博 士（情報科学） 池 田 盛 一

学 位 論 文 題 名

情報量規準に基づくモデル選択を用いたカーネル多変量解析法に関する研究

学位論文内容の要旨

近年インターネットをはじめその他の多くの分野で、高次元で多種類のデータの混在した大量のデータを扱うことが多くなった。そうした中で、それらのデータの隠れた構造を明らかにし、真の分布を推定することによって、予測・制御・情報の抽出・検定リスク評価・意思決定には適切な統計モデルを構成することがますます重要になってきている。

古典的な多変量解析手法においては、データの構造に線形性を仮定しているため、複雑な構造をもつデータの特徴を的確に捉えることはできなかった。説明変数に非線形の項を取り入れたり、また、非線形変換を用いることで線形ではない構造を見出そうとする接近法も数多く提案されているが、これらの手法においても、モデルに取り入れることができる非線形の項は、多項式や三角関数などあらかじめ想定されたものであるため、高度な非線形性を有するデータには適用が難しいという限界がある。

一方、ニューラルネットワークやカーネル法など、モデルを適応的に決定する方法がデータの予測や判別に用いられ、古典的な多変量解析では対応できなかった高度の非線形性をもつデータのモデリングを行うことが可能となった。これらの手法によるモデルは推定するために用いるデータに対しては当てはまりがよいが、未知の観測値に対する予測精度が必ずしも高くないことが指摘されている。未知の観測値に対する予測性能は汎化性能とよばれ、これらの手法を用いる場合には、データの当てはまりのよさと汎化性能をどのように評価するかが問題となる。

柔軟なモデリング手法による統計モデルは、データに対する当てはまりはよいが、汎化性能が劣るという問題は過学習とよばれ、この問題を回避するための手法が数多く提案されている。例えばニューラルネットワークにおいては、交差検証法や情報量規準を用いて中間層の数を抑制することで、複雑すぎるモデルを選択しにくくする手法や、正則化項を付与することで、データに対する当てはまりのよさをある程度犠牲にすることで汎化性能を向上させ、過学習を回避する手法がある。カーネル法も高度な非線形性を表現することが可能な方法である。未知の関数をカーネル関数のもつ再生性とよばれる性質を利用して表現する方法であり、カーネル法を用いた回帰分析や判別分析など、古典的多変量解析手法を拡張することで多くの研究が行われている。カーネル法においてもデータのもつ高度の非線形性を再現できる性能の反面、汎化性能が必ずしも高くないことが指摘されており、ニューラルネットワークと同様の考え方で過学習を回避する手法が提案されている。こうした中で、クラスタリングと情報量規準を併せて用いることで、高い汎化性能をもつモデリング手法が提案されてきた。この手法では、クラスタリングを用いてモデルの複雑さを抑制することで過学習の問題を回避している。

本研究では、サンプルサイズとカーネル法による関数近似の関係に着目して、過学習の問題を考察した。データが誤差なく観測できる場合には、未知の関数をカーネル関数を用いて完全に再現することができるための必要十分条件が知られている。この条件が満たされる場合には、過学習の問題が生じないことを数値実験により確認した。この結果に基づき、観測値に誤差が含まれる場合には、カーネル法における未知の関数の推定誤差が、サンプルサイズが大きくなっても減少しない場合があることを示した。これは、サンプルサイズが大きくなることで過学習が生じることの一つの原因と考えることができる。

さらに、この結果に基づいて、過学習を回避する手法を提案した。提案した手法は、カーネル法において、サンプル点の内積からなる行列を回帰分析における計画行列とみなすことにより、情報量規準を用いて最もよいモデルを選択するものである。推定すべき未知の関数がモデルに含まれていないことを考慮し、情報量規準として、TIC(Takeuchi Information Criterion) を用いた。この手法を用いた数値実験により、カーネル回帰分析では予測精度を小さくすることができることを示した。また、カーネル判別分析においては、予測誤差を減少させることができることを示した。この結果は、有効なサンプル点だけを用いることで、過学習を回避することが可能であることを示している。

この手法を機械学習のベンチマークテストとして広く用いられているデータベースの判別問題に適用し、観測値の誤差に対する仮定が成り立たない場合に、本提案手法がどの程度有効であるか検証を行った。

学位論文審査の要旨

主　査　　教　授　　今　井　英　幸

副　査　　教　授　　宮　腰　政　明

副　査　　教　授　　工　藤　峰　一

学位論文題名

情報量規準に基づくモデル選択を用いたカーネル多変量解析法に関する研究

近年、インターネット上の文書や大規模データベースなど、複雑な構造をもつ大量のデータを入手することが容易になり、このようなデータを解析するために様々な手法が提案されてきた。こうしたデータでは、質的変数、離散型変数、連続型変数など多種類の変数が混在した観測値として記録され、また、データ全体としては、高度の非線形性などの複雑な構造を有するものであることが多い。こうしたデータに内在する構造を明らかにし、将来発生する事象に対する予測や、予測に基づく意思決定を精度よく行うためには、適切な統計モデルを構築することが必要である。

統計モデルはある時点で入手しているデータによって推定される。統計モデルの推定においては、モデルの構築に用いられるデータをいかによく説明しているか、というデータに対するモデルの当てはまりのよさと、将来発生するであろう未知のデータに対する予測精度という二つの面から評価される。モデルの構造を複雑にすることにより、データに対するモデルの当てはまりは改良することができるが、複雑すぎる統計モデルによる予測精度は必ずしも高くないことが指摘されている。データに対するモデルの当てはまりを重視した場合に、予測精度が悪くなる現象は過学習の問題とよばれ、これを回避するために多くの研究が行われている。

本論文では、カーネル法を用いた多変量解析手法に関して、過学習が起きる理由を理論的に解析した上で、その結果に基づいて、真の分布との Kullback-Leibler 距離を最小にすることにより、予測精度の高いモデルを構築する手法を提案している。また、数値的にその有用性を確認している。

本論文は7章からなる。第1章では、本論文の導入部分として研究の目的および研究の学術的な意義を述べている。

第2章では、現在広く用いられている手法について、それらの概要を、データに対するモデルの当てはまりと、予測精度の観点から説明している。特に、多くの先行研究で有効性が確認されているサポートベクター回帰モデルと動径基底ネットワークモデルについて詳しく記述している。過学習を回避するための手段として、サポートベクター回帰モデルではマージンの設定、動径基底ネットワークモデルでは動径基底の個数の決定についても詳しく述べられている。

第3章では、本論文において提案する手法の基礎となる、非線形関数の再生核による近似の理論について説明している。データの観測過程で加法的雑音がある場合に、関数の推定において雑音が与える影響を再生核の理論を用いて解析し、その結果として、サンプルサイズと関数の近似精度に

トレードオフの関係が生じる可能性があることを示している。この結果は、サンプルサイズが大きい場合に、必ずしも予測精度が高くないという現象を説明する一つの根拠となるものである。

第4章は、第3章の解析に基づいてモデルを構築する際に必要となる情報量規準に関するものである。本章では、提案手法の説明で必要となるAIC(Akaike Information Criteria)と、比較手法の一つである動径基底ネットワークモデルで必要となるGIC(Generalized Information Criteria)について詳しく述べている。

第5章と第6章は、本論文の主要な成果である情報量規準を用いたカーネル回帰モデルの構成方法と、カーネル判別分析法に関するものである。カーネル回帰分析は、通常の回帰分析モデルにおける母数推定の問題と考えることができる。回帰分析モデルの母数推定における変数選択が、カーネル回帰モデルにおいてはサンプルの選択に相当することに着目し、第3章で述べたモデルの当てはまりのよさと予測精度の関係を利用して、情報量規準により最適なモデルを構築するアルゴリズムを示している。また、複数の極値がある関数と、単調な関数の二種類の関数の推定に対して本論文の提案手法を適応し、従来法による統計モデルよりも予測精度の高いモデルを構築することができることを示している。さらに、カーネル回帰モデルにおいて有効であったモデル選択の考え方が、正準判別分析にも適用できることを示し、情報量規準によるモデル選択により判別関数を構成することができることを示している。

第7章において、本論文の総括を行っている。本論文で提案された手法についてその意義を明らかにするとともに、これから解明することが必要な課題について述べている。

これを要するに、著者は、多変量の回帰あるいは判別において、カーネル法に代表される十分な複雑さを保有する統計モデルが陥る過学習という問題に対して、その原因を数理解析的に明らかにするとともに、情報量規準に基づくモデル選択を通して解析結果の実用面での有用性も示した。これは、データを扱う多くの分野において重要な示唆を与える成果であり、情報科学、特に、多変量データ解析の分野に貢献するところ大なるものがある。よって、著者は北海道大学博士(情報科学)の学位を授与される資格があるものと認める。