様式 C-19

科学研究費補助金研究成果報告書

平成２２年６月１０日現在

研究種目：若手研究（B）
研究期間：２００８～２００９
課題番号：２０７００２０３
研究課題名（和文）　ラフ集合理論を応用したソフトな特徴選択手法の開発

研究課題名（英文）　Developing Soft Feature Selection Method by Applying Rough Set Theory

研究代表者
天元　宏（TENMOTO HIROSHI）
釧路工業高等専門学校・情報工学科・准教授
研究者番号：８０３２１３７１

研究成果の概要（和文）：ラフ集合理論での粒度の考え方を学習理論へ応用し、識別対象のデータを識別に最適な粒度で離散化する手法を検討した。粒度の最適化は学習サンプルの識別状況を情報量基準で評価することで行った。最適化した各特徴の粒度がその特徴の識別への貢献度評価を与えることを確認できた。従来法では各特徴の取捨選択の指標しか得られないのに対し、提案手法では特徴そのものをどの程度の精度で記述するべきかという指標を得ることができた。

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	１，４００，０００	４２０，０００	１，８２０，０００
２００９年度	８００，０００	２４０，０００	１，０４０，０００
年度			
年度			
年度			
総　計	２，２００，０００	６６０，０００	２，８６０，０００

研究分野：総合領域
科研費の分科・細目：情報学・感性情報学・ソフトコンピューティング
キーワード：①感性情報学　②画像・文書・音声等認識　③機械学習　④ソフトコンピューティング　⑤アルゴリズム

１．研究開始当初の背景

パターン認識の研究分野は、画像等の観測データを直接扱う信号処理寄りの特徴抽出と、特徴抽出後の特徴空間における識別機の構成や誤差評価などを扱う学習理論に大きく分けられる。

一般に、前者の特徴抽出で測定された特徴は大規模となる傾向があるが、大規模な特徴はまた、後者の学習理論において次元の呪いと呼ばれる性能低下を引き起こす。そのため、前者の特徴抽出で測定された大規模な特徴集合から、識別に貢献する少数の特徴を抜き出す特徴選択に関する研究が国内外にて盛んである。

特徴選択の主要な手法は、元の大規模な特徴集合から一部の特徴を仮に選択し、それによる識別性能が向上するように、選択した部分特徴集合を修正して行くものである。しかし、それらの手法では、識別性能を向上させるために各特徴を使うべきか捨てるべきかという指標を得ることはできるものの、特徴

そのものをどの程度の精度で記述(表現)するべきかという指標を得ることができない。

そこで本研究では、近年ファジイ理論のラフ集合の分野で研究が盛んな粒度(各特徴の測定精度)調整の考え方を学習理論へ応用し、識別対象となるデータの識別に最も適した粒度で離散化して識別を行う手法を検討する。

提案手法における粒度の調整は、任意の粒度における学習サンプルの識別状況を情報量基準により評価することで実行する。

提案手法により最適に調整された各特徴の粒度はまた、その特徴の識別への貢献度評価を与える事から、提案手法は特徴選択を一般化した手法と見なせ、申請者らはこの手法をソフト特徴選択と呼んでいる。この事実は、提案手法がパターン認識システムにおける特徴抽出と学習理論を統一的な見地から捉え、システム全体のコスト削減、高速化、そして高性能化を一体として統合的に遂行できる枠組みとなる可能性がある事を示している。

さらに、システムを利用するユーザーの観点より評価すると、生成される識別規則の可読性も重要となる。近年隆盛な SVM やニューラルネットワークといった手法ではユーザーには直感的に理解不能な規則を生成してしまう。これに対し、提案手法の規則は各特徴の区間(高い・低い等の程度)を条件とした IF-THEN 列となり、極めて可読性が高い識別規則となる。

２．研究の目的

提案手法の基本アイデアは、すでに論文(M. Kudo and H. Tenmoto, Optimal Division for Feature Selection and Classification. Proceedings of the International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics, 2005, 106-107.)にて公開しており、その有効性をある程度示している。本研究の研究期間内には、そのアルゴリズムの更なる精密化、及び理論的な特性解析を行う。また、同時に、大規模計算機実験による性能実証や、識別構造の可視化手法などへの応用の可能性を探る。以下にその具体的な内容を述べる。

本手法はまず、一般に連続値で表現される特徴を軸ごとに任意の間隔で離散化し、特徴空間をセル(小領域)の集合に分割する。各セルは学習サンプルの分布状況により(a)単独のクラスの学習サンプルから成る清潔な領域と、(b)複数のクラスの学習サンプルが混在する汚れた領域とに分類される。

離散化の精度を上げてセルの分割を細かくすれば、(a)の領域が増加し(b)の領域が減少する。結果として学習サンプルの識別精度は向上するが、クラス間を分離する識別境界は学習サンプルに極度に依存した複雑なものとなり、逆に未知サンプルに対する汎化性能を低下させる結果となる。

本手法は、このセルの分割状況に関するすべての情報をビット列で表現し、情報量基準の見地からそのビット長を評価することで、対象の識別に最も適した離散化、すなわち識別情報の粒度を追求する。

先の論文にて本手法を公開した時点では、セルの分割状況に関する情報を以下の各項目に分けて評価し、その総和が最小となる分割を最適なものとして採用していた。

(1) 各セルにおける(a)(b)の分類情報
(2) (a)のセルに対するクラスラベル(単独)の情報
(3) (b)のセルにおける具体的な誤識別情報(どのサンプルをどのクラスへ誤識別しているか)
(4) 軸毎の分割数に関する情報

ここで(4)と(3)はトレードオフの関係にあり、(4)を増加させると(3)が減少し、逆に(4)を減少させると(3)が増加する。(2)に関しては、素朴に評価すれば(4)の増加と共に増加してしまうが、同じクラスラベルを持つ相異なるセルが多数存在する状態となるため、クラスラベルが同じであるという情報を利用すれば、更なる情報圧縮の可能性がある。

そこで本研究の研究期間に行う第１の内容として、上記(2)に対し、部分クラス法(M. Kudo, S. Yanagi and M. Shimbo, Construction of Class Regions by a Randomized Algorithm: A Randomized Subclass Method. *Pattern Recognition*, **29**, 4(1996), 581-588.)の応用による更なる圧縮手法の検討を行う。

部分クラス法は学習サンプル集合に対して他クラスを排他する極大部分集合族を求める手法であり、クラスタ間の重複を積極的に認めるクラスタリング手法と見なすことができる。提案手法におけるセルを学習パターンの代表点と捉えれば、(b)のセルを排他しつつ(a)のセルを可能な限り多数含む巨大なクラスタを求めることになり、そのクラスタに対して一つのクラスラベルを割り当てることにすれば、上記(2)を大幅に圧縮できる。

また、学習サンプル数が無限大に向かう際の提案手法における学習誤差のベイズ誤差への収束率の理論的な解析も行い、それによる各種の識別機との比較も行なう。

さらに、提案手法を実際の様々なパターン認識データに適用し、その現実的な適用限界を明らかにする。

３．研究の方法

平成２０年度にまず、これまでの提案手法に、部分クラス法を応用してさらにビット列を圧縮する手法の細部を検討し、計算量やベイズ誤差への収束率の算出など理論的な適用限界の評価を行う。

その後、計算機プログラムへの実装を行うなど、現実のパターン認識データに対する計算機による大規模性能評価実験を行うための準備も進める。

部分クラス法の組込みまでが終了した段階で、申請者が第１回よりこれまで継続的に参加し、研究発表を行ってきている国際会議 SPR(Statistical Techniques in Pattern Recognition)2008 にて成果を発表する。この SPR は国際パターン認識連盟(IAPR; International Association of Pattern Recognition)の第１専門委員会(TC1)が中心となって開催している統計的パターン認識に関する国際会議であり、世界中で本研究課題に関する最も有力な評価及びアドバイスが得られる。

提案手法の基礎理論は平成 17 年４月にアメリカ合衆国カリフォルニア州ニューポートビーチにて開催された特徴選択に関する国際会議 FSDM2005 において発表を行っており、それに対して、データマイニングの分野において単独の特徴量(1 変量)データの離散化手法に関して多くの業績を持つ Marc Boulle 博士(フランステレコム)より強い関心を寄せられている。そこで、この時点での成果について、電子メールにて博士からのコメントを得、それ以降のより精密な研究遂行へとフィードバックさせる。

また、統計的パターン認識及び画像認識の分野で顕著な成果を挙げているチェコ科学アカデミーパターン認識部門の Michal Haindl 博士、Pavel Pudil 博士、Jiri Grim 博士、Iana Novovicova 博士らと交流を深め、本研究課題に関して、代表の Michal Haindl 博士に研究協力を依頼し、承諾を得ているため、この時点までの成果に対し、チェコ科学アカデミーの研究者らと議論し、統計的パターン認識及び画像認識の分野から本研究課題へのフィードバックを行なう。

以上の成果発表及び議論の成果を総合し、提案手法の細部の再検討を行うと共に、それまでの研究成果をまとめた論文原稿の作成に着手する。

その他、適宜、電子情報通信学会の全国大会やパターン認識・メディア理解研究会等の国内学会へ出席し、関連研究の情報を収集する。

平成２１年度には、現実の様々なパターン認識データに対し、提案手法の適用実験を行う。パターン認識の研究分野で実証試験に幅広く用いられている UCI(University of California, Irvine)の Machine Learning Database をはじめとし、画像認識関連など、実際に社会で必要とされている大規模パターン認識データを対象に提案手法の有効性について実験的に評価を行う。

その他、提案手法の、パターン認識データにおける識別構造の可視化手法への応用を検討する。この分野においては千葉大学画像情報工学科の森助教並びに北海道大学大学院情報科学研究科の工藤教授が先進の研究成果を挙げているため、両研究者との研究交流を進める。

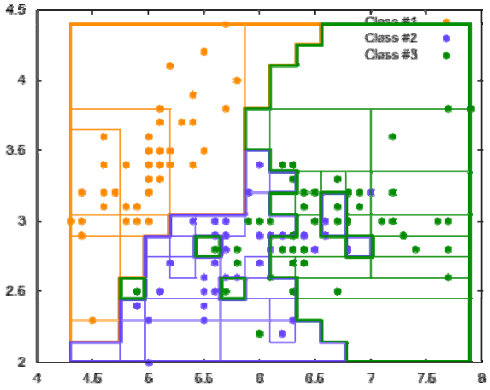
その他、適宜、電子情報通信学会の全国大会やパターン認識・メディア理解研究会等の国内学会へ出席し、特に提案手法の応用面での関連研究について調査を行う。

最後に、本研究課題の最終成果として、論文を Pattern Recognition 誌へ投稿する準備を進める。

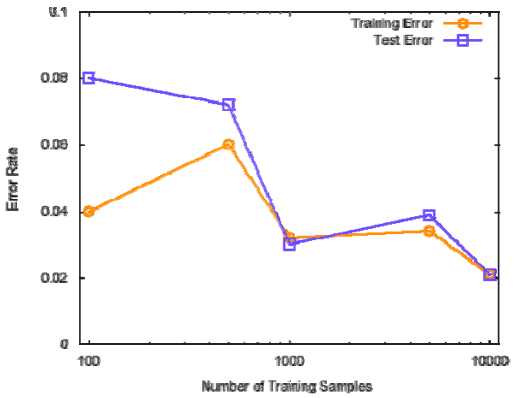
４．研究成果

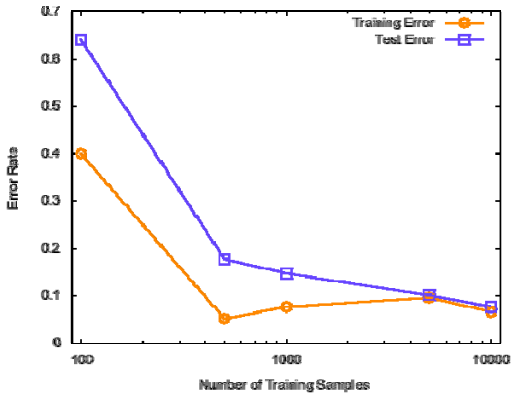
平成２０年度は、申請者のこれまでの提案手法に部分クラス法を適用して高い精度で識別する手法を検討し、計算機プログラムへの実装を行い、現実のパターン認識データに対する計算機による性能評価実験を行った。

実行結果の一例として、パターン認識手法の試験データとして広く用いられている Fisher のアヤメデータの最初の２特徴に対する提案手法による結果を以下の図に示す。



部分クラス法の適用により、粗い粒度で分割されたセルの集合を少数の部分クラスで排他的に被覆していることが確認できた。





また、学習サンプル数を増加させることによる誤差の変化を、２次元及び１０次元の人工データに対して実行して確認した図を以上に示す。学習サンプル数を増加させることにより学習誤差と検査誤差が一致に向かっており、提案手法が汎化に成功していることを確認できた。

さらに、その結果を国際会議にて発表し、本研究課題に関する評価及びアドバイスを得、今後の画像認識等への応用の可能性を検討した。

平成２１年度は、それまでの成果を総合し、提案手法の細部の再検討を行うと共に、パターン認識の研究分野で実証試験に幅広く用いられている UCI(University of California, Irvine)の Machine Learning Database をはじめとし、画像認識関連など、実際に社会で必要とされている大規模パターン認識データへの提案手法の適用可能性について実験的に評価を行った。

また、提案手法のパターン認識データにおける識別構造の可視化手法への応用の検討を目的として、国内の先進の研究者との研究交流を進め、今後の発展課題の検討とまとめを行った。

本研究で提案する手法は特徴(特徴空間の軸)毎に識別に最適となるように分割数を決定することから、分割が不要な特徴を特定・検出することができ、汎化性能の高い識別機の構成という側面のみではなく、一般にかなりの高次元(多変量)データとなるパターン認識データに対する特徴選択を学習と同時に、また、効率的に実行するという特色を備えている。

また、ラフ集合理論の見地から、過剰な粒度で収集された特徴を、より荒い適度な粒度で捉えなおす、という側面も併せ持っている。このような粒度による特徴選択の一般化は、パターン認識研究の分野では極めて独創的なものであり、近年行き詰まりを見せている特徴選択の分野にも新たな貢献ができた。実際に、発表した文献は特徴選択の研究者にも注目されていた。

さらに、提案手法では各特徴の区間(高い・低い等の程度)を条件とした IF-THEN 列から成る可読性が高い識別規則の生成が可能である。従来の決定木や決定リストと同様な可読性を、それらとは異なる粒度による特徴選択の観点から生成できることは、多様な判断基準をユーザーに提示する上でも大変有効であった。

また、Kudo らが 1996 年に提案した部分クラス法は、未知サンプルを含めたサンプルの分布が特徴空間内でクラス間に完全分離していることを前提としているが、現実的にはその仮定は妥当ではない。それに対し、本研究の手法はクラス間での重複をある程度まで認めた上で、精密に分布構造を解析する手法であり、部分クラス法のより現実的な発展形をなすものとすることができた。

５．主な発表論文等

〔雑誌論文〕（計１件）

① H. Tenmoto and M. Kudo, Soft Feature Selection by Using a Histogram-Based Classifier. Lecture Notes in Computer Science, Advances in Pattern Recognition, 査読有り, 5342, 2008, 582-591.

〔学会発表〕（計１件）

① H. Tenmoto and M. Kudo, Soft Feature Selection by Using a Histogram-Based Classifier. Joint IAPR International Workshops on Syntactic Pattern Recognition and Statistical Pattern Recognition, 2008年12月5日, Orlando, Florida, USA.

６．研究組織

(1)研究代表者

天元　宏（TENMOTO HIROSHI）
釧路工業高等専門学校・情報工学科・准教授
研究者番号：８０３２１３７１