情報電子工学科　論文発表

題名	標本マハラノビス距離の識別性能改善に関する研究
掲載雑誌	帝京大学大学院理工学研究科へ提出された博士論文
著者	小林靖之
概　要	本研究は、機械学習における識別器として用いられる標本マハラノビス距離(SMD)の識別性能改善を目的として、SMDのもつ問題点①「学習サンプル数nが不十分であるために計算したSMDが事前に予想される確率分布に従わない。」と問題点②「変数個数である次元数pが増えると一部の変数に生じた異常が検知できない。」に関する検討と対策の提案を行ったものである。 目次は以下のとおりである。本研究の背景と目的を述べた第1章に続き、SMDの問題点①はnがpよりも十分に大きくないために起こるので、第2章から第6章までにおいて、nの範囲に応じて対策を提案した。またSMDの問題点②はSMDを部分系に分割したに正確な確率分布モデルがないために起こるので、第7章と第8章において対策を提案した。最後に第9章において結論と今後の展望を述べた。 第2章においては、打切り誤差を考慮したQ統計量の主成分項の従う確率分布の近似モデルを示した。n≤pにおいては正確多重共線性によりSMDが定義できないから、対応する母固有値λが0である主成分項を抽出してQ統計量を新たに定義する必要がある。そこでQ統計量を構成する主成分項の近似モデルを考察し、Q統計量にすべき主成分項の標本固有値lの閾値の決定や、主成分項について統計的検定できるようになった。 第3章においては、標本共分散行列の条件数と浮動小数点実数型変数の打切り誤差との関係を示した。n>pにおいては起こりうる準多重共線性から生じる異常に小さいlがSMDの計算を不安定化するから、異常に小さいlの値を推定する必要がある。そこで、計算機における数値誤差や分析対象データベクトルの精度との関係から、計算が安定になる場合のSMDにおけるlの条件を明らかにした結果、左記条件を満たさないlに当たるSMDの主成分項を除外できるようになった。 第4章においては、学習サンプル特有の標本マハラノビス距離に関する確率分布の偏りを示した。n>pにおいては準多重共線性が起こらない場合においてもSMDの過学習現象が影響を与える。そこで、SMDにおける過学習現象を抑制に必要なnの下限を見積もった結果、この下限よりも多いnを事前に用意すればSMDの過学習現象が抑制可能になった。 第5章においては、標本マハラノビス距離の主成分に対するデルタ法による改良を示し、続く第6章においては、標本マハラノビス距離の主成分項に対する母固有値を用いない改良法の検討を示した。nが有限値である限り、完全には過学習現象を避けられないため、nが無限大に当たる母マハラノビス距離(PMD)への補正をSMDに施す手法の改良を提案した。まず、第5章でSMDの主成分項に対し統計学のデルタ法によりλを含む補正法を提案した。さらに第6章で正確な推定が困難であるλを推定せずにStein推定法を用いる補正法を提案した結果、SMDの持つnへの依存性が緩和されPMDに性質が近づき、過学習現象を緩和可能になった。 第7章においては、標本マハラノビス距離の主成分項の球状化を用いた変数の影響度評価法を示し、続く第8章においては、標本マハラノビス距離の主成分項とその部分和の従う確率分布の近似モデルを示した。SMDを構成するlの主成分項のモデルを検討した結果、第7章で学習サンプルのSを用いて球状化した主成分変数が従うt分布を用いて、観測対象ベクトルの各変数を球状化する異常度計算法を提案した。さらに第8章でt分布によるモデルを精密化し、球状化した主成分変数の2乗についてχ^2分布を用いてさらに正確な近似モデルを提案した。この結果、球状化したベクトルの各変数による検定計算は高速かつ変数間の相関も考慮でき、SMDの主成分項の確率分布に関するさらに正確な近似モデルから、元の変数に関する影響度をより正確に評価できるようになった。 以上のとおりSMDのもつ問題点を解決する提案がなされた。SMDのもつ問題点がすべて解消された訳ではないが、既存研究には無い成果が得られた。SMDは基本的な識別器であるから、これら提案手法は、他の高度な識別器のもつ問題点の解決にも適用できる可能性がある。