

学位論文題名

次元縮小における統計的感度分析法に関する研究

学位論文内容の要旨

近年、情報科学において解析対象とするデータの次元数が著しく増加している。次元数の増加に伴い、データの本質的な部分を抽出する次元縮小法に基づく手法の需要は増えており、その有用性は広く認知されている。

しかし、次元縮小法に基づく手法にはいくつかの問題点が残されている。その一つは外れ値の問題である。高次元データにおける少数個の外れ値により、解析結果が悪化することが多い。解を適切に表現する統計量が定義できる手法に対しては、その統計量に対する影響関数を導出し、外れ値の検出に利用することができる。しかし、統計量が明確に定義されない次元縮小法に基づく手法も数多く存在する。そこで、本研究では、このような場合でも感度分析を用いることができるよう新たに統計量を定義し、データに混在する外れ値を検出する方法を提案した。

また、入力されるデータの変化に伴う問題が存在する。例えば、外的基準を有する解析手法を考える際、学習対象である標本が逐次的に得られる場合において、得られる標本の特性がある時点から変わるようなときには、学習した規則に対して、その変化が適切に反映されなければならない。このような状況において、標本の特性が変わった時点、すなわち変化点の検出は重要課題である。本研究では、変化点の検出を行うため、統計的感度分析法に基づく逐次診断法を提案した。また、実際例を用いてその有効性を検証した。

さらに、次元縮小後に得られる低次元空間の解釈の問題がある。多くの場合、低次元空間を張る基底ベクトルに基づきデータの解釈が行われるが、解釈の付与が困難な場合も少なくない。本研究では空間を張る基底ベクトルの符号の出現頻度の感度に基づき、解釈が容易となる回転方法を導出した。

本論文は6つの章から構成される。

第1章では、本研究の背景、及び目的について説明した。

第2章では、本研究で扱う従来手法として、部分空間法と主成分分析法を説明した。部分空間法は、クラスごとにデータを次元縮小し、縮小された空間を張る基底ベクトルから計算した射影行列を用いて効率的に判別を行う手法である。主成分分析法は、分散が最大となる直交した少数の軸を導出し、次元数を減らす手法である。得られた部分空間の解釈として、主成分係数行列や主成分負荷量行列を用いて空間を解釈する方法を示した。さらに、本章では次元縮小に基づく手法における問題点

を解決するための方法である統計的感度分析法について述べた。特に標本が統計量に対して与える影響を評価する影響関数を中心に説明した。

第3章では、部分空間法で代表的な CLAFIC 法に対して、各クラスの射影行列の偏差に注目した統計量を定義した。これを用いて経験影響関数及び標本影響関数を導出し、影響的な標本を検出する方法を示した。さらに、単数の外れ値だけではなく、複数個の外れ値の影響を評価する方法についても考察した。実画像データの判別を例として、CLAFIC 法に対する感度分析の結果を示し、誤判別率の低下につながる外れ値を同定し、有効性を示した。

第4章では、部分空間法において、トレーニングデータが逐次的に入力されたときの、統計的感度分析に基づく変化点の検出法を提案した。具体的には、ベクトルで得られている統計量に対して標本が与える影響を、同様にベクトルで表わし、新たに入力された標本が与える変動を評価した。多変量正規分布を用いて発生させたシミュレーションデータを用い、提案手法を CLAFIC 法に適用した例を通じて、その有効性を示した。

第5章では、次元縮小に基づく多変量統計解析の代表的手法である主成分分析法において、その解空間を回転させ、低次元空間において解釈が容易となる軸の導出を検討した。このような着想に基づく先行研究には、主成分係数行列を回転対象とする研究がある。これに対して本研究では、主成分負荷量行列の回転を考え、各軸における係数値の分散が最大でかつ、係数値の符号が同符号となる軸を解釈が容易な軸とした。そこで、直交回転後に得られた軸のうち、絶対値の大きな係数値の、同一軸における異符号の出現頻度の感度について検討した。符号変化に関するシミュレーションを行い、主成分負荷量を回転における基準とする方が解釈が容易な軸を得やすい傾向があることを示した。

第6章では、まとめと今後の課題について述べた。

学位論文審査の要旨

主 査 教 授 水 田 正 弘
副 査 教 授 赤 間 清
副 査 教 授 高 井 昌 彰
副 査 准教授 南 弘 征

学 位 論 文 題 名

次元縮小における統計的感度分析法に関する研究

本論文では、高次元データに対する有効な解析手法である次元縮小およびその関連手法に対して統計的感度分析を適用した研究成果について論じている。次元縮小を利用して、大規模で複雑なデータから有益な情報を抽出するための重要な課題に、外れ値の検出、逐次入力される教師データの構造変化の検出、データが潜在的に有する情報を適切にあらわす軸の同定がある。これらの課題に対する従来の方法は、非線形な構造を有するデータや、構造の急激な変化を伴う時系列データに適さない場合がある。そこで、申請者は、各課題に対して感度分析法を利用することに着目して、理論的な考察とともに、多様なデータに対して有効な解析技法を提案した。

提案技法の適用例として、画像データの判別問題における判別率の改善が挙げられている。また、本技法は、情報ネットワークにおいて重要な課題である、環境の急激な変化、例えば、スパムメールの傾向の変化などを検出する方法としても適用可能であると判断できる。

本論文の構成および内容は以下の通りである。

第1章および第2章では、本研究の位置づけを述べるとともに、関連する従来の手法を総括している。主成分分析に対しては、解析の過程で用いられるいくつかの特徴的な行列を紹介するとともに、それぞれの解釈の方法について独自の視点から記述している。また、パターン認識において広く利用されている部分空間法について、統計学の立場から記述している。

第3章では、情報処理分野において広く利用されている部分空間法における感度分析を扱っている。部分空間法は、画像データの判別に有効であることがよく知られている。しかし、少数個の外れ値がその判別結果を悪化させることも同様によく知られている。そこで、感度分析に用いる統計量として、部分空間法の解に対する影響関数を導出し、これに依り外れ値を検出する方法を提案した。まず、単数の外れ値を検出する方法を提案し、その有効性を理論的考察およびシミュレーションにより評価した。次に、複数の個体が外れ値と考えられる場合に、部分空間法の結果に悪影響を与える状況を検討し、これらを検出する方法を提案した。複数の外れ値を検出する場合、組み合わせの数が極端に大きくなることから、実行可能性が問題となる。そこで、個々の外れ値が及ぼす影響の方向を考慮したクラスター分析法を利用することで、計算量を大幅に削減する方法も提案した。

第4章では、逐次的に個体の観測値が入力される状況を検討し、個体の状況が変化する時点の検

出を試みている。近年、個体が一度に観測されず、逐次的に得られる場合が多く報告されている。このような場合、特に、個体の状態が変化した時点(変化点)を検出することは、きわめて重要な問題であり、多くの先行研究が報告されている。直観的にも自然な方法として、前時点との差分の利用が考えられるが、変化点の検出能力が不十分な場合が多い。そこで、本研究では、感度分析を利用した変化点の検出方法を提案している。具体的には、逐次的に観測値が入力される場合における部分空間法への適用を想定し、変化点を容易に同定できるような手法を開発し、シミュレーションによりその有効性を評価している。

第5章では、第2章で言及した、主成分分析の過程で用いられる種々の行列に関する記述を踏まえ、主成分分析により得られた部分空間に対して、解釈が容易となる座標軸の回転方法について議論している。主成分分析はデータ解析における次元縮小法として基本的な手法であり、広く利用されている。類似した次元縮小法である因子分析では、解の回転について多くの議論があるが、主成分分析において、回転に関する研究は少ない。代表的な理由として、軸の適切な回転についての基準が明確ではないからである。そこで、本研究では、軸を表現する係数の符号に着目し、異符号となる係数の個数に着目し、その感度を考慮した、座標軸の回転方法を提案している。また、手法の有効性を、従来手法との比較により示している。

第6章では、本研究の総括と今後の課題について述べている。

これを要するに、著者は、解析が困難である高次元データに対する解析について、次元縮小および感度分析の立場から効果的な解析方法を提案し、その有効性を示しており、多次元データ解析を含む情報科学に対する貢献は大なるものがある。よって、博士(情報科学)を授与するに値すると認める。