HOKKAIDO UNIVERSITY

Title	A Study of Efficient and Robust Clustering for Large-Scale Datasets [an abstract of dissertation and a summary of dissertation review]
Author(s)	劉, 浩
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11521号
Issue Date	2014-09-25
Doc URL	https://hdl.handle.net/2115/57167
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Hao_Liu_review.pdf, 審査の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　劉 浩

審査担当者　主　査　教　授　栗原 正仁
副　査　教　授　山本 雅人
副　査　教　授　鈴木 恵二
副　査　教　授　小野 哲雄
副　査　准教授　小山 聡

学位論文題名

A Study of Efficient and Robust Clustering for Large-Scale Datasets
（大規模データセットに対する効率的で信頼性の高いクラスタリングに関する研究）

クラスタリングは, データマイニング, 機械学習, パターン認識, 情報検索など, 多数の分野で応用されている重要な技術である。しかし, 近年はビッグデータと呼ばれる大規模なデータセットを扱う必要性が強まっているために, これまで開発されてきたロバスト性に優れた良いアルゴリズムは, 十分に効率が良いとは言えない状況になってきている。一方, 効率の良いアルゴリズムは, 特にノイズのある環境において, 新たなデータを学習する際に, ロバスト性の観点から十分な信頼性を示さない。

本論文は, この問題を解決するために, 十分に効率が良く, ロバスト性のある 2 つの新しい方法を提案している。第一の方法は, 入力されたデータセットを良く表現できる「ランドマーク」という概念を用いた, 密度基盤ファジィ・クラスタリングの効率良いアルゴリズムである。ここでは, ランドマークの数が実際のデータの数よりはるかに少ないため, 計算量が大幅に削減され, 効率化されている。第二の方法は, 漸増的な自己組織化ニューラルネットワークのロバストなアルゴリズムである。ここでは, アルゴリズムがネットワークに含めるべきニューロンの適切な数を動的に決定する際に, ノイズに対する安定性が保たれるような柔軟なネットワーク構造を構築することで, ロバスト性が保たれている。

第 1 章では, 本研究の背景として, クラスタリングの目的と現行の技術状況が簡潔に紹介された後, 大規模なデータセットに対するクラスタリングの問題が議論され, 本研究の目的および本論文で提案する 2 つの方法の概要が紹介されている。

第 2 章では, 本論文に関係する基礎知識が説明されている。古典的なクラスタリングとして, 分割的, 階層的, および密度基盤の方法と各々の特徴が説明され, その後, 自己組織化ニューラルネットワークなど, 比較的最近の漸増的なクラスタリング手法について, 議論されている。

第 3 章では, 既存の FN-DBSCAN アルゴリズムの効率改善を目指す密度基盤クラスタリング手法として, ランドマーク FN-DBSCAN アルゴリズムが提示されている。ランドマークの概念によってデータセットの部分集合が表現され, データセットが圧縮表現されている。理論解析によると, その計算量は, 時間および空間ともに, データセットの規模に比例する。実験結果によれば, 本アルゴリズムは,FN-DBSCAN の優れたクラスタリング品質を保ちつつ, それよりも高速に動作する。しかし, ランドマークの数が極めて少ないときには, そのクラスタリングの結果は, 入力されるデータ

の順序に影響を受ける点に欠点があることから, 著者はその解決を次章以降につなげている。

第 4 章では, ランドマークの生成タスクの問題を解決する局所適応漸増 LBG アルゴリズム (LAI-LBG) が導入されている。これは, 前章末で述べられた欠点を解消するために, 既存の LBG アルゴリズムを改善したもので, 事前にランドマークの数を固定する必要がなく, 漸増的にランドマークを挿入あるいは削除するとともに, 初期設定やデータセットの入力順序には感度が低いという性質をもつ。したがって, このアルゴリズムは, データセットを少数のランドマークによりバランス良く表現することができる。また, 同程度の品質をもつ既存のアルゴリズムよりも高い効率を実現している。

第 5 章では, 前章の方法を第 3 章の方法に応用して大幅に改良した新しい密度基盤クラスタリングアルゴリズムを提案している。この新しいアルゴリズムは, 良いクラスタリング結果を少数のランドマークにより得ることができ, 効率がさらに改善されているとともに, データセットの入力順序には感度が低いため, ロバスト性が向上している。

第 6 章では, 新しい漸増的な自己組織化ニューラルネットワークとして, 強化漸増成長ニューラルガス (Hi-GNG) が提示されている。ニューロンの誕生と死滅, ニューロン間結合の生成と除去, およびその強度の調整がモデル化され, 強化された競合ヘビアン学習 (CHL) が提案されている。それに基づき,Hi-GNG アルゴリズムは, 柔軟なトポロジー構造を, 適切な数のニューロンを用いて自動的に生成することができる。さらに, それはノイズデータに対してロバストでもあり, 既存の GNG アルゴリズムよりも効率が良い。

第 7 章では, 本論文で提案された方法が総括され, 今後の研究課題が議論されている。

これを要するに, 著者は, ランドマークの概念とニューラルネットワークの柔軟な漸増的構築を工夫することで, 大規模データセットに対して十分に効率が良く, ロバスト性の高いクラスタリングの方法を設計し, その性能について新知見を得たものであり, 知能情報学に対して貢献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 (情報科学) の学位を授与される資格あるものと認める。