

学位論文題名

Visualization Techniques of Multi-Dimensional Data
for Pattern Recognition

(パターン認識における多次元データの可視化技法に関する研究)

学位論文内容の要旨

近年、コンピュータ技術の急速な発展に伴い、高速の中央演算処理装置 (CPU) や十分な容量の記憶装置を利用できるようになり、かつてなかった規模のデータの蓄積と処理が可能になって来ている。中でも、人工衛星から送られて来る計測データの処理、癌診断などに利用される X 線画像の解析、遺伝子データに基づく各種機構の解析、さらには多様な環境下での音声情報処理や手書き漢字認識など、大規模なパターン認識の需要が高まってきている。また、近年では、マーケティング・リサーチなどの分野において、顧客の購買履歴などのデータを分析し、その背後に隠された重要な規則や属性間の関係を明らかにすること (データマイニング) を目的とした高度な技術の開発が望まれてきている。

本研究では、データがカテゴリ別に与えられている場合に、新しく与えられたデータを適切にそのカテゴリのいずれかに割り当てることを目的とした、いわゆる、「パターン認識」の問題に焦点を絞り、データやカテゴリの構造を明らかにすることで、より高速・高性能な識別を行うための方法論を述べる。これはまた、データの背後にある規則を明らかにすることにも相当するため、同様の方法がデータマイニングを含む多方面で利用できると期待される。

一般に、パターン認識で扱われるデータは多くの属性値 (一般に「特徴」と呼ぶ) により表現され、結果として有限次元ベクトル空間 (「特徴空間」と呼ぶ) 上の多次元ベクトルの集合として表される。しかし、測定される属性は経験的に決められることが多いため、識別に無関係である冗長な特徴を含む場合や、反対に、誤りの少ない識別を保障する十分な特徴を含んでいない場合がある。そのため、現在の特徴空間の十分性を評価し、その有効性を解析することは、高性能な識別規則を構成するための重要な課題の一つとなっている。しかしながら、通常、特徴空間は非常に高次元であることが多く、我々人間はデータの構造を直接観察し把握することはできない。そこで、本研究では、パターン認識における特徴空間の評価の問題を多次元ベクトル集合の可視化技術によって解決することを考え、その方法論と応用について論じる。高次元空間に散らばるデータを二次元または三次元空間に可視化することで、我々は直接そのデータを視覚的に捉えることができるようになる。このことを利用して、人間の優れた分析能力の支援の下で、データを解析し、それによって高性能な識別システムを構成できると期待される。

そのような可視化の試みは、主成分分析に基づく方法やニューラルネットワークを用いた方法など、従来から多くの研究がなされているものの、1) パターン認識における可視化を利用した識別規則の改善では適用可能な規則の種類が限定され、高性能な規則の構築が期待できない、2)

データ集合の非線形的な構造 (例えば, 一つのカテゴリが複数のクラスタから成る場合) を適切に表現できない, 3) カテゴリ間の分離情報を適切に表現できない, 4) データ数が非常に多い場合に射影結果の可読性が著しく損なわれる, などの問題が存在する. そこで, 本研究では, 従来の多くの可視化技法の特性を目的別に整理し, それぞれの技法の得失を明らかにする. さらに, 幾つかの問題点を克服するために, グラフ表現に基づく可視化技法を提案する. これにより, 従来の手法では表現することが難しかった情報, 特に, カテゴリ間およびカテゴリ集合間の分離度を適切に表現することができることを示す. 続いて, カテゴリ数が十を越すような識別問題において, 本質的な識別の難しさが何処にあるのかを明らかにする方法論と, その方法論に基づいて, より高性能な識別規則を構成するための一つの方法を検討する. また, これらの提案手法の有効性を人工データと手書き文字データを用いた実験結果に基づいて示す.

本論文の構成は以下の通りである. 第1章で, パターン認識の概略, および, 可視化技法が要求される背景について述べる. 第2章では従来の可視化技法を紹介し, 目的別に性質を整理する. さらに, 明らかになった問題点を踏まえて, 第3章で新たな可視化技法を提案する. 続く第4章で, 提案手法の有効性を検証するとともに従来手法との比較を行う. 第5章で, カテゴリの数が多い場合に, 問題を決定木として表現し直すことの有効性を論じ, その表現を利用した手書き文字認識について述べる. 最期に第6章で本論文の総括を行い, 将来解決すべき問題を挙げる.

現在のパターン認識において望まれることは, 従来の方法論がそのまま適用できないような大規模な問題 (カテゴリ数や特徴数が多い場合) に対して, 十分効率良くかつ高精度な識別システムを構築することである. この問題に対して本研究では, 多くの特徴で表現されたデータを適切に解釈する一つの方法論を与え, また, カテゴリ数が多い問題に対して, 本質的な問題の所在を明らかにする表現を与えたものである. 言い換えると, それ程無理のない仮定の下で, より大規模な問題を取り扱うための一つの方向性を示したものである. 今後, これらの提案を既存の各種技法と組み合わせることで, より一層効果的なデータの解釈とパターン認識システムの構築が可能になると期待される.

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

Visualization Techniques of Multi-Dimensional Data for Pattern Recognition

(パターン認識における多次元データの可視化技法に関する研究)

各種の対象を特定の目的において分類する、いわゆる「パターン認識」の機構を計算機上で実現するための技術開発が始まって既に30年以上経つ。パターン認識は我々がやすやすと日常的に行う処理であるものの、計算機でその処理を模倣するためには、我々自身の認識機構そのものを客観的に分析し数学的にモデル化する必要がある、これまでの研究によりその難しさが徐々に判明し、さらなる研究と技術開発が待たれている分野である。特に、近年のコンピュータハードウェアの急速な発展に伴い、かつてなかった規模のデータに関するそのような技術の需要が増してきている。人工衛星から送られて来る計測データに基づく土地利用の分類、癌診断などに利用されるX線画像の解析、遺伝子データに基づく各種遺伝的機構の解析、さらには多様な環境下での音声情報処理や手書き漢字認識などである。しかし一方では、問題の規模の増大に伴い、既存の多くの方法論において計算量的なあるいは結果の解釈や表現方法において限界が見られ始めている。

本研究は、このようなパターン認識の現況にあって、カテゴリのラベルを付与されているデータを可視化することで、利用者の優れた直感や分析能力をデータの表現空間である「特徴空間」の理解のために利用する方法論を検討したものである。具体的には、特徴空間におけるカテゴリ間の分離性や各カテゴリの構造を可視化を通して利用者が視覚的に理解できるようにするための技法を論じている。これにより、現在構築中の識別システムの改善可能性を見極め、またどのような方向で改善策をとることができるかを議論するための一つの手がかりを与えることが可能となる。

そのような可視化の試みは、主成分分析に基づく方法やニューラルネットワークを用いた方法など、従来から多くの研究がなされているものの、1) パターン認識における可視化を利用した識別規則の改善では適用可能な規則の種類が限定され、高性能な規則の構築が期待できない、2) データ集合の非線形的な構造(例えば、一つのカテゴリが複数の成分分

布から成る場合)を適切に表現できない、3) カテゴリ間の分離情報を適切に表現できない、4) データ数が非常に多い場合に射影結果の可読性が著しく損なわれる、などの問題が存在する。そこで、本研究では、従来の多くの可視化技法の特性を目的別に整理し、それぞれの方法の得失を定性的に明らかにし、さらに、この分析によって明らかになった問題点を克服するために、グラフ表現に基づく新しい可視化技法を提案している。これにより、従来の手法では表現することが難しかった情報、特に、カテゴリ間およびカテゴリ集合間の分離度合を適切に表現することができることを示している。また、提案した可視化技法に基づいて、カテゴリ数が多い場合の識別問題を扱う場合に、本質的な識別の難しさがどの部分のカテゴリ集合にあるのかを明らかにする一つの表現方法を提案し、その表現を分析することで総合的に識別規則の性能向上が図れることを手書き文字データを例として示している。

本論文による成果は以下にまとめられる。

1. パターン認識の分野において、従来十分に検討されていなかった特徴空間の十分性の評価を利用者自身が可視化技法を通して行うための一つの方法論を与えた。
2. 従来の可視化技法において不十分であったカテゴリ間の分離性の表示を効果的に行える手法を提案した。
3. カテゴリ数や各カテゴリのデータ数が膨大になった場合にも有効となる可視化技法を今後開発する上での一つの基礎を与えた。
4. カテゴリ数が多い問題において、全体の識別性能向上に対して特にどの部分問題を具体的に検討すべきであるかを個所とともに示す決定木表現を提案し、実際に識別性能の向上が図れることを手書き文字認識において示した。

現在のパターン認識技術において望まれることは、従来の方法論がそのまま適用できないような大規模な問題(カテゴリ数やデータ数、あるいは特徴数が多い場合)に対して、十分効率良くかつ高精度な識別システムを構築することである。この点に対して本研究は、すぐれた識別規則を構成するための前提となる十分な識別情報を有する特徴空間を構成する問題を、大規模データの仮定の下で論じ、一つの方法論を提示したものである。また、提案手法の有効性を多面的かつ客観的に評価している。この技術進歩により、現在の識別システムの限界を利用者が自ら視覚的に判断し、より高性能な識別システムを構成する上で考慮すべき問題が何であるかをより適切に理解することが可能となる。

これを要するに、著者は、パターン認識の分野において、今後ますます重要性を増す大規模データの扱いに関して一つの方向性を示したものであり、数理情報工学及びパターン認識の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。