氏　　名	中 山 優 吾
学 位 の 種 類	博 士 (理 学)
学 位 記 番 号	博 甲 第　9362　号
学位授与年月日	令和2年3月25日
学位授与の要件	学位規則第4条第1項該当
審 査 研 究 科	数理物質科学研究科
学 位 論 文 題 目	

Support vector machine in high-dimension, low-sample-size settings
(高次元小標本におけるサポートベクターマシン)

主	査	筑波大学	教 授	博士(理学)　青嶋 誠
副	査	筑波大学	准教授	博士(理学)　小池 健一
副	査	筑波大学	准教授	博士(数学)　塩谷真弘
副	査	筑波大学	准教授	博士(理学)　矢田和善

論　　文　　の　　要　　旨

本論文は、現代科学に見られる高次元小標本データに対して、共分散行列の検定とそれに基づいた判別分析の問題を考えている。高次元小標本データは、例えば遺伝子発現データのように、数万という膨大な次元数をもつにもかかわらず標本数は数十程度の小標本であるため、大標本漸近理論に基づく従来の多変量統計解析では対処できない。膨大な次元数をもつデータを処理するためには計算コストが低く、さらに、小標本であっても推測精度が高い水準で保証されるような、新しい解析手法を開発することが望まれる。本論文は、機械学習分野で開発されたサポートベクターマシン(SVM)とよばれる分類器について、高次元データのスパイクノイズの検定から始め、高次元小標本における SVM の漸近的性質を明らかにしている。統計的な性質を導く上で、母集団の分布型や共分散行列の共通性は仮定していない。

本論文は、次の4つの章から構成されている。

第1章高次元小標本における共分散行列の球形性検定

第2章高次元小標本におけるハードマージン線形 SVM の漸近的性質

第3章高次元小標本におけるソフトマージン線形 SVM の漸近的性質

第4章高次元小標本における非線形 SVM の漸近的性質とカーネル関数の選択

第1章では、高次元小標本の共分散行列について球形性検定が考えられている。クロスデータ行列法というノンパラメトリックな手法を用いて、従来にはない新しい検定統計量が導出されている。帰無仮説のもと、検定統計量の分布に漸近正規性が証明されている。これをもとにして、検定の精度について与えられる有意水準と検出力を満たす検定方式が、二段階推定によって構築されている。検定方式は、高次元

データのスパイクノイズの検出に応用され、実データ解析における有用性が示されている。

第2章は、ハードマージン線形 SVM について、高次元小標本における漸近的性質を論じている。まず、高次元小標本においては、すべてのデータがサポートベクターになるという興味深い結果を導き出している。この結果から、2つの母集団の標本数についてある均衡条件を導き出し、この条件を満たさない場合にハードマージン線形 SVM は高次元データの判別に強不一致性をもつ、ということを証明している。つまり、ハードマージン線形 SVM は、2つの母集団の標本数が不均衡であると、高次元における誤判別確率が漸近的に1になってしまう。強不一致性は、そのような状況において判別手法が機能しないことを意味する。本論文は、強不一致性の原因を理論的に調査し、高次元小標本においてはハードマージン線形 SVM に巨大なバイアスが生じることを突き止めている。このバイアスを補正するために、新たにバイアス補正線形 SVM という判別手法を提案している。バイアス補正線形 SVM は、2つの母集団の標本数が不均衡であっても、誤判別確率が漸近的に0になる、という一致性が証明されている。さらに、バイアス補正線形 SVM は、母集団が 3 個以上ある場合の多クラス分類にも有効であることが示されている。

第3章は、ソフトマージン SVM について、高次元小標本における漸近的性質を論じている。ソフトマージン SVM の精度は、制約の強さを調整する正則化パラメータに大きく依存する。まず、正則化パラメータを介して、ハードマージン SVM とソフトマージン SVM の関係性を理論的に明らかにしている。ソフトマージン SVM は、ハードマージン SVM と同様に、2つの母集団の標本数が不均衡であると強不一致性が生じることを明らかにしている。そこで、不均衡な状況に頑健な判別手法となるようにバイアス補正法を与え、それが一致性を有することを証明している。

第4章では、非線形 SVM について、高次元小標本における漸近的性質を論じ、カーネル関数の選択基準を与えている。まず、高次元データの球面集中現象に着目し、母集団の判別に球の半径の差異を利用した非線形 SVM を考えている。高次元においてグラム行列がある特徴的な構造を有することに着目し、カーネル関数を利用した非線形 SVM の漸近的性質を一般的に与えている。その結果、非線形 SVM も2つの母集団の標本数が不均衡であると強不一致性が生じることを明らかにしている。そこで、バイアス補正非線形 SVM を与え、その一致性を証明している。カーネル関数の選択も研究され、多項式カーネルが線形カーネルと同様に 2 つの母集団の平均ベクトルの差異を利用するものであること、一方で、ガウスカーネルは平均ベクトルの差異だけでなく共分散行列の差異も利用するものであることを示している。非線形 SVM の精度は、カーネル関数に含まれるパラメータに強く依存する。ガウスカーネルの場合について、高次元小標本の漸近理論を用いて最適なパラメータの選択法が与えられている。最後に、癌患者の遺伝子発現データの判別に、最適パラメータをもつガウスカーネルに基づいたバイアス補正非線形 SVM を適用し、その有用性を交差検証法で確認している。

審　　査　　の　　要　　旨

〔批評〕

本論文は、高次元データの解析において、世の中に広く普及しているサポートベクターマシン（SVM）について、高次元小標本の漸近理論を用いることで性能を理論的に評価し、巨大なバイアスを補正することで SVM の性能を改善させ、さらに、カーネル関数の漸近的性質を明らかにすることでカーネル関数の

選択基準を与えたものである。SVM は機械学習分野で開発され、スパース分類器として広く使われているが、高次元小標本データに対する性能について、理論的な調査は必ずしも十分には行われていなかった。本論文は、高次元小標本データに対する SVM の性能を理論的に調査し、その漸近的性質から SVM の欠点を明らかにして、SVM の補正法を与えている。

漸近的性質は、高次元データのスパイクノイズの検定に基づいて導かれる。固有値のスパイク構造は遺伝子発現データなど多くの高次元データに見られる構造であり、スパイクノイズの検定は高次元データの推測に精度保証をする上で重要な位置づけになる。本論文は、スパイクノイズの検定法を共分散行列の球形性検定から導いており、大変に興味深い。通常、高次元小標本における共分散行列の検定は、計算コストと推測精度の両面で困難を極めるが、本論文はクロスデータ行列法と二段階推定法を駆使することで、それらの問題を解決している。

SVM の漸近的性質が、高次元小標本の漸近理論を用いて明らかにされている。特に、2つの母集団の標本数が不均衡な場合に強不一致性という最悪の判別結果が導かれることは興味深い。これは、高次元データの球面集中現象に起因しており、data piling による過剰適合が原因である。ハードマージン SVM に対して巨大なバイアスを評価しバイアス補正法を導く一連の方法論は、ソフトマージン SVM、そして、非線形 SVM に自然な形で拡張されている。導かれたバイアス補正法は計算が簡単であり、すでに広く普及している SVM の解析ツールに若干の修正を施すだけで済む。そのような簡単なバイアス補正であっても効果は絶大であり、実データ解析でしばしば直面する標本数が不均衡な場合にも、判別結果に一致性が保証される。この社会的な意義は極めて大きい。

本論文は、計算量が膨大となり得る高次元データ解析において、低い計算コストで高精度な統計解析に成功している。その一つの着眼点は、高次元におけるグラム行列の構造にある。グラム行列は SVM に限らずカーネル法全般に関わることから、今後、機械学習分野で高次元統計解析を展開する上で、重要な視点になるだろう。また、カーネル関数の選択について、高次元小標本において漸近的性質を論じることで基準を与え、さらに、カーネル関数のパラメータの選択について、交差検証法による膨大な計算を介することなく理論的に最適値を決定している。これらの新規なアプローチも大変に興味深い。

本論文は、SVM に理論に基づく簡便な補正法を与え、推測精度の保証と計算コストの削減という意義のある成果を与えている。学術的なインパクトは大きく、社会的な波及効果も期待され、大変に高く評価できる。

〔最終試験結果〕

令和2年2月12日、数理物質科学研究科学位論文審査委員会において審査委員の全員出席のもと、著者に論文について説明を求め、関連事項につき質疑応答を行った。その結果、審査委員全員によって、合格と判定された。

〔結論〕

上記の論文審査ならびに最終試験の結果に基づき、著者は博士（理学）の学位を受けるに十分な資格を有するものと認める。