博 士（情報科学） 青 木 和 昭

学 位 論 文 題 名

Multi-Class Classification with Class-Dependent Feature Subsets and Class-Dependent Classifiers

（クラスに依存した特徴集合および
クラスに依存した識別子を用いた多クラス識別に関する研究）

学位論文内容の要旨

一般物体認識や大規模画像認識など、近年のパターン認識においては取り扱うクラス数は日々増大している。しかし従来研究の多くは理論構築と分析のしやすさからクラス数を二としており、多クラス問題へは二クラス識別子を組み合わせることで対処している。二クラス識別子を多クラス問題に適用する一般的なアプローチとして、c クラス問題の場合、あるクラスとそれ以外のクラスを識別する二クラス識別子を c 個構成し結果を統合する“one-against-other”型と、全てのクラス対を識別する二クラス識別子を $c(c-1)/2$ 個構成し結果を統合する“one-against-one”型がある。しかし、直線上に並んだ三クラスを識別する場合、中央のクラスがマスクされてしまうなど、単純に one-against-other を適用しただけではうまくいかない場合がある。つまり、多クラス問題を二クラス識別子を用いて解く方法は便宜的であり、必ずしも多クラス問題を適切に扱っているとは言えない。本質的に多クラス問題を扱うことがパターン認識において重要である。

本研究ではその一般的な方法を論じている。本研究では、多くのクラスを各内部ノードで二つの素な部分問題に分ける「クラス決定木」を方法の中心に据える。また、部分問題毎に識別に有効な特徴を選択することを提案している。従来は全クラスに対して共通の特徴集合を選択していた。しかし、識別に有効な特徴は部分問題毎に異なる。本研究ではこのような特徴集合を「クラスに依存した特徴集合」と呼び、その有効性を解析している。同様に、部分問題毎に識別子を選択する「クラスに依存した識別子」に関しても同様の解析を行っている。

本研究の主な内容を以下に要約する。

・クラス決定木を用いた汎用的に適用可能な多クラス識別手法を提案した。
・クラスに依存した特徴集合およびクラスに依存した識別子の有効性を実験的に示した。
・問題の特徴や性質を可視化する目的での、クラス決定木の有効性を示した。

本論文の構成は以下の通りである。

第 1 章では、本研究の背景として、パターン認識における多クラス問題を紹介するとともに特徴選択について述べている。

第 2 章では、多クラス問題を解くための従来研究、特に決定木を含む階層的識別に関するこれまでの研究を概説している。特徴選択を階層的識別に適用した研究事例についても触れている。さらに、現在最も有力視されている二クラス識別子であるサポートベクターマシンを多クラスに拡張するためのこれまでの方法論を整理している。

第 3 章は中心となる章であり、第 1 章で示した問題意識の下、多クラス問題を部分問題の集合として扱う一般的方式を議論している。本研究では「クラス決定木」を用いて部分問題を階層的に表現することを提案している。従来の決定木は一つのクラス中に複数のクラスタを見い出す方法であり、結果として一つのクラスが複数の葉ノードに多く散らばった表現となっている。この方式はクラス数が限られている場合には比較的有効であるものの、クラス数が百を超えるような場合には解釈が困難である。そこで、本研究ではクラスタではなくクラスを単位としたマクロ的な扱いを重視して「クラス決定木」を提案している。下位のクラス数が限られた部分問題においては、通常の決定木を使うことができる。クラス決定木は以下の特徴を持つ:

・クラスの全体集合を根に持ち、根から葉へと辿ることでクラスを決定する。
・内部ノードは一つの部分識別問題を表し、クラスの集合を二つの素な部分集合に分ける。
・クラス数分の葉ノードを持つ。
・各内部ノードでは任意の特徴集合と識別子を用いることができる。

第 3 章ではさらに、クラス決定木の得失を明らかにしている。第一の利点は、対象とする識別問題の難しさが視覚化されることである。各ノードでの分離度 (識別率) を調べることで「どの部分問題がどの程度難しいか」を一覧できる効果は大きい。このことを利用して、識別が困難である部分問題に対して適切と思われる特徴を新たに追加することなどを検討することが可能となる。全クラスを一括して考察した場合には難しいこのような検討が、考察すべきクラス集合が限定されることで比較的容易になる。また、マスキング問題を階層的に問題を解くことで回避できる点も利点として挙げられる。識別性能という点では、決定木は本質的には一括で識別を行う方法に比べて劣るおそれがある。これは、上位ノードでの識別誤りが下位へ行くにつれ累積するためである。しかし、実際には、ノードでの特徴選択および識別子選択を行うことで性能を向上させられる場合が殆どであることを実験的に示した。加えて、本研究では二種類のクラス決定木の構成法を示しており、これらの構成法の有効性を従来の同様の試みとの比較において定性的に示している。

第 4 章では、実データに対する計算機実験の結果を述べている。実験の結果、クラスに依存した特徴集合およびクラスに依存した識別子をクラス決定木に適用することで、8 個の実データセットに対して識別性能向上を確認している。一括して全クラスを識別する手法と比較して、ボトムアップ構成を行った場合に 5.65%、トップダウン構成を行った場合に 6.80%、識別率を改善した。また、手書き文字認識を例として、問題を決定木として可視化することでの有効性を論証している。

第 5 章では結論と今後の課題について述べている。

本研究の貢献は、クラス決定木表現が多クラス問題における問題解釈と問題点克服において有効な表現であること、クラスに依存した特徴集合およびクラスに依存した識別子が性能向上に非常に有効な手段と成り得ることを示したことである。提案手法は制約無しで汎用的に多クラス問題に適用可能であり、識別規則を人が解釈可能なクラス決定木として表現できる。課題としては、クラス決定木の構成にかかる時間計算量が大きい点が挙げられる。さらに、可視化された識別規則をもとに、識別性能の改善を図る具体的な方法についても検討を行う必要がある。

学位論文審査の要旨

主　査　教　授　工　藤　峰　一
副　査　教　授　宮　腰　政　明
副　査　教　授　原　口　　誠
副　査　教　授　今　井　英　幸

学位論文題名

Multi-Class Classification with Class-Dependent Feature Subsets and Class-Dependent Classifiers

(クラスに依存した特徴集合および クラスに依存した識別子を用いた多クラス識別に関する研究)

人間の高次情報処理の一つであるパターン認識をコンピュータに行わせる研究は,近年,応用範囲が広がり重要性を増している.パターン認識は,物理的に測定可能な特徴量を基に対象に名前をつける処理ということができる.その意味では,一般物体認識やトピック認識など数多くの応用で,これまでの方法論が前提としていた数をはるかに超える“名前”(クラス)を扱うことが求められるようになっている.

当初,クラス数の増大は計算量の問題を引き起こすだけと捉えられていた.実際,殆どの識別子(識別を行う機械)は2クラスの問題を解くものであり,3クラス以上の問題には2クラス用識別子を組み合わせることで対処してきた.しかし,実際に解いてみると,なかなか良い識別子を構成することはできなかった.この原因が,組み合わせの仕方によるものか,あるいは対象とする問題の性質が変化したのか,を明らかにすることが重要な課題となっていた.本研究はその課題に対して一つの解答を与えたものである.

本研究では,所与の多クラス問題を部分問題の集合に分割する.それらの部分問題間の関連を木で表した構造を「クラス決定木」と呼んでいる.さらに,データからクラス決定木を構築する二種類の方式を提案している.クラス決定木はクラス数分の葉を持つ二分木であり,内部ノードで上位のスーパークラス(クラスの部分集合)を2つの部分スーパークラスに分割する.根にはすべてのクラスからなるスーパークラスがある.識別によく使われる通常の「決定木」が識別を多段階に分けて行うものに対して,クラス決定木は最適な部分問題分割の方式を木の形で示す.つまり,メタレベルの決定木である.

クラス決定木の最大の利点は,現下の多クラス問題を理解する手がかりを与えてくれることにある.一例として,類似文字を処理した際に得られた知見を挙げる:1) 類似文字対が隣接する葉に配置されたことで,現在利用している特徴空間の妥当性が確認された;2) どの部分問題も平均的に高い識別精度で認識されているものの,類似文字対の認識精度が低いことが確認された;3) 漢字群とひらが

ら群が上位ノードで分割されていることが観察された;4) ひらがな同士の識別率の低さの原因が直線に基づく特徴にあることが突き止められた;5) ノード毎に特徴選択を行うことで識別率を高められることが確認された.

クラス決定木により一つの問題は階層的に二つの部分問題へと繰り返し分割される. その場合, 識別性能の向上を意図して部分問題毎に最適な特徴集合を選択すること, また, 最適な識別子を選択することは自然なアプローチである. 著者は,「クラスに依存した」との修飾語をつけてこれらの方法の有効性を実際に検証した. 問題は, 木の構築, 特徴選択, 識別子選択, をどの順でどう行うべきであるかであり, 組み合わせ的な検査を要することにある. 本研究では, これらを系統的に検討し, 結論として, 木の構成には提案した二つの方法のどちらのでも構わない一方, 識別子選択には大きな効果があることを示した.

多クラス問題が単にクラス数が 2 より大きい問題ではないことを明らかにしたことが著者の貢献である. 準最適に選ばれた特徴集合が部分問題毎に異なったという事実は, 部分問題毎に着目すべき性質が異なることを意味している. また, 部分問題毎に最適な識別子が異なることは, 難しさの異なる部分問題が統合されて多クラス問題を構成していることを意味する. つまり, 多クラス問題は単にクラス数が多い問題ではなく, 異種の問題が複雑に混在している問題であることがこの研究により明らかになった.

クラス決定木の構築に関しては 2002 年に著者が世界で初めて提案したものである. しかし, その後 1996 年頃から別の研究者が同様の試みを提案していることがわかった. それでも本研究の重要性は揺るがない. 問題分析手法としてその重要性を主張したのは著者がはじめてであり, 一般的な構築法を示したのも, 識別性能向上の基盤としたことも著者の貢献である.

本論文による成果は以下にまとめられる.

1. 多クラス問題の問題分析法として「クラス決定木」という表現を提出し, その具体的構築方法を二種類与えた.
2. クラス決定木の各ノード, つまり, 各部分問題において特徴選択行うことの有効性, さらに, 識別子選択を行うことの有効性を実験的に示した.
3. 多クラス問題を解く上では安直に各種識別子を適用することに先んじて問題分析を行うことの重要性を示した.

これを要するに, 著者は, パターン認識における多クラス問題の扱いにおいて, 伝統的な方式が抱える潜在的問題を明らかにするとともに問題分析を先んじて行うことの重要性を論じ, さらに, そのための系統的な方法論を与えた. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.