研　究　報　告　書

「密度比推定による大規模・高次元データの知的処理技術の創生」

研究タイプ：通常型
研究期間：　平成 21 年 10 月～平成 25 年 3 月
研 究 者：　杉山　将

1．　研究のねらい

社会・経済構造の急激な変化，高齢化社会，環境問題，領土問題などへの迅速な対応が叫ばれている昨今，コンピュータで処理すべき情報量は爆発的に増加している．情報処理の立場からは，この問題は「データの数」，「データの次元」，「データ構造の複雑さ」が著しく増加していると解釈できる．このような背景のもと，データ処理に統計的な手法が積極的に用いられるようになってきており，特に**機械学習**と呼ばれる分野が学術界・産業界から大きな注目を集めている．本研究のねらいは，複雑な構造を持つ超高次元かつ莫大な量のデータから，その背後に潜む有用な規則を自動的に見つけ出す機械学習技術の新しいパラダイムを創生することである．

機械学習の目的は，与えられたデータからその背後に潜む一般的な規則を自動的に獲得することである．例えば，キーボードやマウスなどを使わずに脳波で直接コンピュータを操作しようというブレインコンピュータインターフェースにおいては，人間の脳波のパターンと人間の意志（カーソルを右や左に動かす等）との関係が学習の対象となる．このような機械学習の研究において，データの統計的な振る舞いを調べることは学習の効率を向上させるために非常に重要である．しかし一方では，データの確率分布そのものを推定することは避けるべきであるという考えがある．なぜならば，データの確率分布の推定は学習問題そのもの（上記の例では脳波と意思とのマッピングの学習）よりも本質的に難しい問題であり，確率分布の推定を経由することにより規則の学習の精度が大きく低下してしまうからである．この原理は数学者 Vapnik により提唱され，この原理に基づいたサポートベクターマシンとよばれるパターン認識アルゴリズムが，非常に優れた性能を示している．

本研究では，この Vapnik の原理を応用し，重点サンプリング，確率分布比較，相互情報量推定，条件付き確率推定などの様々な機械学習タスクを，**「確率密度比」**の推定により統一的に解決する枠組みを提案する．そして，密度比推定に基づいた新しい機械学習アルゴリズム群を開発し，その数理的な性質を明らかにする．更に，信号画像処理，ロボット制御，脳波解析，自然言語処理，生命情報解析などの分野に応用する．

2.　研究成果

（1）概要

Vapnik の原理に従えば，確率密度比の推定は確率分布の推定より容易に行うことができる（図1）．なぜならば，各密度が分かればそれらの比がわかるが，比が分かっても各密度は特定できないからである．そのため，密度比推定を通して機械学習を行えば，密度推定を用いるよりも精度が良くなると期待される．

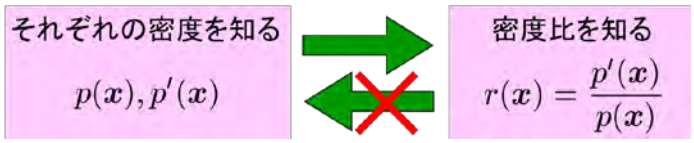

図１：密度推定と密度比推定

本研究の開始にあたって，重点サンプリング，確率分布比較，相互情報量推定，条件付き確率推定などの様々な機械学習タスクが，確率密度でなく確率密度比のみを推定することにより解決できることを示した．そのもとで，研究テーマ A「確率密度比の直接推定手法の開発」，研究テーマ B「密度比推定の数理的性質の解明」，研究テーマ C「密度比推定に基づく機械学習のアルゴリズム開発」，研究テーマ D「実応用」の研究を行った．

（2）詳細

研究テーマ A「確率密度比の直接推定手法の開発」

まずは，カルバック・ライブラー情報量に基づく密度比の直接推定法 KLIEP を開発し，密度比推定という新しい研究分野を切り開いた．その後，二乗損失に基づく簡便な密度比推定法 uLSIF を開発した．この手法により，密度比推定にかかる計算コストを大幅に軽減することができるようになり，様々な機械学習アルゴリズムへの応用可能性が広まったとともに，理論解析を行うための数理的な基礎も築いた．また，uLSIF 法に基づき，高次元空間での密度比推定に次元削減を組み合わせる新しいアルゴリズムも開発した（論文3）．

研究テーマ B「確率密度比推定の数理的性質の解明」

上記の KLIEP 法と uLSIF 法が，パラメトリック推定とノンパラメトリック推定の両方の枠組みにおいて，理論的に最適な近似精度を達成していることを証明した．更に uLSIF 法は数値的安定性の意味でも最適な手法になっていることを示した．また，KLIEP 法と uLSIF 法を含む，これまでに提案されてきた全ての密度比推定法を含む統一的な枠組みを考案し，各手法の相互関係を明らかにした（論文2）．

研究テーマ C「確率密度比に基づく機械学習のアルゴリズム開発」

大きく分けて，重点サンプリング，確率分布比較，相互情報量推定，条件付き確率推定に関する4つの研究を行った．「重点サンプリング」では，データを生成する確率分布が変化する非定常環境下での適応学習アルゴリズム（図2）とモデル選択手法，能動学習アルゴリズムなどを開発した．「確率分布比較」では，異常検出（図3），二標本検定（論文4），時系列の変化点検知などのアルゴリズムを開発した．「相互情報量推定」では，独立性検定，特徴選択，特徴抽出，独立成分分析，因果推定，オブジェクト適合，クラスタリングなどのアルゴリズムを開発した（論文1，図4）．「条件付き確率推定」では，条件付き密度推定，確率的パターン認識（論文5），マルチタスク学習，マルチラベル学習などのアルゴリズムを開発した（図5）．

研究テーマ D「実応用」

重点サンプリングに基づく機械学習手法を，音声からの話者認識，日本語文書の単語分割，ブレインコンピュータインターフェース（図6），ロボット制御などに応用した．確率分布比較に基づく機械学習手法を，製鉄プロセスの異常診断，光学部品の品質検査，画像中の注目領域の抽出，動画やツイッターからのイベント検出などに応用した．相互情報量推定に基づく機械学習手法を，遺伝子解析，製鉄プロセスの解析，医療画像処理，モーションキャプチャデータの解析などに応用した．条件付き確率推定に基づく機械学習手法を，製鉄プロセスデータの可視化，顔画像からの年齢認識，加速度センサからの行動認識などに応用した．顔画像からの年齢認識装置は2012年春に NEC ソフトより商品化された．

その他「広報活動」

本研究で得られた成果をまとめた英語専門書を MIT Press 社より出版した（著書1，図7）．応用分野を開拓するため，国内外の企業・研究所・大学にて60回以上の招待講演を行った．

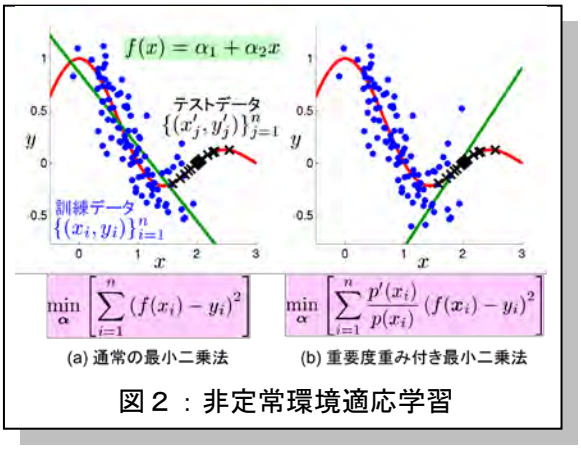

図２：非定常環境適応学習

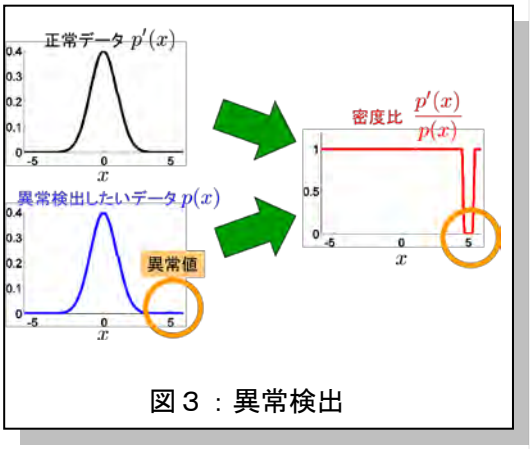

図３：異常検出

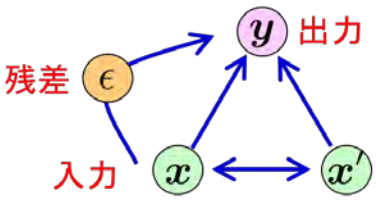

図４：相互情報量推定

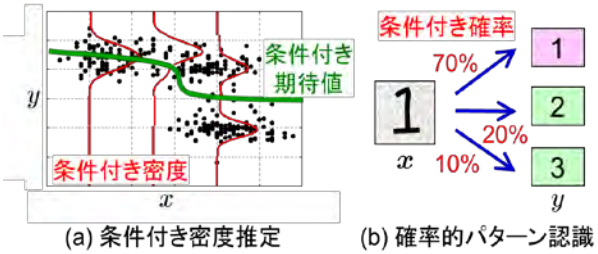

図５：条件付き確率推定

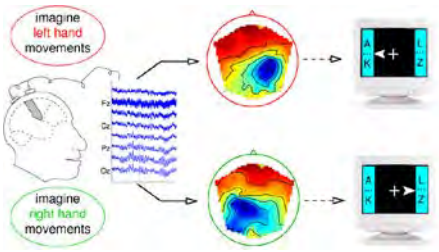

図６：ブレインコンピュータインターフェース

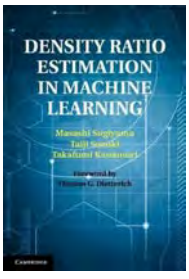

図７：英語専門書

3.　今後の展開

機械学習技術はこの数年で一気にコモディティ化が進み，研究者だけでなく企業の技術者にもそのキーワードが認知されるようになった．そのため，さきがけ研究期間中に多くの企業と共同研究をする機会に恵まれた．ビッグデータ時代のこれからも，機械学習技術の重要度は益々高まっていくものと考えられるため，今後も機械学習技術の実用化を更に加速させていきたい．

情報技術の応用パートナーといえば，数年前までは IT 系の企業が多かったが，近年は材料メーカや物流など多岐にわたる企業が機械学習技術に興味を持ちつつある．そこで，特に非 IT 系の企業とのコラボレーションを加速させて，新しい応用分野を開拓していきたい．その際には，基礎的な技術の普及が重要なカギを握るため，機械学習の入門書を執筆したり，チュートリアルをしたりという広報・教育活動も積極的に行いたい．

このように色々な分野の研究者や技術者が入り混じる機械学習分野における研究は，もはや個人で行うのは不可能だと感じられるほど裾野が広がってきている．そこで，基礎から応用まで幅広くカバーできるチームを作って研究をすすめていきたい．その際，現行技術の改良だけでなく，次世代の独創的なパラダイムを生み出せるよう地道な基礎研究をしっかりと行なっていきたい．

4.　自己評価

本研究プロジェクトでは，密度比推定という新技術を軸に，その技術が応用できる実問題を探すというスタイルで研究を進めてきた．幸い多くの企業に興味を持ってもらえ，多数の応用事例が見つかった．しかし，技術に基づくボトムアップ型の研究スタイルであるため，基本的には解ける問題を探して解いていたということになる．ビッグデータ時代には，これまでの技術では到底解決できないような難問が次々と現れることが予想されるため，世の中の風潮としては，応用に基づくトップダウン型の研究スタイルが好まれるようになってきていると感じている．

応用に基づいて研究が発展していく潮流そのものを否定するものではないが，このままでは技術に基づく研究がないがしろにされてしまうのではないかという懸念を強く感じている．情報技術は，色々な応用分野で活用されることによって，その研究成果を実世界に還元できるものであるが，情報技術そのものの基礎研究を軽視すべきではないと強く感じる．

本研究の成果は，密度比を使って簡単に解ける問題を見つけて解いたと解釈できる．しかし，問題がもともと簡単であったわけではなく，密度比を用いることによって，難しかった問題を簡単な形に変換できたため，結果としてうまく解けたのである．このような研究成果が得られたのは，密度比推定という技術が軸になっている研究計画に対して，さきがけの強いサポートを受けられたからである．応用研究がより重要視される中，今後も一定量の情報学の基礎研究をサポートする体制が存続して欲しいと心より願う．

一方，自分の作った理論を誰かが見出して応用してくれると信じている理論研究者が，自分の世界に閉じこもってしまっているのも事実である．従って，理論と応用の橋渡しをより積極的に進めていく必要がある．そのためには，前述したように基礎から応用までをカバーする研究チームを作ることが重要であると考えている．この理想的な研究体制が実現できるよう，更に自己研鑽を続けていきたい．

5.　研究総括の見解

データ解析の分野で、密度比推定という枠組みをさまざまな応用分野に適用可能にするという研究である。

この研究では、密度比推定という新技術を軸に、その技術が応用できる実問題を探すというスタイルで研究を進め、多くの企業と共同で多数の応用事例を解決している。また、この技術を普及させるための教科書も出版している。着実に成果をあげていることを評価したい。

6. 主な研究成果リスト

(1)論文(原著論文)発表

1. Sugiyama, M. Machine learning with squared-loss mutual information. Entropy, vol.15, no.1, pp.80-112, 2013.
2. Sugiyama, M., Suzuki, T., & Kanamori, T. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. Annals of the Institute of Statistical Mathematics, vol.64, no.5, pp.1009-1044, 2012.
3. Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. Neural Networks, vol.24, no.2, pp.183-198, 2011.
4. Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., & Kimura, M. Least-squares two-sample test. Neural Networks, vol.24, no.7, pp.735-751, 2011.
5. Sugiyama, M. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. IEICE Transactions on Information and Systems, vol.E93-D, no.10, pp.2690-2701, 2010.

(2)特許出願

1.
発　明　者: 植木 一也, 伊原 康行, 杉山 将.
発明の名称: 目的変数算出装置, 目的変数算出方法, プログラムおよび記録媒体
出　願　人: NEC ソフト, 東京工業大学
出　願　日: 2009年9月28日
出 願 番 号: 特願 2009-221989, 特開 2011-070471

2.
発　明　者: 平田 丈英, 河原 吉伸, 杉山 将.
発明の名称: パターン自動抽出方法およびパターン自動抽出システム
出　願　人: JFEスチール株式会社
出　願　日: 2010年5月25日
出 願 番 号: 特願 2010-119743, 特開 2011-247696

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

著書1:Sugiyama, M., Suzuki, T., & Kanamori, T. Density Ratio Estimation in Machine Learning, 344 pages, Cambridge University Press, Cambridge, UK, 2012

受賞1:2010年5月13日　電子情報通信学会 PRMU 研究会　研究会奨励賞

受賞2：2010年5月14日　人工知能学会 DMSM 研究会　研究会優秀賞
受賞3：2011年6月2日　情報処理学会　長尾真記念特別賞
受賞4：2011年12月16日　日本神経回路学会　論文賞
受賞5：2012年4月14日　船井情報科学振興財団　船井学術賞
受賞6：2012年6月19日　電子情報通信学会 IBISML 研究会　研究会賞ファイナリスト