博士（工　学）　天　元　　宏

学 位 論 文 題 名

MDL 基準を用いた識別規則の構成方法に関する研究

学位論文内容の要旨

近年、計算機の性能は飛躍的に向上し、従来予想し得なかった程の高い処理性能を持つに至った。しかし、最新の計算機においても、その機能は人間の持つ高度な情報処理能力には遠く及ばない。人間の持つ情報処理能力のうち、最も特徴的なものの一つにパターン認識がある。これを計算機上で実現しようとする試みが計算機支援のパターン認識の研究であり、計算機科学の黎明期から様々な努力が続けられてきた。しかし、従来のパターン認識技術では現実のデータに対して精度と処理速度の両面において十分であるとは言い難く、そのため、近年、パターン認識技法を見直す気運が高まってきている。

パターン認識において、識別規則の構成は最も重要な課題の一つである。そのため、これまで様々な識別規則が提案され、改良が続けられてきた。それらはパラメトリック識別規則とノンパラメトリック識別規則の二つに大別できる。パラメトリック識別規則では、サンプルの分布に統計モデルを仮定し、そのモデルのパラメータを訓練サンプル集合に基づいて統計的に推定することで、ベイズ定理を通して最終的な識別規則を得る。一方、ノンパラメトリック識別規則では、サンプルの分布に特定の統計モデルを仮定せず、訓練サンプル集合に基づいて、分布よりもむしろ識別境界を推定する。

これまでの識別規則は基となる分布が単純なデータに対しては十分な識別性能を発揮するものの、実際のデータに対しては不十分なものが多く、その原因は次のように考えられる。(1) 特徴空間における各クラスの分布は一般に複雑な形状をしており、互いに重なり合う部分も存在する。そのため、正規分布などの単純なモデルを採用した場合、真の分布からかけ離れる傾向が強い。(2) 訓練サンプルの過度な信頼から生じる性能劣化も無視できない。外れ値あるいは境界付近のサンプルの僅かな変動はノンパラメトリック識別規則には大きい影響を与える。(3) 利用できる訓練サンプルは少数に限られる。そのため、識別規則を訓練サンプル集合に適合させ過ぎると未知サンプルに対する識別性能は逆に低下する (汎化の問題)。

これらの問題点を解決するためには、識別規則が (1) サンプルの複雑な分布に十分対応できる柔軟な表現力、(2) 訓練サンプルの局所的な性質だけでなく、ある程度大局的な性質を反映できる能力、さらに、(3) 訓練サンプル集合への過適合を回避する能力を総合的に持つ必要がある。そこで、本研究では、訓練サンプル集合に対する誤差とその識別規則自身の複雑さの両方を同時に評価する方法論の一つである MDL(最小記述長) 基準を用いて、適切な複雑さを持ち、十分な表現能力を持つ識別規則の構成方法を考える。本研究では、特に、ノンパラメトリック識別規則である区分的線形識別規則および、パラメトリック識別規則とノンパラメトリック識別規則の中間的な性質を持つ識別規則である混合モデルに基づく識別規則の二つに関して検討を行った。

本研究では、区分的線形識別規則と混合モデルに基づく識別規則それぞれに対し、MDL 基準を用いた新しい構成方法を提案し、その有効性および特性、問題点を明らかにすることを目的とする。本論文は五つの章から成り、その概要は以下の通りである。

第 1 章は本研究の序論である。パターン認識の研究を概観することから始め、パターン認識にお

ける識別規則の重要性を論じる。その後、これまでに提案された識別規則の問題点を指摘し、それらの問題点を解決するための方針を検討する。続いて、本研究の目的を述べ、本研究で提案する新しい識別規則の構成方法の概略を述べる。

第2章では、主に従来の代表的な識別規則の再検討を行う。まず、準備として、数学的な記法の定義を行う。その後、従来の代表的な識別規則を概観し、それらの特性や問題点を考察することで、大多数の従来手法の抱える共通の問題点を明らかにする。続いて、本研究で扱う二種類の識別規則に関して、これらの手法に着目した理由および提案法による改善点に関して論ずる。

第3章では、区分的線形識別規則の新しい構成方法を提案する。区分的線形識別規則は真の識別境界を複数の超平面で近似するノンパラメトリックな識別規則である。最初に、この方法論の概略とその特性を論じる。続いて、これまでに提案された様々な構成方法を述べ、特に最も実用的とされる Park and Sklansky の構成方法を紹介する。さらに、この方法では訓練サンプル集合に対する識別性能を制御できず、高い汎化能力を持たせられないことを指摘し、訓練サンプル集合に対する誤差を制御する新しい構成方法を提案する。従来法と提案法において共通する基本構成アルゴリズムを述べた後、提案法における具体的な構成アルゴリズムを説明する。ここで、誤差の制御パラメータの決定に MDL 基準を用いる。人工データおよび実データに対する実験により、提案法の有効性を示すとともに、その特性、問題点を明らかにする。

第4章では、混合モデルに基づく識別規則の新しい構成方法を提案する。この識別規則はクラス条件付き確率密度関数の近似に複数の正規分布の混合モデルを用いる、パラメトリック識別規則とノンパラメトリック識別規則の中間的な性質を持つ方法である。最初に、この識別規則は混合数(混合する成分分布の数)を増すことで、単純なパラメトリック識別規則と比較して複雑な分布をより柔軟に表現できること、また、適切な混合数を選択することで、従来の識別規則の持つ問題点を解決できる可能性があることを指摘する。その後、提案法の理念を述べ、従来法との考え方の違いを論じる。従来は各クラスに与える混合数をクラス毎に独立して、尤度を評価する MDL 基準を用いて決定していたのに対し、提案法では各クラスに与える混合数の最適な組合せを、識別性能を評価する MDL 基準を用いて決定することを述べる。さらに、従来法と提案法における混合数の選択方法を具体的に述べる。最後に、人工データおよび実データに対する実験により、提案法の有効性および特性、問題点を明らかにする。

第5章は本研究の結論である。総括として、第3章、第4章で得た結論を基に本研究の成果と問題点をまとめる。本研究で提案した識別規則の構成方法が、識別規則自身の適切な複雑さを選択する機構を持つことにより、従来の識別規則の持つ問題点を解決し、高い識別性能を発揮できたことを述べる。また、その特性に関して明らかにしたことを述べる。さらに、今後の課題をまとめる。

学位論文審査の要旨

主　査　教　授　新　保　　勝
副　査　教　授　伊　達　　惇
副　査　教　授　宮　腰　政　明
副　査　教　授　佐　藤　義　治

学位論文題名

MDL基準を用いた識別規則の構成方法に関する研究

パターン認識における識別規則の構成は基礎となる最も重要な課題である。しかし、その構成方法はこれまで各種の手法が提案されているものの、現実的なパターン認識問題に対して十分な識別性能を発揮できていないのが現状である。

本論文は、実用的な識別規則は問題に固有な複雑さを推定する必要があることを指摘するとともに、識別規則自身の複雑さを制御する機構を備えた新しい構成方法を提案し、その有効性および適用限界を明らかにしたものであり、主要な成果は次の点に要約される。

(1) パラメトリック識別規則とノンパラメトリック識別規則の代表的な手法について複雑さの面で検討し、従来の識別規則の抱える問題点がサンプルの複雑な分布形状に対する表現能力の不足、および個々の訓練サンプルに対する過度な信頼、少数の訓練サンプルに対する過度な適合にあることを指摘した。その上で、識別規則の複雑さをMDL(最小記述長) 基準を用いて制御することにより、これらの問題点を総合的に解決する手法を提案した。

(2) 従来の区分的線形識別規則は複雑さの制御に必要な訓練サンプル集合に対する識別性能を任意に指定できないことを指摘し、訓練サンプル集合に対する識別誤差を一定の閾値以下に抑える新しい構成方法を提案した。また、MDL基準を用いてその閾値を汎化能力の高い値に設定する方法を示した。

(3) 従来の混合モデルに基づく識別規則は混合する基本成分数の選択で、本来の目的である識別を考慮していないことを指摘し、訓練サンプル集合に対する識別性能と混合モデルの複雑さの両方を識別指向のMDL基準で評価することにより、識別に適した成分数を選択する新しい方法を提案した。

(4) 上記二つの新しい識別規則の構成方法を種々の人工データおよび実データに対して適用の上、その有効性を確認し、適用限界を明らかにした。

これを要するに、著者は従来の識別規則の構成方法が抱える問題点を明らかにし、識別規則の複雑さに着目した新しい構成方法を提案したものであり、パターン情報処理工学の発展に寄与するところ大である。

よって著者は、北海道大学博士 (工学) の学位を授与される資格があるものと認める。