

学 位 論 文 題 名

A Study on Classifier-Independent Feature Selection

（識別子独立型特徴選択に関する研究）

学位論文内容の要旨

パターン認識において特徴選択は、与えられた特徴集合から識別に必要な特徴のみを選ぶ処理である。特徴を選択することで、(I) データの測定コストを削減できる、(II) 識別子の性能を向上させられる、(III) 識別子の構成及び識別時に要する計算時間を削減できる、(IV) データ構造の可視化・理解を容易にする、等の効果がある。最近では、パターン認識の応用範囲は (i) テキスト分類、(ii) 遺伝子配列の解析、(iii) 画像認識・画像検索、(iv) 医療診断等に広がってきており、それらの分野では千を超える特徴数を扱う必要があるため、従来の特徴選択手法のスケーリングが問題となっている。

特徴選択手法は大きく二つに分けることができる。一つは識別子依存型特徴選択であり、これは予め指定された識別子に密接に関係した評価関数の値（例えば推定識別率）を基準として特徴集合を求める手法である。二つ目は識別子独立型特徴選択であり、これはどの識別子にも有効な特徴を選択する手法である。使用するべき識別子が既知である場合には識別子を特定する識別子依存型特徴選択が有効であるものの、一般的な要請に答えるためには識別子独立型の特徴選択が望まれる。また、識別子独立型特徴選択の手法は識別子依存型特徴選択の手法に比べ、少ない計算時間で処理できる点でも好ましい。よって、本論文では主に識別子独立型特徴選択を考える。

識別子独立型特徴選択では、クラス毎の分布を出来るだけ正確に推定した後、それぞれの分布が互いに出来るだけ離れるような特徴の部分集合を求める。クラス毎の分布を正確に推定できれば最適な分類器であるベイズ識別子を構成できることから、識別子独立型特徴選択はベイズ識別子を仮定した識別子依存型特徴選択と見なすこともできる。識別子独立型特徴選択手法として、1) 確率密度関数をパラメトリック或いはノンパラメトリックな分布で推定した後、分布間の分離度により特徴を評価する方法、2) ベイズ識別境界を推定して境界線上の点における法線ベクトルの向きから特徴を評価する方法、の大きく分けて二つのアプローチがある。

識別子独立型特徴選択の従来研究はそれ程多くないものの、1) のアプローチに基づく手法として DIV 法 (Novovicova et al., 1996) があり、2) のアプローチに基づく手法として RFI 法 (Holz et al., 1994)、SUB 法 (Kudo et al., 1993)、PRISM 法 (Singh, 2003) が挙げられる。著者はまず分布推定に基づく DIV 法 (Novovicova et al., 1996) に着目した。DIV 法は分布推定後に擬似ベイズ識別子を構成できるという点から、提案者は意識していなかったが最も直接的に識別子独立型特徴選択を行うアプローチの一つである。しかし、残すべき特徴の個数

を利用者が指定する必要がある、識別子独立型特徴選択手法として利用する上で大きな問題となっていた。そこで、著者は分布間のダイバージェンス分離度に基づいて残すべき特徴の個数を自動決定し特徴選択を行う $mDIV$ 法を提案した。複数の実データや人工データを用いた実験の結果、この方式により複数の識別子の性能を維持または改善できること、二段階特徴選択（識別子独立型特徴選択の後、識別子依存型特徴選択を実行する手法）により識別子依存型特徴選択を直接利用する場合と比べて少ない計算時間で有効な特徴集合を発見できること、の二点を確認した。

一方で、この方式では依然パラメトリックな分布が十分な近似を与えない場合があること、またそれにより、特徴評価の性能が分布推定の精度に強く依存すること等の問題が残った。そこで、もう一方の方式であるベイズ識別境界の推定方式に基づく識別子独立型特徴選択手法を更に検討した。これは、 k 個の近隣サンプルを用いてベイズ識別境界上の複数点とそれらにおける法線方向を推定することで各特徴の寄与度を測る手法（ $kNFS$ 法）である。実験により、 $mDIV$ 法及び線形カーネルに基づく手法（ $LSVM$ 法）と比べて $kNFS$ 法が有効であることを確認した。特に $mDIV$ 法と比べて、設定すべきパラメータ数が少なく済むため高次元データに対しても利用し易く、また $LSVM$ 法に見られる識別境界に対する線形性の制約を受けないため多様な分布に従うデータにおいても適切な特徴選択が行える、という二点を確認した。

本研究における成果は以下にまとめられる。

- A) その重要性にも関わらずこれまでに提案された識別子独立型特徴選択手法は少なく、また比較・検討も十分ではなかった。これに対して本研究では、分布推定に基づく手法と境界推定に基づく手法の二つの観点から新たな識別子独立型特徴選択手法を提案し、これらの方式が従来の方式を性能と効率の両方において凌ぐものであることを確認した。また、少数サンプルにおいて見せかけの識別情報を示す特徴に対して、これらの示す量を適切に補正する方式も提案し、その有効性を確認した。
- B) 特徴数の増大に対する一つの有効な方法として、識別子独立型特徴選択の後に識別子依存型特徴選択を実行する二段階特徴選択が提案されている。しかし、これまではその効果の検証と有効な組合せの提案が不十分であった。これに対して本研究では、提案手法を前段階に用いることで二段階特徴選択が優れた方式であることを定量的に示すことに成功した。

今後の課題として、近年大規模データを扱う機会が増えていることから、特徴数に対してより計算量の少ない識別子独立型特徴選択手法の開発が望まれる。また、得られる訓練サンプルが限られている場合において、見せかけの有効性を示す特徴が存在する事実に対し、より適切な判定方法を確立することが挙げられる。

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	栗原 正仁
副査	教授	佐藤 義治
副査	助教授	中村 篤祥

学位論文題名

A Study on Classifier-Independent Feature Selection

(識別子独立型特徴選択に関する研究)

様々な認知的対象を事前に与えられた幾つかのカテゴリに分類するパターン認識の問題は、コンピュータの利用が単なる計算から人工知能への利用へと大きく拡大した30年前から広く工学における重要な課題となっている。その応用範囲も、初期の音声認識や文字認識などから、現在は、医療画像診断や衛星画像による土地利用の分類、撮像された画像からの興味ある対象の抽出、遺伝情報の機能解明から知識発見やデータマイニングにまで、当初の想定を大きく超えて広がっている。しかし一方では、応用範囲の広がりにつれ問題の規模が増大し、これまでの基本的な方法論をそのままでは計算量的に適用できない状況も多々生まれている。このため、より広範囲に適用可能な理論を再構築する必要性が高まっている。本研究は、その中であって、パターン認識装置の性能向上に極めて重要な働きをする特徴選択という問題に関して、基礎理論を深く検討したものである。

パターン認識における特徴選択とは、予め与えられた数多くの特徴(属性)のうち識別に本質的に貢献するものだけを選出する試みである。通常、どの特徴が識別に有効であるかを事前に判定するのは難しいため、まずは、有効と思われるなるべく多くの特徴が集められる。しかし、識別系の構成に際して利用できる「訓練サンプル」の数は通常限られるため、大規模な問題(特徴数が百を越すような問題)においては、サンプル数に対する特徴数の比が非常に大きくなり、結果として、「次元の呪い」という言葉に象徴されるように多くの識別規則はその性能を著しく低下させることになる。このことから、特徴選択の目的も、初期の測定コストの削減や計算量の削減といった目的から識別規則の性能向上を目的とするようにシフトしてきている。また、特徴選択に注目する研究分野も、初期のパターン認識分野や統計の分野から知識発見や学習理論の分野へと広がり、異なる分野でその重要性が同時発生的に再認識されつつある。

方法論としての特徴選択としては、これまで数多くの手法が提案されており、優れた比較研究もある。これまでに提案されている方式は大きく、特定の識別規則を仮定してその

識別規則の性能を最も高めるような特徴集合を求める「識別子依存型」と、特定の識別規則を仮定せずどんな識別規則にも有効となる「識別子独立型」がある。個々の識別性能を向上させるには依存型が良いものの、実際には事前に使うべき識別子が定まらないこと、また、可視化などの予備調査のためにも独立型が良く、さらに計算量の面でも有利である。そのため、本研究ではもっぱら「識別子独立型」を考察している。これらの事実は最近になって認識されつつあるもので、それ故にこれまで提案されている手法は限られている。本研究では、識別子独立型の有効性を論証し、一般の問題に有効に働く方式を提案するものである。具体的には、1) 分布推定と分布間の分離測度に基づく方式、ならびに、2) 最適な識別境界の推定とその境界上の法線ベクトルに基づく方式、の二手法を提案している。前者は既に提案されていた識別子依存型の手法を識別子独立型として利用できるようにしたものであり、識別子独立型の特徴選択手法として標準的な方式を提案するものである。

一方、後者は、前者の手法で扱い難いようなクラスの分布形状が複雑な問題にも適用可能な手法となっている。この方法は、最適なベイズ識別子の境界を推定し、境界上の数点における法線ベクトルの方向から各特徴の重要性を測るものである。世界標準の実データを用いた実験を通して、多くのデータにおいて前者の方法を凌ぐことを確認している。これらは、現在の識別子独立型特徴選択手法の中で最も優れたものと言える。また、問題規模の拡大への対処方法として、前段で識別子独立型の特徴選択を行い、その後、後段で識別子依存型の特徴選択を行う、「二段階特徴選択」を考察し、提案したそれぞれの手法が前段の方式として有効であることを数値的に示した。基本提案は既にあったものの、手法の多様な組み合わせの比較において、提案する方式を取り入れることが最も有効であることを示した功績は大きい。

本論文による成果は以下にまとめられる。

1. これまで十分な検討がなされていなかった識別子独立型の特徴選択手法に関して、その重要性を論証するとともに、標準的かつ実用的な二手法を提案し、それらの有効性を現実的なデータを用いて示した。
2. 限られた個数のサンプルから特徴選択をする場合、各特徴がそれらの本来持っている以上の分離性能を示すことがあることを指摘し、その対処法を考案した。
3. 近年の問題規模の増大に対する実用的な対処法として二段階の特徴選択方式の有効性を実証し、今後の多くの応用に対する有望な方式となることを明らかにした。

現在、パターン認識の問題はその応用範囲を大幅に拡大してきており、従来方式では計算量的にも性能の面でも有効な特徴選択が行えない状況が増えてきている。本研究は、この情勢に対して、現実的な見地からこれまでの方法論を見直し、基礎的な考察を加えた上で優れた方式を提案したものであり、その貢献は大なるものがある。この成果は今後の多くのパターン認識応用において一つの基盤技術として幅広く使われるであろう。

これを要するに、著者は、パターン認識における特徴選択の問題において、新たな方法論の必要性和重要性を明らかにし、また、その具体的な方法論を示すことでこの分野の重要な基盤技術を着実に進めたものであり、パターン認識分野の進歩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あるものと認める。