別紙4

報告番号	※乙　　第4870号

主　論　文　の　要　旨

論文題目 有限学習標本に基づく統計的パターン分類器の設計と評価

氏　名　竹下　鉄夫

論　文　内　容　の　要　旨

本論文では，パターンの母集団の確率分布として多次元正規分布を仮定し，有限学習標本に基づく統計的パターン分類器の設計と評価に関して論ずる．パターン認識の最終性能は分類器の平均誤り率によって評価すべきである．パターン集合の確率分布に関する情報が完全に既知であれば，分類器の平均誤り率を最小にすることが可能である．平均誤り率最小を達成する決定をベイズ決定と呼び，このとき達成される最小平均誤り率をベイズエラーもしくはベイズ誤り率と呼ぶ．また，最小平均誤り率を達成する分類器をベイズ分類器という．

パターンの確率分布に多次元正規分布を仮定するとしても，実際のパターン認識においてはパターン分類器の設計・評価に使用できる標本数は有限であり，パターン集合の確率分布に関する構造も有限個の標本を用いて推定するしかない．このため，有限個の標本を用いて設計された分類器の識別性能はベイズ分類器よりも劣るのが普通である．また，有限個の標本から設計された分類器の識別性能は学習標本の選び方によって確率的に変動するため，その振る舞いに関しては解明されていないことが多い．

パターン分類器の識別性能を評価する方法としては，実際にテスト標本を用いて識別実験を行う方法と，テスト標本に母集団のパラメータを用いて数値積分により評価する方法がある．前者には，再代入法(Resubstitution method)，分割法(Holdout method)，一つ取って置き法(Leave-one-out method)などが広く知られている．最近では，Efronがブートストラップ法と呼ぶ評価方法を提案し，識別性能の評価に関する分散を小さくすることに成功している．これらの性能評価法は母集団の分布として特別な分布を仮定せず，分類器の設計と評価を同一セットの標本を用いて行うことができる．しかしながら，この方法はベイズ誤り率が小さいところでは，膨大なテスト標本を用意する必要が生じる上，テスト標本の有限性に起因する誤り率の推定誤差要因を除くことができないため，学習標本の有限性のみに起因する認識システムの識別性能の評価が難しい．

一方，あらかじめ母集団のパラメータを設定しておけば，母集団の分布に正規分布を仮定すると，パターン分類器の性能評価に数値積分を導入することにより正確な識別性能の評価ができる．しかしながら，特徴パターンの次元数が増加するとともに計算量の増加を招き，さらに，識別性能を示す誤り率自身が学習標本の選び方に従って確率的に変動するため，誤り率を確率変数として考えなければならなくなり，計算量の増大は甚だしいこととなる．Fukunagaらは数値積分により誤り率を求める方法を示しているが，多次元正規分布において共分散行列が異なる場合のベイズ誤り率を求めているに過ぎない．また，FukunagaやRaudysらは学習標本の有限性に起因するベイズ誤り率からの識別性能の低下の程度を示す近似式の導出に関連して，共分散行列が等しい場合について数値積分により誤り率を求めている．共分散行列が異なる場合も含み，より一般的に統計的パターン分類器の性能評価を数値積分により包括的かつ系統的に実施した例は国の内外を問わず見当たらない．

パターンの確率分布の未知パラメータを標本パターンから推定する問題は統計的推定の一問題である．しかし，推定の良否が認識システムの識別性能に及ぼす影響については，前述したように識別性能を示す誤り率自身が確率的に変動するため，その評価は困難な問題である．平均値ベクトルの推定量として標本の単純平均がよく用いられているがSteinは二乗誤差最小という基準のもとでは最良でない(admissibleにならない)ことを示し，統計学の分野でセンセーションを巻き起こした．したがって，母集団の分布に正規分布を仮定して設計する線形識別関数や２次識別関数についても，平均値ベクトルや共分散行列の推定値として標本平均値ベクトル，標本共分散行列を用いることが一般的に行われているがこれが最良の方法であるという保証はない．

学習標本が少ないとき，特に共分散行列の推定精度が認識システムの識別性能に悪影響を与えていることは従来から多くの研究者によって指摘されてきた．しかしながら，それらの研究はいずれも共分散行列それ自体の推定精度を検討するというよりも，認識システムの識別性能との関連で共分散行列の推定精度についての評価をしている．最終的には，推定の良否は認識システムの識別性能により評価することは必要であるが，一般的には前述したように膨大な計算を必要とする．従って，従来共分散行列の推定精度を上げるための試みは個別的な認識システムの識別性能の実験的な評価にとどまっており，理論的もしくは系統的に研究された例は見当たらない．

本研究においては，２次識別関数の主要な部分を成すマハラノビス距離に着目し，これの推定精度の向上が認識システムの識別性能の向上につながると予想した．すなわち，共分散行列の推定精度をマハラノビス距離と関連付けることにより，２クラス以上の認識システムを構成することなく，マハラノビス距離の推定精度を評価することにより，共分散行列の推定の良否を評価するようにした点に本研究の特徴の一つがある．

統計的パターン分類器を設計する際に必要となる主要なパラメータとしては，平均値ベクトル，共分散行列がある．これらのパラメータ自身の推定に関しては統計学を通して研究がなされているものの，これらの推定パラメータを用いて構成されるマハラノビス距離や統計的パターン分類器の誤り確率という諸統計量は，その振る舞いが複雑であり，十分に調べられているとは言い難い．

本論文は，標本平均値ベクトルや標本共分散行列などの推定精度や，それらを用いて構成される上記諸統計量について，主としてコンピュータによりモンテカルロ・シミュレーションを実施し，それに基づいて統計的パターン分類器の性能を評価したものである．以下に，各章の概要を述べる．

第１章では，研究の意義，多次元正規分布と識別関数の基礎，本研究に関連する国内外の研究について述べる．

第２章では，２次識別関数の主要な要素を占めるマハラノビス距離の推定誤差について論ずる．Steinが示した，“３次元以上のパターンでは，単純標本平均が平均二乗誤差最小という基準のもとでは最良の推定を与えない”という現象について確認したのち，さらに，Steinの提唱した平均値ベクトルに関する補正の効果も確認する．

さらに，Steinの補正は平均値ベクトルのみではなく，標本共分散行列の推定精度の向上をももたらし，その推定値を用いて構成されるマハラノビス距離についても，母集団の平均値ベクトルの絶対値が小さいほどSteinの補正の効果が大きいことを示す．

第３章では，マハラノビス距離の推定量の確率分布を理論的に明らかにし，有限学習標本を用いて推定した標本平均値ベクトルと標本共分散行列がマハラノビス距離の推定量に与える影響の程度を定量化する．マハラノビス距離の推定量の確率分布を前記二つの推定パラメータの成分に分解し，各々の成分毎の平均値と分散に関する理論式を導出する．その結果，標本共分散行列の推定誤差がマハラノビス距離の推定に大

きな影響を与えていることを示す.

第４章では，有限学習標本から母集団のパラメータを推定し，その推定パラメータを用いて，線形識別関数と２次識別関数を構成し識別関数の性能を評価する．特徴パターンの次元数，学習標本数，および平均認識率の間の関係を明らかにし，RaudysおよびFukunagaらの提案する近似認識率と比較し，それらの評価式の適用限界を明らかにする．さらに，認識率の低下分に関する評価を行い，ベイズ誤り率が高く，特徴ベクトルの次元数が高くなるに従い，認識率の低下分の分布はχ^2分布に近くなることを示す.

第５章では，２次識別関数の値を実質的に支配するマハラノビス距離の推定値について，有限学習標本に起因して母集団のパラメータが正確に推定できないことでマハラノビス距離の推定誤差が増大する主たる要因は標本共分散行列の推定誤差にあることを第３章で明らかにしたので，標本共分散行列の推定誤差に起因して，識別関数の性能が低下すること防止するため Friedmanが提案したRegularized Discriminant Analysis（以下ＲＤＡ法と呼ぶ）により，有限学習標本に基づくパターン分類器の識別性能の低下を評価する．その結果，２次識別関数が学習標本数に敏感なパターン分類器であることを確認する．またＲＤＡ法の有効性を確認したのち，ベイズ認識率の高い状況ではＲＤＡ法の一方のパラメータがあまり有効に作用しないことを示す．ＲＤＡ法が有している二つのパラメータの自由度の内，より有効なパラメータを指摘し，より優れた統計的パターン分類器を設計できる見通しを得る．その後，ＲＤＡ法の実用的なパターン認識システムへの効果を調べるため，手書き数字認識実験を行う.

第６章では，本論文を総括し，今後の展望と課題について述べる.