氏名	いわ むら まさ かず 岩 村 雅 一
授与学位	博士（工学）
学位授与年月日	平成15年3月24日
学位授与の根拠法規	学位規則第4条第1項
研究科，専攻の名称	東北大学大学院工学研究科（博士課程）電気・通信工学専攻
学位論文題目	パターン認識における確率分布推定法に関する研究
指導教官	東北大学教授　阿曽 弘具
論文審査委員	主査　東北大学教授　阿曽 弘具　東北大学教授　阿部 健一 東北大学教授　牧野 正三　東北大学助教授　大町 真一郎

論文内容要旨

第1章 序論

確率論，数理統計学の基礎的研究によって得られた知見は，確率分布を仮定したモデルとして，工学など，幅広い分野で用いられている．これらの理論はその根本でサンプル数が十分多い状況を仮定しているため，少数のサンプルしか利用できない実際の状況にはそぐわない部分があり，その影響はパラメータの推定誤差や確率分布の推定値の偏りとして現れる．特にパターン認識では，学習用に多くのサンプルを集めることが非常に困難で，少数のサンプルから高い認識性能が得られるような技術が常に求められている．

本論文では，少数サンプルで推定した場合の推定誤差の原因を追究し，確率論，数理統計学の知見を活用して，推定誤差を補償する方法，推定誤差を軽減する方法，必然的な偏りを補正する方法を研究し，理論式に基づいた手法を開発した．

第3章，第4章では誤差を含む推定値から誤差の影響を少なくする手法，第5章では確率分布の偏りに対処する手法を提案する．

第2章 パラメトリックな確率分布とパラメータの推定

本章では，パラメトリックな確率分布の推定法を概観し，各種の固有値補正に基づく共分散行列の推定法とその根拠を紹介した．最初に統計的パターン認識において広く用いられているベイズ決定則について述べた後，ベイズ決定則で必要となる確率分布の推定法について述べた．本論文では主に正規分布を仮定して導かれる識別関数を扱うため，平均ベクトルと共分散行列の推定法については特に詳しく述べた．

確率分布の推定法は，パラメトリックモデルとノンパラメトリックモデルの2つに大きく分けられる．本論文ではパラメトリックモデルについて扱い，パラメトリックモデルのパラメータをいか

に高精度に推定するかに焦点を当てる．パラメータの推定法には，点推定の枠組みに属する不偏推定や最尤推定と，区間推定の枠組みに属するベイズ推定などがあり，パターン認識などの応用で広く用いられる．

共分散行列の推定に最尤推定法を用いる場合には，その固有値の誤差の傾向が広く知られ，統計学の立場からより良い共分散行列の推定値を得る方法が数多く提案されており，その中の代表的な手法について述べた．これらの手法の一部は，本論文で提案する手法の有効性を示す実験で比較対象とした．

ベイズ推定については，第5章でベイズ推定を用いたパターン認識の問題点を指摘する．そのため，ベイズ推定に関する基本的な事柄を本章に記した．

第3章 固有ベクトルの誤差を補償する真のマハラノビス距離の推定法

ベイズの定理から導かれる統計的識別関数は，パターンの確率分布が正しく与えられたときに，誤認識によって生じる損失の期待値を最小に留める最適性を持つ．しかし，ほとんどのパターン認識問題ではパターンの確率分布は未知であるため，パターンの確率分布として特定の分布を仮定し，分布パラメータを学習用のサンプルから推定することが多い．このとき，学習サンプル数が不足すると，パラメータの推定誤差のために認識性能が低下することが知られている．これは，特徴量の次元数が増えると，分布の推定に必要な学習サンプル数が指数関数的に増加することが原因で，次元の呪いとして知られている．

パターンの確率分布を多次元正規分布と仮定した識別関数の場合，多次元正規分布の分布パラメータである平均ベクトルと共分散行列を学習サンプルから推定する必要があるが，推定値である標本平均ベクトルと標本共分散行列はそれぞれ誤差を含む．これまで，平均ベクトルや標本共分散行列の固有値については推定誤差を補正する手法が提案されているが，固有ベクトルの推定誤差に関してはほとんど考慮されてこなかった．たとえ真の固有値を正しく推定して用いたとしても，固有ベクトルが推定誤差を含む場合にはマハラノビス距離が正しく計算されないことが考えられる．

本章では，共分散行列の固有ベクトルの推定誤差を考慮することで，学習サンプルが十分用意できない場合においても高精度な認識を行うことを目的とし，マハラノビス距離を正しく推定するために真の固有値を修正する手法を提案した．

提案手法を文字認識に適用した結果，マハラノビス距離が正しく推定されること，認識性能が改善されることを確認できた．マハラノビス距離の推定値を改善する効果はパラメータの推定に用いる学習サンプルが少ない程大きく，提案手法は学習サンプルを十分用意できない場合において特に有効である．多くの手法では認識実験や設計者の知識によって定めるハイパーパラメータを必要とするが，提案手法では一切不要である．

第4章 頑健な分布の推定法

パターン認識で用いられる主要な識別関数は共分散行列の逆行列を必要とするため，固有値展開によって共分散行列の固有値と固有ベクトルに分解し，固有値と固有ベクトルから逆行列を構成することが多い．ところが，学習用のサンプルから求めた標本共分散行列が推定誤差を含む場合には標本共分散行列の固有値と固有ベクトルが誤差を含み，大きな固有値がより大きく，小さな固有値がより小さく偏るため，認識性能が低下することが知られている．

このような問題に対しては，第3章で述べた方法も一つの解決策であるが，第4章では第3章の方法とは全く異なるアプローチで解決する方法を示した．すなわち，最初に固有値の偏りがどこで生じるかを実験的に調査し，固有値の偏りは固有ベクトルの誤差（真の固有ベクトルと標本固有ベ

クトルのずれ）が原因で，固有値展開で生じると考えられることを確認した．そして，従来のように固有値展開後に固有値を補正するのではなく，固有値展開の際に固有値が偏りにくくなるよう共分散行列を変換する手法を提案した．

具体的には，特徴量の次元数が大きくなると固有ベクトルの推定誤差が大きくなり，このことによって固有値が偏ると考えられることから，標本共分散行列を調べて，分布が等方性に近いと考えられる（超）平面を固有値展開前に縮退させる．縮退した共分散行列に対して固有値展開を行い，得られる縮退した固有値と固有ベクトルを元の次元数に展開する．以上の手順を経ることにより，固有ベクトルの誤差に頑健に分布を推定できると考えられる．

認識実験により，提案手法が標本共分散行列を用いた場合に比べて真の分布をより正しく推定し，認識性能を改善することを確認した．

提案手法の導出過程で，分布を特定するような仮定を一切用いていない．そのため，幅広い応用が期待できる．一方，提案手法の 2 つのパラメータは真の分布の違いを吸収しているものと考えられるが，この定め方を導くのは今後の課題である．

第5章 ベイズ推定における尤度の偏りの補正法

学習用の標本から分布パラメータを推定する際，未知パラメータを定数として推定する最尤推定法に対し，未知パラメータを確率変数と考えて，その分布を推定し，パラメータの分布をもとに次のサンプルの予測分布を求めるベイズ推定がある．ベイズ推定は，標本が与えられる前に持っている真のパラメータに関する知識を「事前分布」という形で表現し，標本から得られた情報と統合した「事後分布」を求め，次のサンプルの分布を「予測分布」として推定する．

共分散行列が未知のときに予測分布を求める方法としては，真の固有ベクトルが特定の行列をパラメータとする逆 Wishart 分布に従うとする Keehn の方法や，パラメータが一様に分布すると仮定した無情報事前分布を用いる Geisser の方法などがある．サンプル数が少ない場合，ベイズ推定を用いたほうが最尤推定よりも優れた学習が行えるとされるが，Keehn の方法，Geisser の方法ともに学習標本の大きさがクラス間で異なる場合のパターン認識では有効でないとされる．様々な予測分布に対して適用可能な一般的な手法はまだ提案されていない．

ベイズ推定を行うと，サンプル数が大きい場合には，未知入力ベクトルの予測分布が漸近的に多次元正規分布に一致することが知られている．一方，サンプル数が十分大きくない場合については，予測分布の期待値が多次元正規分布と比べて小さく偏ると考えられる．この偏りはサンプル数によって異なり，一様でない．そのため，各クラスのサンプル数が等しい場合には，偏りに（あまり）差が生じないため，認識性能に影響を及ぼさないが，サンプル数が異なる場合には偏りに差が生じて，認識性能が低下するものと考えられる．

学習サンプル数がクラス間で異なる場合に認識性能が低下する原因が，サンプル数を n としたとき，$n \to \infty$ の近似が成り立たないときに生じる予測分布の偏りの差にあるならば，予測分布の偏りによって生じる尤度の偏りを補正することで認識性能が改善するはずである．

本章では，Geisser の方法について，サンプル数に応じた尤度の理論式（厳密解）を導出し，サンプル数が小さい場合の偏りの程度を定量的に評価し，誤認識が起こりやすい状況についての根拠を明らかにするとともに，解決法を与えた．まず，サンプル数が少ない場合に尤度の期待値が偏ることをサンプルから求めた平均値を用いて確認した．次に，尤度の理論式（厳密解）を導出し，サンプル数がクラス間で異なる場合には，導いた理論式で尤度を補正すれば認識精度が向上することを実験によって示した．これは，ベイズ推定を用いたパターン認識では予測分布の偏りが認識性能を悪化させていることを検証するものである．さらに，本章で導く理論式を用いて，過去の実験的な研究で指摘されている「分散の小さなクラスの学習サンプルが少ないほど認識率が低下する」という経験的に得られた知見に理論的根拠を与えた．

なお，本章で示す尤度の理論式の導出法は，別の事前分布を仮定した場合にも容易に適用できる高い一般性を持つ．

第6章 結論

本論文では，サンプル数が少ない場合に生じる分布パラメータの推定値の誤差の原因を追究し，誤差や偏りの評価を理論的に行うとともに，より正確な推定値を得るための補正法を提案し，検証実験，認識実験によってその有効性を実証した．

本論文で行った研究は応用面を殊更に意識したものではなく，むしろ基本となる原理の考察に重きを置いている．したがって，本論文の主な成果は，パターン認識の分野に限らず，他の工学的な応用，また工学以外への応用がそのままの形で，もしくは，若干の修正によって可能であると考えられる．

論文審査結果の要旨

パターン認識では，パターンが確率的に発生するとモデル化して，確率論や数理統計学の知見を利用して識別を実現している．確率分布を仮定したモデルはパターン認識だけでなく他の幅広い分野で有用である．確率モデルを仮定する場合，モデルパラメータはサンプルデータから推定することになるが，サンプル以外のデータに対しても有効な汎用モデルにするには十分な数のサンプルが必要である．しかし，パターン認識など工学的応用ではデータの次元数に応じた十分な数のサンプルを得ることが難しく，少数サンプルで推定することになり，推定パラメータは多くの誤差を含み，モデルの性能が発揮できない．著者は，少数サンプルで推定した場合の推定誤差の原因を追究し，確率論，数理統計学の知見を活用して，推定誤差を補償する方法，推定誤差を軽減する方法，必然的な偏りを補正する方法を研究し，理論式に基づいた手法を開発した．本論文はその成果をとりまとめたもので，全編6章よりなる．

第1章は序論で，研究の背景と目的を述べている．

第2章では，パラメトリックな確率分布の推定法を概観し，各種の固有値補正に基づく共分散行列の推定法とその根拠を紹介している．

第3章では，共分散行列を用いる識別関数であるマハラノビス距離の計算の際に標本共分散行列の固有ベクトルが関わる誤差に着目し，その誤差が固有値をゆがめるため真の共分散行列とは異なってしまうこと，その誤差の分布が真の固有ベクトル自体には依存しないことを導き，そのゆがみを補償してマハラノビス距離のより正確な値を求める方法を提案している．固有ベクトルの誤差の分布の理論式を導き，ゆがみ補償を求める具体的方法を示した．少ないサンプルでもマハラノビス距離が正確に求まることを実験的に確認するとともに，認識実験により高精度な認識ができることを示し，有効性を実証した．これは有用な成果である．

第4章では，誤差が入っている標本共分散行列から厳密な数値計算で固有値展開するため固有値の誤差が大きく現われてしまうと考え，標本共分散行列自体の数値的ゆれを考慮してあらかじめ誤差を軽減しておく方法を追究し，標本共分散行列を縮退させて固有値展開し元に戻すという手法で固有値の新しい推定法を導いている．これは固有値が等しい場合固有ベクトルが一意に決まらないという性質を利用したもので，多次元の列ベクトルを2個ずつ組にして縮退可能かどうか調べることに基礎をおいている．この推定法の確からしさを実験的に確認するとともに認識実験により従来手法と比較して高精度な認識が達成できることを示し，提案手法の有効性を実証した．

第5章では，ベイズ推定における予測分布の期待値の偏りに関して，サンプル数に応じたその偏りの程度を評価する理論式を導いている．認識対象クラス間でサンプル数が異なるとき，予測分布の偏りによる尤度の偏りをこの理論値で補正することにより認識の際に公平な比較が可能となることに着目して，この補正を加えた認識手法を提案し，認識実験によりその有効性を実証している．今後の展開の基礎となる興味深い成果である．

第6章は結論である．

以上要するに本論文は，サンプル数が少ない場合に生じる分布パラメータの推定値の誤差の原因を追究し，誤差や偏りの評価を理論的に行うとともに，より正確な推定値を得るための補正法を提案し，検証実験，認識実験によってその有効性を実証したもので，情報通信工学，パターン認識理論の発展に寄与するところが少なくない．

よって，本論文は博士（工学）の学位論文として合格と認める．