

学 位 論 文 題 名

パターン認識における k 近隣法の高速化に関する研究

学位論文内容の要旨

近年、情報処理技術の進歩に伴い人工知能の実現に向けて認識問題の自動化に対する気運が高まっている。認識問題では、画像認識、音声認識などの各種応用面で個別に技術開発が多く行われているものの、これらの技術の根底にはパターン認識における基礎的手法が用いられる。パターン認識に関しては様々な研究が為されてきた。中でも 1960 年代から 1970 年代にかけて行われた k 近隣法とその関連の多くの試みは、現在のパターン認識技術の礎を築き、その後も地道な努力の積み重ねで多方面に大きな進歩を遂げてきている。また、近年、計算機のパフォーマンスは大きく向上し、認識機能を搭載した機器が相次いで実用化されることに伴いその基盤であるパターン認識技術への関心も益々高まっている。

パターン認識において古典的な手法である k 近隣法は未知サンプルの識別を以下の手順で行う。(1) 未知サンプルと全ての訓練サンプルとの間の距離を計算する、(2) 距離の小さい順に k 個のサンプルを選び出す、(3) 選出した k 個のサンプルで多数決を行い、その中で最も数が多かったクラスをもって未知サンプルのクラスとする。このように k 近隣法は非常に簡単なアルゴリズムによって行われるにも関わらず、性能評価が解析的に行え、また容易に実現できることから実用性は非常に高く、パターン認識における重要な手法となっている。この手法は認識以外にも応用があり、例えば、密度推定問題などでは上述のアルゴリズムの (1) 及び (2) の処理を行うことにより得られる k 個のサンプルを用いて分布密度を推定することができる。

しかし、この k 近隣法はアルゴリズムの性質上、訓練サンプル数が増加するとそれに比例して計算量が増していくという特長があり、訓練サンプル数が非常に大きな場合においては大きな障害となる。そのために計算量の削減を目的として様々な研究がされてきた。初期には、識別時に用いる訓練サンプル集合の大きさを減らす研究が主に行われ、多くの手法が提案された。それらの手法は、計算量はかなり小さくなるものの、元となる訓練サンプル集合を正しく識別することに主眼を置いていたために、未知サンプルを持ってきて識別した結果、本来の k 近隣法と異なる識別規則を構成するという事態が起り得た。その後、次第にサンプル自体を減らすのではなく、不必要な計算を減らす事により計算量を削減する手法が相次いで提案された。今日に至ってもその傾向の元に様々な努力が続けられている。本研究でも本来の識別結果を保持しながら計算量の削減を行うことを目的とし、訓練サンプルの探索中に即座に処理を終了できるような条件を付加する事によって高速化を計る。

本研究では識別結果が本来の k 近隣法と同じであるという条件の下で計算量を減らすことを目的とする。実際には全ての訓練サンプルとの距離計算を行うことなしに真の k 近隣点を見い出すことで計算量の削減を行う。本手法では k 近隣点を探索する処理の途中で、探索を打ち切っても解が保証される条件を考察し、その有効性及び特性を明らかにする。なお、この条件は分枝限定法などの探索手法と組み合わせることで効果を発揮する。

本論文は全部で七つの章から成り、その概要は以下の通りである。

第 1 章は本論文の序章である。パターン認識問題における従来の試みについての概説を行う。その後、本論文の構成を示し、目的を述べる。

第2章では、まず、数学的な準備として本論文中で使用する用語の説明を行う。その後、本論文の主題である k 近隣法の従来研究の概説を行い、それらの特長を示すと共に問題点を指摘する。特に、 k 近隣法の改良手法を分類し、本研究との関連する手法に焦点を当ててそれらの詳細を述べる。

第3章では、提案手法の説明を行う。まず、 k 近隣法の問題点を指摘する。その後、提案手法の基本アイデアを説明する。提案手法は未知サンプルの識別時に複数の条件を付加することにより任意の探索手法の探索途中に処理を打ち切れるようにしたものである。ここで付加する条件は十分条件であり、必ずしも有効であるとは限らない。しかし、多数の未知サンプルに対して適用した場合、平均的に計算量の削減が期待できる。始めにクラスラベルがない場合について適用可能な条件を説明する。クラスラベルがない場合はパターン認識以外の用途にも用いられ、様々な分野で幅広い応用が考えられる。また、クラスラベルがない場合については即座に探索を中断する条件の他にも、実際に計算を行うサンプルの対象を小さくする条件も提案し、その有効性に関する検討を行う。その後、クラスラベルの付いたパターン認識問題のみに適用できる条件の提案を行い、その概説を行う。

第4章では、提案手法の理論的な解析を行う。提案手法の条件を満たし得る理論的な確率について論述すると共に、人工データを用い実験を行い、実際に条件を満たした確率を調べ、理論上の確率との差異について議論し、提案手法の特質を明確にする。

第5章では、まず、 k 近隣法の計算量を削減する手法の一つである分枝限定法の概要を説明し、提案手法と組み合わせた場合のアルゴリズムを述べる。そして、両者を組み合わせた場合において実際の自然データを用い実験を行い、提案手法の効果と特質を明らかにする。実際に提案手法の有効性を示すために、分枝限定法と分枝限定法に提案手法を組み合わせた場合の比較結果を示す。また、前章で述べた論理的な確率との比較検討を行う。

第6章では、前章で用いた分枝限定法と同様に k 近隣法の計算量を削減する手法の一つである大町らの手法を用い、提案手法と組合せることを考える。大町らの手法は分枝限定法と同じように探索木を作成する手法であり、クラス毎に部分木を作成することに特長がある。まず、大町らの手法の概略を説明し、その後、提案手法と組み合わせたアルゴリズムを示す。そして、第5章と同様に実験を行い、大町らの手法と大町らの手法に提案手法を組み合わせたものとの差異を調べて、前章の結果との比較検討を行い、組合わせる手法によって提案手法の効果が変化するかを明らかにする。また、実際に識別を行う前の段階における、提案法の有効性の推測について議論を行う。

第7章は終章である。本論文の結論を示し、提案手法の特質、問題点を述べる。そして、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 佐 藤 義 治
副 査 教 授 宮 腰 政 明
副 査 助 教 授 工 藤 峰 一

学 位 論 文 題 名

パターン認識における k 近隣法的高速化に関する研究

パターン認識問題では、画像認識や音声認識などの各種応用面で個別に技術開発が多く行われているものの、これらの技術の根底には対象をその物理的性質に基づいて幾つかのカテゴリあるいはクラスに分類するパターン認識の基礎的手法がある。

パターン認識の古典的な手法である k 近隣法は、クラスが不明の未知サンプルとクラスが既知の訓練サンプル間の距離を計算し、距離の小さい k 個の訓練サンプルが属するクラスの多数決を採って、未知サンプルのクラスを決定する。 k 近隣法はアルゴリズムの簡単さにも関わらず、性能評価が解析的に行えることと実現が容易であることなどから広く応用に用いられ、パターン認識における重要な手法となっている。しかし、 k 近隣法はアルゴリズムの性質上、訓練サンプル数の増加に比例して計算量が増す問題があり、実用性の観点から計算量の削減が望まれている。

本論文では、このような背景から、 k 近隣点を求める探索処理で探索時に終端条件を用いて高速化する手法を提案し、その有効性を検証したものであり、成果は以下の点に要約される。

(1) 探索木を利用する代表的な方法について、探索の途中で真の解を発見した場合にその時点で処理を終了できる十分条件を提案し、探索を最後まで行う従来手法の高速化を計った。

(2) 提案手法の理論的解析を行い、特徴数(次元数)と訓練サンプル数に基いて上記の条件により探索効率が改善する割合を上界として評価した。また、次元数が大きいとき、この評価式が十分によい推定値を与えることを示した。

(3) 幾つかの実用的なデータベースを用いて、この手法が最大で 37%、平均で 17% の計算量を改善することを示した。

これを要するに、著者は、情報処理の有効な技法として、パターン認識のみならず密度推定問題などにも応用可能な k 近隣法の改善方法を提案し、実例による検証を行ったものであり、数理情報工学ならびにパターン認識工学の分野に寄与するところ大なるものがある。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。