審 査 の 結 果 の 要 旨

氏　　名　　平井　聡

大量のデータからの知識を獲得するための機械学習・データサイエンスにおいて、近年では学習対象が時間と共に変化する動的な設定を扱うことが多くなってきている。例えば、時系列データからのクラスタリングの問題では、クラスターの数が動的に変化するような状況がこれに相当する。このような動的な学習においては、学習対象の構造の変化をいかに正確に、いかに早期に検知するかが重要な問題となっている。本論文では、上記の問題を克服するために、記述長最小化原理（Minimum Description Length（MDL）principle）に基づいて、データの確率モデルの変化とその予兆を検知する方法論を開発した。MDL原理とは、与えられたデータに対して総記述長を最小にするモデルを最適なモデルと見なすモデル選択原理である。一般に、モデルは、クラスター数のような離散値で指定されるので、変化は突発的に起こるとみなされてきた。しかしながら、本論文では、モデルの変化は潜在的な空間の中で漸進的に起こっていると考え、そのような変化を捉えるために3つの連続値指標を提案した。それは（１）「構造的エントロピー」（２）「逐次的MDL変化統計量」、（３）「カーネル複雑度」　である。いずれもMDL原理に基づいて構成される量である。これらの3つの指標に基づいて、モデルの変化とその予兆を検知するための具体的アルゴリズムを構築し、理論的考察と実験的検証を行うことにより、動的な設定における機械学習の体系を築いた。

本論文は「Detecting Model Changes and their Early Warning Signals with the Minimum Description Length Principle」（記述長最小原理に基づくモデル変化と早期警戒信号の検知)と題し、7章からなる。

第１章「Introduction」（序）では、時系列データからのモデル変化検知の問題、およびその早期警戒信号の発信の問題を提起し、その工学的意義を唱えている。

第２章「Preliminaries」（準備）では、本論文を貫くモデル化の原理としてMDL原理を説明し、そこにおける基本的な概念として正規化最尤符号長（Normalized Maximum Likelihood Codelength：NML）と動的モデル選択を導入している。

第３章「Structural Entropy」（構造エントロピー）では、モデル変化の不確実度合いを測る連続値指標として「構造的エントロピー」を提案している。各時刻に複数のデータが与えられ、これらをクラスタリングする設定を考える。クラスターは潜在変数である。クラスター数(モデル)が時間と共に変化する場合、その過渡期にモデルがどの値

をとるかについて不確実になる状況が現れると考える。その不確実度合いを定量化したのが構造的エントロピーである。これは潜在変数モデルに対するデータのNML符号長と温度パラメータを用いた確率分布に対するエントロピーとして定義された。そこで、温度パラメータを適切に設定すると、構造的エントロピーの値が高くなる時点を検知することにより、信頼性の高いモデル変化の早期警戒信号を発信できることを理論的に示した。人工データを用いて、ガウス混合分布の混合数やARモデルの次数が漸進的に変化するときの予兆を検知できることを示した。さらに現実のビール購買データからのユーザの購買パタンの変化予兆検知や、電力消費データからの電力消費パタンの変化予兆検知に応用して、その有効性を検証した。

第４章「Sequential MDL Change Statistics」（逐次的MDL統計量）では、漸進的モデル変化の度合いを測る連続値指標として「逐次的MDL変化統計量（SMCS）」を提案した　MDL変化統計量とは、ある時点のモデル変化度合いを、その時点の前後で別々のモデルを用いてデータを圧縮したときと、同一のモデルを用いてデータ圧縮した時のNML符号長の差として定義されたものである。この概念は既に存在していたが、これを潜在変数モデルに対して、かつデータが逐次的に入力される毎に計算するように修正したものがSMCSである。SMCSの立ち上がりの状況を見て、それが一定の閾値を超えたときにモデル変化の予兆を検知するアルゴリズムを提案した。そこで閾値パラメータの適切に設定することにより信頼性の高い早期警戒信号を発信できることを理論的に示した。また、この方法により、SEと同等以上の性能を有するモデル変化予兆検知が実現できることを、3章で扱ったのと同様の人工データ及び現実のデータを用いて実験的に検証した。

第５章「Kernel Complexity」（カーネルコンプレキシティ）では、データをノンパラメトリックにモデリングし、その構造の複雑性を表す連続値指標として「カーネル複雑さ」を提案している。カーネル複雑さは、GINI指標と呼ばれる情報量の偏りの程度を示す量において、情報量をカーネル密度に対するNML符号長として計算したものとして定義される。時系列データのノンパラメトリックな構造的変化を、カーネル複雑さの時間変化を追跡することにより検知するアルゴリズムを構成した。これにより、通常の混合分布を用いては表せないノンパラメトリックな構造の変化を検知できることを実験的に示した。

第６章「Conclusion」（結論）では全体を総括し、将来の展望を与えている。

以上を要するに、本論文は、動的設定におけるモデルの変化とその予兆の検知といった、機械学習・データサイエンスの分野の中でも新しい問題に対して、MDL原理に基づく方法論を体系的に提示しており、数理情報学の発展に大きく寄与している。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。