密度比推定に基づく統計的機械学習

Statistical Machine Learning based on Density Ratio Estimation

杉山 将（ Masashi SUGIYAMA Dr. Eng.）

東京工業大学 准教授

（ Associate Professor, Tokyo Institute of Technology ）

電子情報通信学会　情報処理学会　人工知能学会　日本応用数理学会　日本統計学会　IEEE

受賞：財団法人船井情報科学振興財団 船井学術賞, 2012　日本神経回路学会 論文賞, 2011　精密工学会 技術賞, 2011　情報処理学会 長尾真記念特別賞, 2011　人工知能学会 研究会優秀賞, 2010　電子情報通信学会 研究会奨励賞, 2010　画像の認識・理解シンポジウム 優秀論文賞, 2008　財団法人安藤研究所 安藤博記念学術奨励賞, 2008　IBM Faculty Award, 2007　人工知能学会 研究会優秀賞, 2007

著書：杉山 将. イラストで学ぶ機械学習：最小二乗法による識別モデル学習を中心に, 講談社, 2013.　Sugiyama, M., Suzuki, T., & Kanamori, T. Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012.　Sugiyama, M. & Kawanabe, M. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT Press, 2012.　杉山 将. 統計的機械学習：生成モデルに基づくパターン認識, オーム社, 2009.　八谷 大岳, 杉山 将. 強くなるロボティック・ゲームプレイヤーの作り方～実践で学ぶ強化学習, 毎日コミュニケーションズ, 2008.

研究専門分野：統計的機械学習　データマイニング　信号画像処理

あらまし　インターネットやセンサーを通して、膨大な量のデータが容易に入手できる「ビッグデータ時代」が到来しつつある。近年、このような大量のデータから、いかにして有用な知識を得るかが重要な研究課題となっており、データの統計的な性質を活用する統計的機械学習が有望な情報処理パラダイムの一つとして注目されている。本稿では、統計的機械学習の基礎技術である「密度比推定」を紹介する。密度比推定とは、確率密度関数の比をデータから推定する技術の総称であり、密度比の推定を通して、非定常環境適応学習、特徴選択、クラスタリング、パターン認識、条件付き確率推定、独立成分分析、異常値検出など、様々なデータ解析タスクを統一的かつ優れた精度で解決することができる。

１．はじめに

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[1][2][3][4]。機械学習の技術は、コンピュータによる自然言語の理解、顔画像の認識、音声の識別、ロボットの制御、DNA の解析、脳機能の解明など、様々な分野に応用されている。

一口に機械学習と言っても、データ解析の目的によって、パターン認識、回帰、クラスタリング、異常検知、特徴選択など、様々なタスクが存在する。機械学習の最も汎用的なアプローチは、データを生成する確率分布を推定することである。なぜならば、データの生成分布を知ることは、そのデータに関する全ての知識を得ることと本質的に等価だからである。しかし、データの生成分布の推定は、統計学的に最も困難な問題として知られている。

一方、パターン認識や回帰などデータ解析の目的が具体的に定まれば、その目的を直接達成するアルゴリズムを開発することが最善のアプローチである。例えば、与えられたデータをそれが属するカテゴリに分類するパターン認識の問題では、各カテゴリに属するデータの生成確率をそれぞれ推定すれば、その確率に基づいて新しいデータが属するカテゴリを正しく予測できる。しかし、パターン認識のためには、必ずしもデータの生成確率を推定する必要はなく、異なるカテゴリ間の境界線さえ求められれば十分である。この考えに基づいて、サポートベクトルマシン[5]とよばれる最先端のパターン認識技術では、異なるカテゴリ間の境界線を直接求めることにより、高い認識性能を達成している。

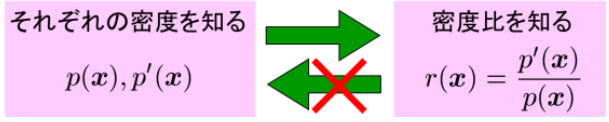
このようにして、パターン認識に対しては優れたアルゴリズムが開発されたが、パターン認識以外にも様々なデータ解析のタスクが存在し、それぞれに対して優れたアルゴリズムを開発することは困難である。実際、サポートベクトルマシンの基礎理論は、既に1960 年頃から研究され始めており、半世紀にも及ぶ研究開発を経て、様々な実データ解析に応用されるようになった。ビッグデータ時代には、大量のデータを前

密度比推定に基づく統計的機械学習

Statistical Machine Learning based on Density Ratio Estimation

に迅速に意思決定を行うことが望まれており、このような長期間に及ぶ基礎研究を行うことは現実的ではない。一方、最も汎用性の高い生成分布推定アプローチでは、精度良くデータ解析を行うことはできない。

そこで筆者らは、生成分布推定アプローチとタスク特化アプローチの中間を考え、ある条件を満たすデータ解析タスクのクラスに対して、アルゴリズムを開発するというアプローチを提案した。具体的には、複数の確率分布が含まれるデータ解析タスクのうち、それらの確率分布そのものは必要なく、確率密度関数の比さえわかればデータ解析を行うことができるというクラスを考えた。このクラスには、非定常環境適応学習、特徴選択、クラスタリング、パターン認識、条件付き確率推定、独立成分分析、異常値検出など、様々なデータ解析タスクが含まる。そして、この確率密度関数の比を、それぞれの確率密度関数を推定することなく直接推定するアルゴリズムを開発し、これら全てのデータ解析タスクを統一的、かつ優れた精度で解決することができるようになった（図 1）。



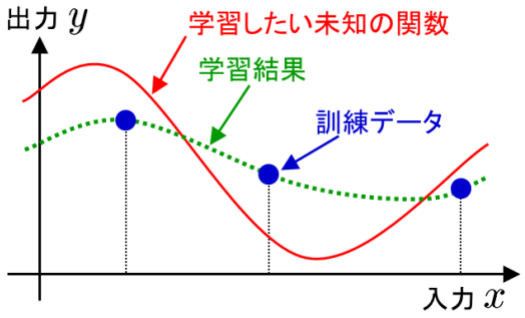
それぞれの密度を推定するよりも密度比を推定するほうが容易である。

図 1　密度比推定

この「密度比推定」の技術的な詳細は、専門書[6]および解説記事[7]をご覧頂くことにし、本稿では密度比推定によって、どのようなデータ解析が行なえるかを概観する。

2．非定常環境適応学習

入力と出力が対になったデータの背後に潜んでいる入出力関係を、推定する問題を教師付き学習とよぶ(図2)。この名称は、入力が生徒の質問、出力が教師の答えに例えられることによる。未知の入出力関係を学習することができれば、学習に用いていない新しい入力に対する出力を予測できるようになる。未知の状況に対して一般化できるということから、これは汎化能力とよばれる。この汎化能力の獲得こそが、教師付き学習の目的である。



入力と出力が対になった訓練データから、その背後に潜んでいる入出力関係を推定する。

図 2　教師付き学習

汎化能力の獲得を理論的に保証するために、学習に用いる訓練データと将来予測を行いたいテストデータが、同じ規則に基づいて生成されているという条件が一般的に仮定される。しかし、近年の機械学習の多くの応用分野では、この基本的な仮定が成り立たたない。例えば、脳波解析では、脳の振る舞いが時間と共に変化するため、訓練データとテストデータの傾向が異なる。音声や画像の認識では、訓練データとテストデータを収集する環境が一般に異なる。また、ロボット制御では、ロボットの行動規則が学習と共に更新されるため、結果としてデータの生成規則が変化する。

一方、訓練データとテストデータが全く別の規則に基づいて生成されると、訓練データからテストデータの情報を予測することは原理的に不可能である。従って、訓練データとテストデータをつなぐ何らかの仮定が必ず必要となる。共変量シフトは、そのような仮定の一つである[8]。共変量とは、入力データの別称であり、共変量シフトとは、入力データの生成規則が訓練時とテスト時で変化するが、入出力関係は変化しないという状況を指す。以下では、ブレイン・コンピュー

密度比推定に基づく統計的機械学習

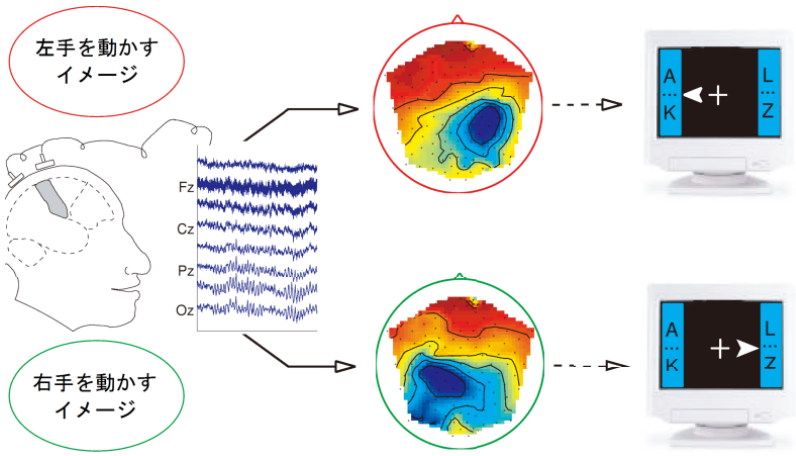
Statistical Machine Learning based on Density Ratio Estimation

タ・インターフェースを例に、共変量シフトに対処するための適応学習技術を紹介する。

ブレイン・コンピュータ・インターフェースとは、脳波によって計算機に意志を伝える技術であり、手足を動かすことのできない人でもコンピュータを操作できるようにするための重要な技術である（図 3）。

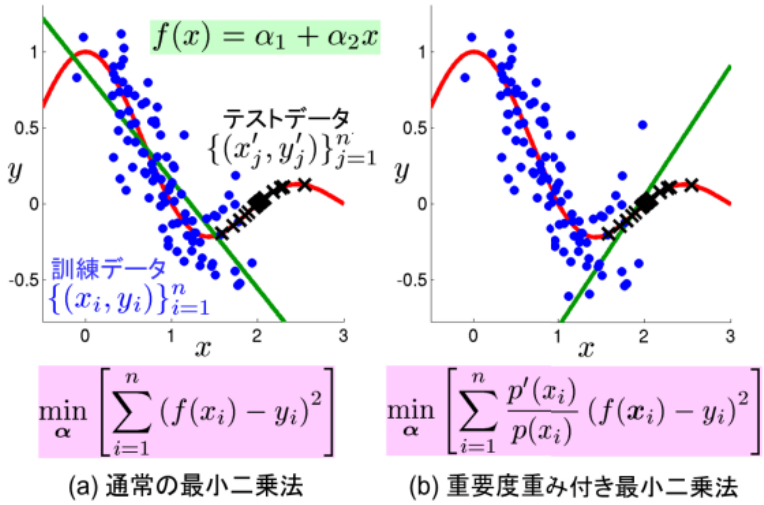
ここでは、脳波でマウスカーソルを左右に動かすタスクを考えることにする。学習の目標は、脳波パターンとその脳波によって伝えようとしている意志（左か右か）が対になった訓練データをもとに、脳波パターンと意志との関係を学習することである。これにより、将来与えられる脳波パターンによって示唆される意志を正しく予測できるようになる。ただし、脳の非定常性のため、学習用の脳波パターンと将来与えられる脳波パターンは、一般に異なる確率分布に従う。

共変量シフト適応の考え方を図 4 に示す。訓練用の脳波パターンとテスト用の脳波パターンの確率密度関数の比によって与えられる「重要度」に従って、学習規準を重み付けする。密度比推定によってこの重要度を推定することによって、共変量シフトに起因する非定常性に対して、適応的に学習を行なえる。



脳波でコンピュータを直接操作する。

図３　ブレイン・コンピュータ・インターフェース



(a) 通常の最小二乗法　　(b) 重要度重み付き最小二乗法

(a) 通常の最小二乗法ではモデルを訓練データに適合させるため、テストデータが訓練データと異なる確率分布に従う場合はテストデータをうまく予想できない。

(b) テストデータに近い訓練データに強い重みをつけることにより、テスト出力に適合させることができる。

図４　重要度重み付き最小二乗法による共変量シフト適応

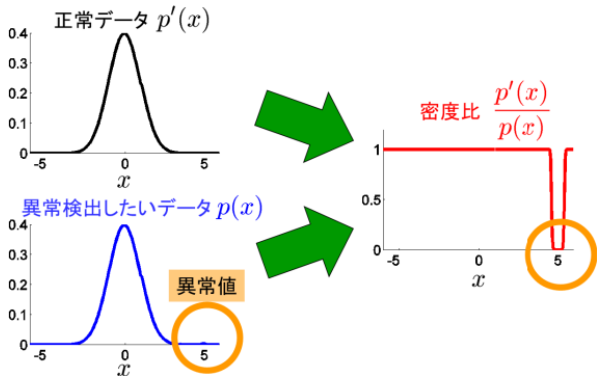
共変量シフト適応技術は、上述したブレイン・コンピュータ・インターフェースにおける非定常環境適応以外にも、ロボット制御における標本再利用や最適データ収集、自然言語処理におけるドメイン適合、顔画像からの年齢予測における照明環境適合、加速度センサーからの行動識別におけるユーザ適合、話者識別における声質適合、半導体露光装置の位置合わせなど、様々な実問題に応用されている。

３．確率分布比較

データ集合に含まれる異常値を見つける問題を異常値検出とよぶ。このような入力データだけからの機械学習問題は、前述の教師付き学習と対比して、教師なし学習とよばれる。一般に教師なし学習は、教師付き学習と比べてデータ解析の目的があいまいである。異常値検出も例外ではなく、どのようなデータを異常とみなすかを決めないと、主観的な議論に陥ってしまう。しかし、異常には様々なパターンが存在し得るため、あらかじめ異常とは何かを厳密に定義することは困難である。

そこで、逆に正常とは何かを定義し、正常でないものを異常とみなすことにする。具体的には、異常を発見したいデータ集合以外に正常データの集合が与えられると仮定し、もとのデータ集合のうち正常データから外れたものを異常値とみなすことにする。この考え方は、異常値を検出したいデータ集合の確率密度関数と正常データの確率密度関数の比を推定し、この比の値が１から大きく離れたデータを異常値とみなすことにより実現できる（図 5）。このような方式に基づく異常値検出は、光学部品の異常検出やローン顧客の審査などに応用されている。

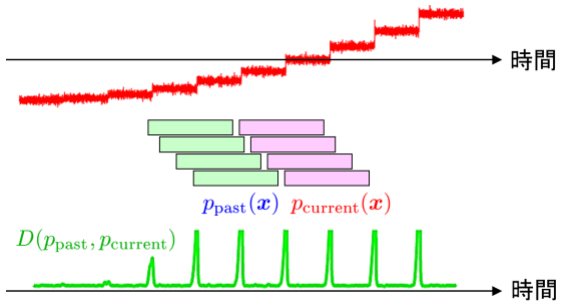
異常値検出は、二つの確率分布の一点を比較することに対応するが、二つの確率分布の全体を比較することも重要である。これは、二つのデータ集合が同じ確率分布から生成されたかどうかを判定する問題に対応し、二標本検定とよばれる。二標本検定は、例えば、カルバック距離やピアソン距離など、二つの確率分布間の距離がある閾値より大きいかどうかを判定することにより実現できる。二つの確率分布間の距離は、密度比推定により精度良く推定できる。二標本検定は、共変量シフトが起こっているかどうかの判定や、異なる状況で採取されたデータを合併して処理して良いかどうかの判定などに用いることがでる。



データ中の異常値を直接検出するのは困難だが（左下）、正常データの密度（左上）との比を取ることにより、異常値が強調され検出が容易になる（右）

図５　密度比に基づく異常値検出

また、過去の時系列データと現在の時系列データが従う確率分布間の距離を推定することにより、時系列の傾向の変化検出を行うこともできる（図 6）。



過去の時系列データと現在の時系列データの確率分布間の距離を推定することにより、時系列の傾向の変化を捉えられる。

図６　密度比に基づく変化検知

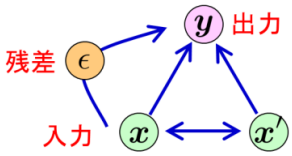
このような変化検出手法は、生体信号からの状態推定、画像中の注目領域の抽出、動画やツイッターからのイベント抽出などに応用されている。

４．相互情報量推定

入出力データが与えられた時、入力と出力に依存性があるかどうかを判定することによって、様々なデータ解析を行なえる。例えば、入力ベクトルの一部の要素が出力と独立であることがわかると、そのような要素は教師付き学習においては無視することができる。これは、出力の予測に役立つ入力変数ベクトルの部分集合を求めることに対応し、特徴選択とよばれる。特徴選択によりデータの解釈性が高まるため、例えば、遺伝子や脳波の解析に応用することができる。一方、出力の予測の精度を向上させるために、入力ベクトルを低次元表現に変換することを特徴抽出とよぶ。特徴抽出は、出力との依存性が最大の低次元表現を求めることにより実現できる。

入力データだけが与えられる場合でも、それらと最も依存性が高い出力ラベルを求めることにより、データのクラスタリングを行なえる。他にも依存性の推定により、ブラインド信号源分離、異ドメイン間オブジェクト適合、独立性検定、因果解析など、様々なデータ解析を行うことができる（図 7）。

二つの確率変数間の依存性は、それらの同時確率密度から周辺確率密度の積までの距離によって見積ることができる。例えば、カルバック距離を用いた相互情報量や、ピアソン距離を用いた二乗損失相互情報量がよく用いられる。これらの情報量は、密度比推定により精度良く推定できる。

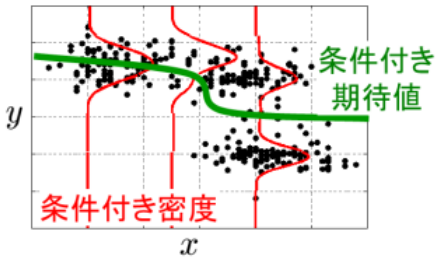


入出力間の独立性判定により、特徴選択、特徴抽出、クラスタリングが行なえる。同様に、入力間の独立性判定により、独立成分分析やオブジェクト適合が行なえる。また、入力と残差との間の独立性を判定することにより、因果推論が行なえる。

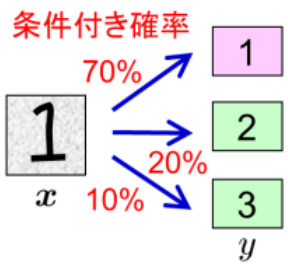
図７　相互情報量に基づく独立性判定

５．条件付き確率推定

回帰とよばれる教師付き学習では、連続値をとる出力変数の、入力が与えられたもとでの条件付き期待値を推定する。しかし、出力の条件付き分布が多峰性や非対称性を持つときは、回帰分析では十分な情報が得られないため、条件付き確率密度そのものを推定することが重要となる（図 8(a)）。



(a) 条件付き密度推定



(b) 確率的パターン認識

(a) 出力変数が連続値を取るときは、条件付き確率密度の推定に対応し、条件付き期待値を推定する回帰分析の一般化になっている。条件付き確率密度の推定は、出力の条件付き確率分布が多峰性や非対称性を持つときに有用である。

(b) 出力変数がカテゴリ値を取るときは、確率的パターン認識とよばれ、カテゴリの予測だけでなく予測したカテゴリの信頼度も同時に得ることができる。

図８　条件付き確率推定シフト適応

このような条件付き密度の推定は、データの可視化や移動ロボットの状態遷移確率などに応用できる。

一方、出力がカテゴリ値を取るとき、条件付き確率はカテゴリの事後確率を表すため、これを最大にするカテゴリを選ぶことによって、パターン認識を行なえる（図 8(b)）。このパターン認識法には、カテゴリの予測だけでなく、予測の信頼度も同時に得られるという特徴があり、顔画像からの年齢予測や加速度センサーからの行動識別などに応用されている。

条件付き確率は、その定義から、確率密度比の形で表すことができるため、密度比推定によって精度よく推定できる。

６．まとめ

本稿では、筆者らが開発した密度比推定に基づく統計的機械学習技術の概要を紹介した。密度比推定によって、様々な機械学習タスクを統一的に解決できるため、密度比推定の精度や計算効率を更に向上させることにより、様々な機械学習アルゴリズムの性能を一挙に改善できる。今後は、密度比推定の基礎技術を更に発展させていくとともに、密度比推定により解決できる新たなデータ解析タスクを開拓し、それらの機械学習技術を様々な実世界問題の解決に活用してくことが期待される。

密度比推定に関する論文やソフトウェアが、著者のホームページ

http://sugiyama-www.cs.titech.ac.jp/~sugi/

からダウンロードできる。

興味を持って下さった方は、そちらも合わせてご覧いただけたら幸いである。

参考文献

[1] 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇（編）. パターン認識と機械学習（上）: ベイズ理論による統計的予測, 丸善出版, 2007.

[2] 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇（編）. パターン認識と機械学習（下）: ベイズ理論による統計的予測, 丸善出版, 2008.

[3] 杉山 将. 統計的機械学習：生成モデルに基づくパターン認識, オーム社, 2009.

[4] 杉山 将. イラストで学ぶ機械学習：最小二乗法による識別モデル学習を中心に, 講談社, 2013.

[5] 赤穂 昭太郎. カーネル多変量解析：非線形データ解析の新しい展開, 岩波書店, 2008.

[6] Sugiyama, M., Suzuki, T., & Kanamori, T. Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012.

[7] 杉山 将. 機械学習入門. オペレーションズ・リサーチ, vol.57, no.7, pp.353-359, 2012.

[8] Sugiyama, M. & Kawanabe, M. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT Press, 2012.

この研究は、平成２０年度ＳＣＡＴ研究助成の対象として採用され、平成２１～２３年度に実施されたものです。