 2011

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
川喜田　雅則（KAWAKITA　MASANORI）
九州大学･システム情報科学研究院･助教
研究者番号：90435496