

(様式2)

学位論文の概要及び要旨

氏 名 加藤 聡 印

題 目 クラスタリング問題への適用に向けた自己組織化マップの学習法改善に関する研究

学位論文の概要及び要旨

本論文は、Kohonenによって提案された自己組織化マップ (Self-Organizing Map : SOM) をクラスタリングに適用する問題に対し、そのクラスタリング性能の改善を目的とした2段階SOM (2-stage SOM) およびその改良手法である拡張2段階SOMの提案を行い、人工的な評価用データセットや、UCI Machine Learning (UCI ML)データベースから引用した実データを用いて、これらのデータセットに対する提案手法におけるクラスタリング性能、ならびに従来のクラスタリング手法との比較検討などを行った研究成果をまとめたものである。本論文は全6章から構成されており、第1章は序論として、本研究に関連する従来の研究概要について述べ、本研究の目的ならびに、本研究を行うに至った背景および、各章の概要を述べている。

第2章では、クラスタリング問題の概要について、種々のクラスタリング手法の分類を行い、代表的なクラスタリングアルゴリズムである k -means法や、階層的クラスタリング手法である最短距離法、最長距離法、ワード法、また、グラフ理論を用いた手法について、それらの具体的なアルゴリズムと問題点について説明している。特に、非階層的クラスタリング手法の一つである k -means法は、クラスタリング結果が初期状態に大きく依存するという問題点があり、また、階層的クラスタリング手法においては、クラスタ間距離の定義の違いによって複数のバリエーションがあるため、適用する手法によってクラスタの併合過程が異なることや、計算量の観点から大規模データへの適用が困難であることを問題点として指摘している。本章では、これら従来のクラスタリング手法に対して指摘された問題点から、本論文で対象とする、SOMを用いたクラスタリングの位置付けを述べている。

第3章では、SOMを用いたクラスタリングの具体的な手法を示し、その問題点を指摘した上で、クラスタリング問題への適用に向けたSOMの改善学習アルゴリズムである、2段階SOMを提案している。SOMの学習アルゴリズムでは、ある競合層セルが受けたコードベクトルの更新が、そのセルに隣接したセルにも影響するという性質がある。これは近傍学習と呼ばれ、学習後のSOMの特徴マップにおける位相保持写像を可能としている。この位相保持写像は、SOMをクラスタリング問題に適用するにあたって非常に重要な特性である。しかしながら、位相保持写像を可能にするための近傍学習によって、クラスタとクラスタとの間にコードベクトルが残留し、これらのコードベクトルを持った不活性セルが、学習後のクラスタ抽出に悪影響を及ぼす。そこで本章では、SOMの基本学習アルゴリズム (BSOM) と、近傍学習にしきい値作用を導入した学習アルゴリズム (THSOM) とを段階的に

適用する手法（2段階SOM）を提案している。2段階SOMは、BSOMの適用によって得られたコードベクトルを、THSOM適用時の初期状態として用いるもので、位相保持写像の実現と不活性セルの抑制とを同時に達成することを目的としている。本章では、この2段階SOMに対して、人工的に作成した入力データによる学習実験およびクラスタリング実験を行い、2段階SOMでは、位相保持写像を維持しつつ不活性セルの発生が抑制されていることを確認している。

第4章では、正規分布型の形状を持つ複数のクラスタから構成されるデータや、任意の分布形状を持つもの、クラスタごとのデータ密度が異なるものなど、いくつかの評価用データセットを人工的に作成し、2段階SOMのクラスタリング性能と問題点について詳細に論じている。本章において、提案手法である2段階SOMは、各クラスタにおけるデータの密度が一定の場合には、従来のSOMやk-means法などと比較して、クラスタリング時における誤分類率の改善が見られることを確認している。一方で、各クラスタにおけるデータの密度が一定でない場合、2段階SOMにおける、他の手法に対する誤分類率改善の程度は高くなく、さらなる改良が必要であることも示している。本章では、誤分類率の改善が見られない原因について、2段階SOMのTHSOM過程におけるしきい値の尺度が、競合層セルのコードベクトル間の距離と、学習時におけるセルの勝利回数の値そのものに基づくことが問題であることを指摘している。

第5章では、前章において示された2段階SOMの問題点について、その改良手法である拡張2段階SOMの提案と、人工データおよび実データを用いた評価について述べている。拡張2段階SOMでは、隣接セルのコードベクトル間距離および、各セルのBSOM過程における勝利回数の変化量に注目し、THSOM過程におけるしきい値の尺度の算出法を変更している。本章では、まず人工的に作成した入力データを用いたクラスタリング実験を行って、拡張2段階SOMは従来型2段階SOMと比較して、各クラスタにおけるデータの密度が一定でない場合に誤分類率が改善されることを確認している。さらに本章では、機械学習の評価用の実データとして良く用いられているUCI MLデータベースのいくつかのデータセットを対象としたクラスタリング性能の評価も行っており、拡張2段階SOMと、従来型2段階SOM、k-means法、および階層的クラスタリング手法とを比較した場合の誤分類率についての議論において、拡張2段階SOMの有効性を示している。

最後の第6章において、本研究のまとめと今後の展望および課題など、全体の総括を行っている。